ARTICLES

The Public Pension Crisis..........................................................Jack M. Beermann

Why *Strickland* Is the Wrong Test for Violations of the Right to Testify ...............Daniel J. Capra & Joseph Tartakovsky

Vertical Boilerplate .................................................................James Gibson

When Certainty Dissolves into Probability: A Legal Vision of Toxic Causation for the Post-Genomic Era .........................Steve C. Gold

Information Privacy and Data Control in Cloud Computing: Consumers, Privacy Preferences, and Market Efficiency ..............Jay P. Kesan, Carol M. Hayes & Masooda N. Bashir

Docket Dividends: Growth in Shareholder Litigation Leads to Refinements in Chancery Procedures ................Donald F. Parsons, Jr. & Jason S. Tyler

Risk-Based Student Loans ......................................................Michael Simkovic

THE WASHINGTON AND LEE LAW ALUMNI ASSOCIATION STUDENT NOTES COLLOQUIUM

Is Hedge Fund Adviser Registration Necessary to Accomplish the Goals of the Dodd–Frank Act’s Title IV? ........................Luther R. Ashworth II

Review: Is Hedge Fund Registration Necessary? .............................................................J.W. Verret

Why Register Hedge Fund Advisers— A Comment ..................................................Lyman Johnson

Sheltering Psychiatric Patients from the *DeShaney* Storm: A Proposed Analysis for Determining Affirmative Duties to Voluntary Patients .........................Claire Marie Hagan

Joshua’s Children: Constitutional Responsibility for Institutionalized Persons After *DeShaney v. Winnebago County* ..................................................Susan Stefan

Coercion, Consent, Compassion ..................................................John D. King

Vol. 70, No. 1(Winter 2013) Lexington, Virginia
WASHINGTON AND LEE UNIVERSITY
SCHOOL OF LAW

The LAW ALUMNI ASSOCIATION assists the School of Law by providing the following:

- Tuition Scholarships for students of the School of Law
- Faculty Fellowships to recognize excellence in scholarship and teaching

THE LAW ALUMNI ASSOCIATION
solicits your membership and support of the Law Annual Fund

Your contributions to the Law Annual Fund go toward such specific programs as Law Alumni Association Tuition Scholarships and Faculty Fellowships. Your contributions also help the School maintain its small classes, low faculty-to-student ratio, beautiful campus, and one-on-one attention. The Law Annual Fund enriches the Law School in hundreds of small but significant ways, and makes a critical difference in our ability to fulfill our mission. You can send your contribution to The Law Annual Fund, Washington and Lee University School of Law, Lexington, VA 24450, or you can give online at http://support.wlu.edu/web/page/normal/giftform.html.

Questions? Need more information? Please contact Elizabeth Outland Branner, Director of Law School Advancement, at (540) 458-8191 or brannere@wlu.edu, Sarah Hughes, Assistant Director of Law School Advancement, at (540) 458-8063 or shughes@wlu.edu, or Darlene Moore, Executive Secretary of the Law Alumni Association, at (540) 458-8571 or moored@wlu.edu.
Administration & Faculty

Washington and Lee University

KENNETH P. RUSCIO, B.A., M.P.A., Ph.D.
President of the University

Deans

NORA V. DEMLEITNER, B.A., J.D., LL.M.
Dean & Roy L. Steinheimer, Jr. Professor of Law

JOHANNA E. BOND, B.A., J.D., LL.M.
Associate Dean for Academic Affairs & Associate Professor of Law

JOHN G. KEYSER, B.S., M.S.
Associate Dean for Administration & Technology, Lecturer in Social Sciences & Law

MARY ZANOLLI NATKIN, B.A., J.D.
Assistant Dean for Clinical Education & Public Service, Clinical Professor of Law

STEPHEN TWITTY, B.A., J.D.
Assistant Dean for Student Affairs

SHAWN D. MCSHAY, B.S., M.B.A.
Assistant Dean for Admissions

Emeritus Faculty

DENIS J. BRION, B.S., J.D.
Professor of Law, Emeritus

WILLIAM S. GEIMER, B.S., J.D.
Professor of Law, Emeritus

EDWARD O. HENNEMAN, B.A., J.D.
Associate Professor of Law, Emeritus

FREDERIC L. KIRGIS, B.A., J.D.
Law Alumni Association Professor of Law, Emeritus

LEWIS H. LARUE, A.B., LL.B.
Class of 1958 Law Alumni Professor of Law, Emeritus

ANDREW W. McTHERAL, JR., A.B., M.A., LL.B.
James P. Morefield Professor of Law, Emeritus

JAMES M. PHENISTER, B.S., J.D.
Professor of Law, Emeritus

ROY L. STEINHEIMER, JR., A.B., J.D.
Robert E.R. Huntley Professor of Law, Emeritus

JOSEPH E. ULRICH, B.S., J.D.
Professor of Law, Emeritus

Permanent Faculty

C. ELIZABETH BELMONT, B.A., J.D.
Director, Community Legal Practice Center and Associate Clinical Professor of Law

DAVID I. BRUCK, B.A., J.D.
Clinical Professor of Law and Director, Virginia Capital Case Clearinghouse

CHRISTOPHER M. BRUNER, A.B., M.Phil., J.D.
Associate Professor of Law

SAMUEL W. CALHOUN, B.A., J.D.
Professor of Law

ROBERT T. DANFORTH, B.A., J.D.
Professor of Law

KEVIN M. CROTTSY, B.A., Ph.D., J.D.
Professor of Classics & Law
Administration & Faculty
Washington and Lee School of Law 2012–2013

Mark A. Druml, B.A., M.A., J.D., LL.M., J.S.D.
Class of 1975 Alumni Professor of Law and Director, Transnational Law Institute

Michelle L. Druml, B.A., J.D., LL.M.
Associate Clinical Professor of Law and Director, Tax Clinic

Joshua A.T. Fairfield, B.A., J.D.
Associate Professor of Law

Jill M. Fraley, B.A., J.D., LL.M., J.S.D.
Assistant Professor of Law

Susan D. Franck, A.A., B.A., J.D., LL.M.
Associate Professor of Law

Mark H. Grunewald, B.A., J.D.
James P. Morefield Professor of Law

Brant J. Hellwig, B.S., J.D., LL.M.
Professor of Law

Margaret Howard, A.B., J.D., M.S.W., LL.M.
Law Alumni Association Professor of Law

Lyman P.Q. Johnson, B.A., J.D.
Robert O. Bentley Professor of Law

Timothy S. Jost, B.A., J.D.
Robert L. Willett Family Professor of Law

John D. King, B.A., J.D., LL.M.
Associate Clinical Professor of Law

Erik Luna, B.S., J.D.
Sydney and Frances Lewis Professor of Law

Timothy C. MacDonnell, B.A., J.D., LL.M., LL.M.
Associate Clinical Professor of Law, Director, Black Lung Clinic

Ann Maclean Massie, B.A., M.A., J.D.
Professor of Law

Russell A. Miller, B.A., J.D., M.A., LL.M.
Professor of Law

David Millon, B.A., M.A., Ph.D., J.D.
J.B. Stombock Professor of Law

James E. Moliterno, B.S., J.D.
Vincent Bradford Professor of Law

Brian C. Murchison, B.A., J.D.
Charles S. Boss Professor of Law

Doug Rendleman, B.A., M.A., J.D., LL.M.
Robert E.R. Huntley Professor of Law

Christopher B. Seaman, B.A., J.D.
Assistant Professor of Law

Joan M. Shaughnessy, B.A., J.D.
Roger D. Great Professor of Law

A. Benjamin Spencer, B.A., M.Sc., J.D.
Professor of Law and Director, Frances Lewis Law Center

Sarah K. Want, B.A., M.L.S., J.D.
Professor of Law

Robin Fretwell Wilson, B.A., J.D.
Class of 1958 Law Alumni Professor of Law and Law Alumni Faculty Fellow

Visiting Faculty
Paul Belonick, B.A., M.A., J.D.
Visiting Professor of Law
Administration & Faculty

Washington and Lee School of Law 2012–2013

MICHAL BUCHHANDLER-RAFAEL, L.L.B., LL.M., S.J.D.
Visiting Assistant Professor of Law

DONALD EARL CHILDESS III, B.A., M.A., J.D., LL.M.
Visiting Associate Professor of Law

ROBERT CREIGH DEEDS, B.A., J.D.
Visiting Professor of Law

J. GORDON HYLTON, A.B., J.D., M.A., Ph.D.
Visiting Professor of Law

JUSTICE DONALD W. LEMONS, B.A., J.D.
Distinguished Professor of Judicial Studies

PATRICK LUFT, B.A., J.D.
Distinguished Professor of Judicial Studies

TODD C. PEPPERS, B.A., J.D., Ph.D.
Visiting Professor of Law

JONATHAN SHAPIRO, B.A., J.D.
Visiting Professor of Law

JAMES E. VIATOR, B.A., J.D.
Visiting Professor of Law

CHRISTOPHER J. WHelan, LL.B., M.A., Ph.D.
Visiting Professor of Law

Adjunct & Extended Faculty

HARLAN RAY BECKLEY, B.S., M.Div., M.A., Ph.D.
Lecturer in Religion & Law, Director, Shepherd Program, Fletcher Otey Professor of Religion

NEIL V. BIRKHOF, B.A., J.D., LL.M.
Adjunct Professor of Law

MICHELE K. BURKE, B.S., J.D.
Professor of Practice

R. EDWIN BURNETTE, JR., B.A., J.D.
Adjunct Professor of Law

ALBERT V. CARR, B.A., J.D.
Professor of Practice

JUDY CLARKE, B.A., J.D.
Professor of Practice

GREGORY J. COOPER, B.A., Ph.D.
Lecturer in Philosophy & Law, Associate Professor of Philosophy

JOHN H. CRADDICK, JR., B.S.E., J.D., LL.M.
Professor of Practice

DAVID DEJONG, B.A., J.D., LL.M.
Adjunct Professor of Practice

DANIEL EVANS, B.A. B.S., J.D.
Professor of Practice

KELLY L. FALGIONI, B.A., J.D.
Professor of Practice

JUDGE JOHN F. FORSTER, JR., B.A., J.D.
Professor of Practice and Director, Judicial Externships

AARON SAMUEL HAAS, B.A., J.D.
Oliver Hill Fellow (Adjunct)

DAVID L. HEILBERG, B.A., J.D.
Adjunct Professor of Practice

JAMES W. JENNINGS, JR., B.A., J.D.
Professor of Practice
Administration & Faculty
Washington and Lee School of Law 2012–2013

TIMOTHY JOSEPH KEEFER, B.A., J.D.
Adjunct Professor of Law

REX LAMB, B.A., J.D.
Adjunct Professor of Law

CHARLES LOWNY, B.A., M.A., Ph.D.
Lecturer in Philosophy & Law, Visiting Professor of Philosophy

TIMOTHY LUBIN, B.A., M.T.S., Ph.D.
Lecturer in Law and Religion, Professor of Religion

A. CARTER MAGEE, JR., B.A., J.D.
Professor of Practice

JAMES E. MAHON, B.A., M. Phil., Ph.D.
Lecturer in Philosophy and the Law, Professor of Philosophy

JOHN LAWRENCE MANNING, B.S., LL.B., J.D.
Adjunct Professor of Law

JOHN J. MILES, B.S., M.A., J.D.
Adjunct Professor of Law

FRANK W. MORRISON, B.A., J.D.
Adjunct Professor of Law

LAWRENCE LLOYD MUIR, JR., B.A., J.D.
Adjunct Professor of Practice

STEPHEN WILLS MURPHY, B.A., M.Phil., Ph.D., J.D.
Adjunct Professor of Law

CAROL D. NEWMAN, B.A., Ph.D., J.D.
Professor of Practice

J. LEE E. OSBORNE, B.A., J.D., LL.M.
Adjunct Professor of Law

G. MICHAEL PACE, JR., B.A., J.D.
Professor of Practice

JAMES H. PANNABECKER, B.A., J.D., M.R.P.
Professor of Practice

HOWARD PICKETT
Adjunct Instructor of Shepard Poverty

THOMAS H. SPEEDY RICE, B.A., J.D.
Professor of Practice, Transnational Law Institute

CHRISTOPHER B. RUSSELL, A.B., J.D.
Adjunct Professor of Law

GILBERT E. SCHILL, JR., B.A., J.D.
Professor of Practice

JAMES S. SEEVERS, JR., B.A., J.D.
Professor of Practice

GUY BRYAN SEREFF II, B.S., J.D.
Professor of Practice

JONNIE SPEIGHT, B.A., J.D.
Adjunct Professor of Law

GREGORY N. STILLMAN, B.A., J.D.
Professor of Practice

ADAM ROBERT TRUSNER, B.A., J.D.
Adjunct Professor of Law

WILSON F. VELLINES, Jr., B.A., J.D.
Adjunct Professor of Law

DANIEL J. VICATOR, B.A., J.D.
Professor of Practice
Administration & Faculty\Washington and Lee School of Law 2012–2013

William P. Wallace, Jr., B.A., J.D.
Adjunct Professor of Law

Mark Allan Williams, B.A., J.D., MBA.
Professor of Practice

R. Craig Wood, B.A., M.Ed., J.D.
Professor of Practice

Robert C. Wood, III, B.A., LL.B.
Adjunct Professor of Law

Henry L. Woodward, A.B., LL.B.
Professor of Practice

Charles L. Woody, B.A., J.D.
Adjunct Professor of Law
Each year, the Frances Lewis Law Center, the Washington and Lee Law Review, and other Washington and Lee University School of Law organizations co-sponsor at least one major symposium examining prospects for change in an important area of the law.

Preceding Major Symposia

Gideon at 50: Reassessing the Right to Council
(November 8–9, 2012)

Regulation in the Fringe Economy
(November 10–11, 2011)

Restitution Rollout: The Restatement (Third) of Restitution & Unjust Enrichment
(February 25, 2011)

Prosecutorial Power: A Transnational Symposium
(April 1–2, 2010)

Protecting Virtual Playgrounds: Children, Law, and Play Online
(October 3, 2008)

The Restitution Roundtable
(December 14, 2007)

Gender-Relevant Legislative Change in Muslim and Non-Muslim Countries
(March 30, 2007)

Understanding Corporate Law Through History
(March 24, 2006)

For information about upcoming symposia, please contact:
Wendy Rice, Frances Lewis Law Center, Washington and Lee University School of Law, 1 Denny Circle, Lexington, Virginia 24450; (540) 458-8962.

WASHINGTON AND LEE LAW REVIEW

Statement of Ownership, Management, and Circulation: The Washington and Lee Law Review (ISSN 0043-0463) is owned by Washington and Lee University and is published quarterly by students of Washington and Lee University School of Law. The Law Review is printed by Western Newspaper Publishing Co., Inc., 537 East Ohio Street, Indianapolis, Indiana 46204, and is entered as a periodical with postage paid at post offices in Lexington, Virginia, and Indianapolis, Indiana. POSTMASTER: Send address corrections to Washington and Lee Law Review, Room 408, Washington and Lee University School of Law, 1 Denny Circle, Lexington, Virginia 24450-2116. Alexander M. Sugzda is the 2012–2013 Editor in Chief. The phone number is (540) 458-8566; the fax number is (540) 458-4679; and the e-mail address is: lawreview@wlu.edu. The average number of copies printed of each issue during the preceding twelve months was 520.

Subscription Information: The subscription price for Volume 70 (four issues) of the Washington and Lee Law Review is $35.00 ($40.00 for subscriptions outside the United States). Please e-mail all subscription requests to the Washington and Lee Law Review at lawreview@wlu.edu. Subscriptions will be renewed automatically unless the Law Review receives notice of cancellation by December 31.

Back Issues: Volumes 68 through 70, issue 1 may each be purchased directly from the Washington and Lee Law Review at the current subscription price of $35.00 ($40.00 for foreign subscriptions) per volume. Individual issues cost $25.00 (includes postage). Volumes 1 through 63 may be obtained from:

William S. Hein & Co., Inc.
1285 Main Street
Buffalo, NY 14209-1987
(800) 828-7571; order@wshein.com.

Copies of Articles: Except as otherwise provided, the author of each article in this issue and the Washington and Lee Law Review grant permission for copies to be made available for classroom use, provided that: (1) the copies are distributed at or below cost; (2) the author and the Law Review are identified on each copy; (3) a proper notice of copyright is affixed to each copy; and (4) the user notifies the Law Review that copies have been made. Articles are also available through Westlaw, LEXIS, and Hein Online.

World Wide Web Address: The Washington and Lee Law Review maintains a website at http://law.wlu.edu/lawreview. The website provides general information regarding the Law Review and may be used to order subscriptions, notify the Law Review that copies have been made for classroom use, and contact individual editors and staffwriters.

AN INVITATION TO AUTHORS

The Editorial Board of the Washington and Lee Law Review invites authors to submit articles and essays on areas of the law that are currently of interest to the legal profession. Student-written work is not accepted.


Submission of Manuscripts: Please submit manuscripts to Joanna Heiberg or Anthony Raucci, Senior Articles Editors, Washington and Lee Law Review, Room 408, Washington and Lee University School of Law, 1 Denny Circle, Lexington, Virginia 24450-2116. Electronic submissions may only be made through our ExpressO webpage at http://law.bepress.com/cgi/submit_lawreview.cgi?a=Washingtonallr. Alternatively, hard copy submissions are accepted. Please do not duplicate your submission by sending a hard copy as well as an electronic copy.

All manuscripts must be typed in Microsoft Word, not converted from any other software. The font of both the text and the footnotes should be 11 pt., Century Schoolbook, and double-spaced on 8.5 x 11 inch paper. Footnotes should appear at the bottom of each page of text. Text and footnotes should conform to The Bluebook: A Uniform System of Citation (19th ed. 2010). Article submissions should be accompanied by an abstract of not more than 300 words.
ARTICLES

The Public Pension Crisis

Jack M. Beermann

Why *Strickland* is the Wrong Test for Violations of the Right to Testify

Daniel J. Capra
Joseph Tartakovsky

Vertical Boilerplate

James Gibson

When Certainty Dissolves into Probability: A Legal Vision of Toxic Causation for the Post-Genomic Era

Steve C. Gold

Information Privacy and Data Control in Cloud Computing: Consumers, Privacy Preferences, and Market Efficiency

Jay P. Kesan
Carol M. Hayes
Masooda N. Bashir

Docket Dividends: Growth in Shareholder Litigation Leads to Refinements in Chancery Procedures

Donald F. Parsons, Jr.
Jason S. Tyler
Risk-Based Student Loans 527

Michael Simkovic

THE WASHINGTON AND LEE LAW ALUMNI ASSOCIATION STUDENT NOTES COLLOQUIUM

Is Hedge Fund Adviser Registration Necessary to Accomplish the Goals of the Dodd–Frank Act’s Title IV? 651

Luther R. Ashworth II

Review: Is Hedge Fund Registration Necessary? 705

J.W. Verret

Why Register Hedge Fund Advisers—A Comment 713

Lyman Johnson

Sheltering Psychiatric Patients from the DeShaney Storm: A Proposed Analysis for Determining Affirmative Duties to Voluntary Patients 725

Claire Marie Hagan

Joshua’s Children: Constitutional Responsibility for Institutionalized Persons After DeShaney v. Winnebago County 793

Susan Stefan

Coercion, Consent, Compassion 823

John D. King
ARTICLES
The Public Pension Crisis

Jack M. Beermann*

Abstract

Unfunded employee pension obligations will present a serious fiscal problem to state and local governments in the not-too-distant future. This Article takes a look at the causes and potential cures for the public pension mess, mainly through the lens of legal doctrines that limit public employers’ ability to avoid obligations. As far as the causes are concerned, this Article examines the political environment within which public pension promises are made and funded, as an attempt to understand how this occurred. The Article then turns to ask if states could implement meaningful reforms without violating either state or federal law. In particular, the Article looks at state balanced budget requirements, state constitutional provisions regarding public employee pensions, and federal constitutional law and asks whether states could significantly reduce their pension promises to public employees without violating the law. The entire analysis is also informed by the concerns of the employees and retirees whose perhaps sole source of retirement income would be reduced by changes in benefit levels. The Article concludes with remarks placing the matter in that context, raising the possibility of a bailout to ameliorate the potentially disastrous consequences of reform to public employees and retirees.

* Professor of Law and Harry Elwood Warren Scholar, Boston University School of Law. Thanks to Nancy Altman, Robert Costrell, Ward Farnsworth, Larry Kotlikoff, Steve Levine, Eric Madiar, and Andrew Maylor for their help with this project. Special thanks to Ted Orson, counsel for the City of Central Falls and the State of Rhode Island in the city’s bankruptcy proceeding for taking the time to educate me on the ins and outs of municipal bankruptcy. For excellent research assistance, I thank Bryn Sfetsios, Boston University School of Law Class of 2013, Marisa Tripolsky, Boston University School of Law Class of 2014, and Matt Martin and Erin Monju, both of Harvard Law School Class of 2012.
I. Introduction

The first decade of the new millennium was a difficult one for state and local government finances, and the second decade has started out even worse. In addition to the difficulty governments at all levels are experiencing in trying to maintain services without raising taxes, some analysts claim that many state and local governments are sitting on a fiscal time bomb—underfunded public employee pension and health care liabilities1 that threaten to destroy the fiscal well-being of many state and local governments. Some accounts predict that absent significant benefits reductions (which may not be legally feasible), state and local governments will soon be devoting an untenably large

1. This Article focuses primarily on unfunded pension liabilities. State and local governments also have substantial unfunded health care liabilities, and a few distinct aspects of that problem are highlighted in this Article.
portion of their budgets to making pension payments and satisfying other obligations to retired workers.

Unfunded liabilities are possible because government pensions are still largely defined benefit plans, and the law generally does not require full advance funding of the projected costs of accrued benefits. In a defined benefit plan, an employee is promised a specific dollar amount of retirement benefits, usually based on the employee’s final salary. These promises are often accompanied by promises of lifetime government-financed health care, without regard to the cost to the public employer. Although states operate under balanced budget requirements, it turns out that underfunding pension obligations does not violate these state law requirements. Thus, current taxpayers are able to push off pension and other promises to retirees to future generations of taxpayers.

Private industry has moved away from defined benefit plans toward contribution plans, under which employers contribute a fixed amount to an employee’s retirement plan and the employee receives retirement benefits based on the performance of the investments purchased with the contributions. The advantage of a contribution plan to employers is obvious—certainty. Once an employer makes the contributions required under the plan, there is no chance that actuarial miscalculations or market downturns will require additional contributions in the future. The employee, not the employer, bears the risk of a market downturn or inflation that might reduce the value of the pension.

Defined contribution plans also have some advantages for employees. First, employees may gain control over their funds and have the power to direct investments to their preferred level of risk. Second, employees’ retirement funds are not subject to the solvency of the employer. There is no opportunity for employers to manipulate contribution levels. Further, once the employer’s money is deposited into the account, the employer cannot raid the fund or take any other action that would prejudice the employees’ ownership of the fund.


3. This is not to say that private contribution retirement plans are risk-
While most public employers and employees in the United States set aside money each year to fund future projected pension obligations, many public pension plans are seriously underfunded either intentionally or due to unrealistic assumptions concerning investment performance and the amount that will be owed over time. This means that unless contributions are increased substantially, future pension payments to retired government workers will be made, at least in part, from current revenues. The problem is thought to be so serious that some local governments may be effectively insolvent. Retirees face the risk of reduced pension payments and current employees face the risk of receiving less generous retirement benefits than the promises that they have been depending upon.

In the private sector, the Employee Retirement Income Security Act (ERISA) and programs administered by the Pension Benefit Guaranty Corporation provide a mechanism to deal with insolvent pension plans and the outstanding pension obligations of bankrupt private firms. The financial consequences of pension plan insolvency to private companies and their employees may be disastrous, but ultimately they can be resolved in an orderly manner without forcing the company to pay all of its obligations. State and local governments have fewer options. State law and the Contract Clause of the United States Constitution may make it impossible for states to enact meaningful pension reform or simply discharge obligations that are too difficult to meet. Even if a state is insolvent, the federal Constitution may demand complete payment of all pension obligations. Bankruptcy may be

free, but federal regulations under ERISA, Employee Retirement Income Security Act of 1974, 29 U.S.C. §§ 1001–1461 (2006), prevent private companies from failing to make required payments. Employees may suffer if their plan is terminated due to the insolvency of the employer or the inability or unwillingness of the employer to continue to contribute, but past contributions are largely safe in private plans.

6. For a suggestion that ERISA be extended to state and local pensions, see Jon G. Miller, Is Your Client's Government Pension Safe?: Making the Case for Federal Regulation, 2 ELDER L.J. 121, 121 (1994).
7. See infra Part III.
8. U.S. CONST. art. 1, § 10, cl. 1.
an option for some municipalities, but this very drastic step is not open to all municipalities, and is not available to the states themselves.

Even if everyone agreed that the best option would be to move away from defined benefit plans to defined contribution plans, implementing this change could be difficult because of the magnitude of unfunded liabilities. If paying current retirees’ benefits depends on contributions from active government employees and current tax revenues, it may be impossible to move current employees to contribution plans without magnifying the crisis beyond manageability.

The public pension crisis raises three separate concerns. The first involves the potential fiscal disaster that some predict will occur years from now, when public employers are required to pay the pension benefits they have been promising to public employees for many years. The second concern is the reduction in government services that may be necessary to make these payments, which could lead to great taxpayer dissatisfaction and political instability. The third concern involves the consequences to public employees and retirees, especially those who did not participate in Social Security, who could be left with insufficient assets for a decent retirement.

Underfunding public pensions is in substance, if not in form, an example of deficit spending in which current taxpayers enjoy the benefits of government services while pushing off some of the costs to future taxpayers. It is a double whammy for those future taxpayers—they will not only be required to pay for the consumption of prior generations, but will also receive reduced government services as state and local governments allocate funds to pensions and health care for retired workers rather than services for current taxpayers.

It should be noted that some analysts deny that there is a crisis in public pension costs looming just over the horizon. In their view, the total unfunded pension and health care liability of state and local governments is relatively small when compared to the overall revenues of state and local government. They also point out that the average pension earned by retired government

---

9. See infra notes 28–36 and accompanying text.
10. See infra note 31 and accompanying text.
workers is small—under $20,000 per year. On this view, the “pension” crisis is an effort by conservative political forces to undermine public employee unions whose members tend to support liberal politicians and views.

Although the matter is not free from doubt, this Article proceeds on the assumption that there is at least some truth to the conclusion reached by many, that pension obligations will present a serious fiscal problem in the not-too-distant future. This Article takes a look at the causes and potential cures for the public pension mess, mainly through the lens of legal doctrines that limit public employers’ ability to avoid obligations. As far as the causes are concerned, this Article examines the political environment within which public pension promises are made and funded, as an attempt to understand how this occurred. The first issue here is whether the promises governments have made to public employees are extravagant in light of the pay, benefits, job security, and opportunities for advancement of state and local government workers as compared to workers in private industry. The Article then turns to ask if states could implement meaningful reforms without violating either state or federal law. In particular, the Article looks at state balanced budget requirements, state constitutional provisions regarding public employee pensions, and federal constitutional law and asks whether states could significantly reduce their pension promises to public employees without violating the law. The entire analysis is also informed by the concerns of the employees and retirees whose perhaps sole source of retirement income would be reduced by changes in benefit levels. The Article concludes with remarks placing the matter in that context, raising the possibility of a bailout to ameliorate the possibly disastrous consequences of reform to public employees and retirees.


12. See infra notes 37–38 and accompanying text (providing examples of optimistic analysts).
II. The Political Economy of Public Pensions

There are at least three separate issues regarding the political economy of public pension funding. First is the basic question of whether unfunded retirement promises to government workers constitute a fiscal crisis or whether the issue has been created as a means of attacking public employee unions or generally attempting to reduce compensation to public workers.\(^\text{13}\) The second issue concerns the nature of retirement promises to government workers: Are the promises excessive and subject to manipulation and abuse, or are they simply part of a perhaps generous, but reasonable overall, compensation package? The final issue is, assuming that public employee retirement benefits are excessive or subject to abuse, how did this happen: Why would elected officials provide excessive retirement benefits to government employees?

\(^{13}\) Attention to the underfunding of public pensions is not new. An early hint at the forthcoming crisis was a 1976 Harvard Law Review Note discussing potential problems that might arise regarding public pensions in difficult fiscal times, such as altering the eligibility and benefits rules and moving investments into state securities. See Note, Public Employee Pensions in Times of Fiscal Distress, 90 Harv. L. Rev. 992, 992–93 (1976) [hereinafter Public Employee Pensions] (noting the rise in public employee pension funds and the riskiness of these programs). In 1978, the Pension Task Force Report on Public Employee Retirement Systems estimated state and local unfunded pension liabilities at $150 to $175 billion. Staff of H. Comm. on Educ. & Labor, 95th Cong., Pension Task Force Rep. on Public Employee Retirement Systems 165 (Comm. Print 1978). A 1979 report to Congress by the Comptroller General characterized the underfunding of state and local pensions as a national problem. See U.S. Gov’t Accountability Office, HRD-79-66, Funding of State and Local Pension Plans: A National Problem 2 (1979) (reporting on “the magnitude of unfunded accrued liabilities, actions or lack of actions being taken to fund the plans on a sound actuarial basis, and the fiscal impact of requiring actuarial funding on state and local governments”). This report noted that most of the pension funds it analyzed were underfunded using ERISA standards. Id. at 19. A 1981 article in the journal Public Choice posited two explanations for continued growth in unfunded pension liabilities: increased income of municipal employees made deferred compensation more attractive to the employees and demand for public services, due to baby boomers going to public schools, grew faster than the tax base, which made deferred compensation attractive to governments. See Dennis Epple & Katherine Schipper, Municipal Pension Funding: A Theory and Some Evidence, 37 Pub. Choice 141, 170 (1981).
A. How Large Is the Potential Fiscal Problem?

The public pension crisis is all over the news. Analysts refer to unfunded pension obligations as a ticking fiscal time bomb likely to cause serious problems in the future. California is the state with the largest unfunded pension obligations, and a recent report predicts that without significant, immediate reform, public services in California will face drastic cuts as more and more of the state’s budget is devoted to making pension payments. Other analysts dispute this and argue that pension obligations constitute a relatively small portion of state budgets and should be manageable over time. Which view is more accurate?

Those claiming that there is a public pension funding crisis seriously outnumber those making the contrary claim that pension debt is manageable. One study reported that unfunded obligations to public school teachers alone have been stated to total $332 billion, but the study’s own calculations put the figure at $933 billion, or nearly a trillion dollars. The Pew Center estimates are on the lower end, with a total of $1.38 trillion estimated to be underfunded for both pensions and retiree health care benefits for all state and local employees. A report by the

14. Problems with funding of public pensions are not confined to the United States. See Eduard Ponds, Clara Severinson & Juan Yermo, Funding in Public Sector Pension Plans-International Evidence 4 (Nat’l Bureau of Econ. Research, Working Paper No. 17082, 2011) (discussing global pension fund issues). Many nations have underfunded public employee pension plans. Id. at 21, 28. Some are completely or partly “pay as you go,” which means by design, no funds are set aside to pay future pension obligations—all benefits are paid out of the current budget. Id. at 7.


Little Hoover Commission, a bipartisan state oversight agency, estimates the unfunded liabilities of California’s ten largest public pension plans (of a total of eighty-seven studied) at $240 billion and predicts that large cities in California will soon be devoting one-third of their operating budgets to pension payments.19 Another study concludes that to achieve full funding, government contributions to employee retirement, including social security and pensions, will have to increase by 250%, representing 14.1% of total revenues.20 A Mercatus Center study has estimated the national gap to be approximately $3 trillion,21 as does a 2012 report by a group chaired by former New York Lieutenant Governor Richard Ravitch and former Federal Reserve Board Chair Paul Volcker.22 A new study, published in

21. See Eileen Norcross & Andrew Biggs, The Crisis in Public Sector Pension Plans: A Blueprint for Reform in New Jersey 1 (Mercatus Ctr., Working Paper No. 10-31, 2010). One problem is that there is no uniform standard for reporting the level of pension funding. See id. (“Using methods that are required for private sector pensions, which value pension liabilities according to likelihood of payment rather than the return expected on pension assets, total liabilities amount to $5.2 trillion and the unfunded liability rises to $3 trillion.”). For a proposal to create a uniform legal standard for reporting pension funding, see Daniel J. Kaspar, Defined Benefits, Undefined Costs: Moving Toward a More Transparent Accounting of State Public Employee Pension Plans, 3 WM. & MARY POL’Y REV. 1 (2011).
July 2012, comes to this startling conclusion: “[T]he average public employee pension plan in the United States is only around 41 percent funded while total unfunded liabilities as of 2011 are roughly $4.6 trillion.”23 Another analysis, by an economist at the Center for Economic Policy and Research, estimates the shortfall at $647 billion, using traditional rates of return for pension fund assets.24 This is a significant shortfall, but much lower than the $3 or $4 trillion figures used by others.

To put the magnitude of underfunding in perspective, the federal government’s total debt, as of March 2012, is approximately $16.5 trillion25 as compared to $3.8 trillion in annual spending, while total state and local spending per year is approximately $3.2 trillion with an estimated $2.99 trillion total debt.26 It is unclear whether this estimate of state and local debt includes unfunded pension liabilities. Assuming it does not, counting $3 trillion in unfunded pension liability and $1 trillion

23. ANDREW G. BIGGS, STATE BUDGET SOLUTIONS, PUBLIC SECTOR PENSIONS: HOW WELL FUNDED ARE THEY, REALLY? 1 (2012), http://www.statebudgetsolutions.org/doclib/20120716_PensionFinancingUpdate.pdf (describing how public pension plans currently value their financial health). This study also observes that the funding problem has gotten much worse relatively recently: “According to standard actuarial accounting, the average public pension has fallen to around 75 percent in 2011, versus 103 percent in 2000.” Id.


in unfunded retiree health care benefits promises would put the total state and local debt at approximately $6.5 trillion, or about 2.5 years of total spending, while the federal government’s debt equals more than 4 years of total federal spending. Websites like pensiontsunami.com (devoted to California’s pension issues) exemplify the near-consensus that pension obligations are a ticking fiscal time bomb for state and local governments.27

The contrary view—that there is no public pension funding crisis—is best exemplified by an article published on the Huffington Post titled, An Overblown ‘Crisis’ For State Pension Funds28 and Monique Morrissey’s study titled Discounting Public Pensions: Reports of Trillions in Shortfalls Ignore Expected Returns on Assets.29 These articles claim that state and local pension obligations are manageable, and that the contrary view is based on conservative analysts using low projected rates of return on pension fund assets to make the funding gap look larger than it actually is.30 Morrissey’s study claims that to meet the actual shortfalls, state and local governments would have to increase their pension funding from 4% of their budgets to 5%, a significant but manageable increase.31 While many studies attack state and local pension funds for justifying low current contributions by predicting an 8% return on investments, Morrissey claims that 8% is historically accurate and more realistic than the much lower Treasury Bill rate used by those claiming that a crisis exists.32

28. Carter, supra note 16.
29. MORRISSEY, supra note 16.
32. Id. at 2–3.
A particularly comprehensive study concluding that unfunded pension liabilities do not present a severe problem was published in 2007 by the Government Accountability Office (GAO), the research arm of Congress. That study found that “the additional pension contributions that state and local governments will need to make in future years to fully fund their pensions on an ongoing basis are only slightly higher than the current contribution rate.” Specifically, the study found that “contribution rates would need to rise to 9.3 percent of salaries—less than a half percent more than the 9.0 percent contribution rate in 2006.” The GAO report was much more concerned about health care costs because many governments do not set aside anything to fund health care promises, and if health care costs continue to rise, it may be difficult for the promises to be fulfilled.

The Huffington Post article reveals the political nature of this dispute. The article characterizes the Economic Policy Institute, which concludes that there is no serious problem, as “partly funded by unions,” and attacks the Mercatus study as unreliable at least partly because the Mercatus Center is funded by the Koch brothers, well-known conservative activists.


34. Id. at 27.

35. Id. This study was conducted before the financial crisis and recession that began in 2008, so it is unclear if these calculations are still accurate. For a slightly more recent study of the funding status of state and local government retiree benefits, see U.S. GOV’T ACCOUNTABILITY OFFICE, GAO-08-223, STATE AND LOCAL GOVERNMENT RETIREE BENEFITS: CURRENT FUNDED STATUS OF PENSION AND HEALTH BENEFITS (2008).

36. Another report concludes that while in 2010 3.8% of state and local budgets were devoted to paying pension costs, that figure would rise to somewhere between 5% and 12.5%, depending on the health of the plan and investment outcomes. See Alicia H. Munnell, Jean-Pierre Aubry & Laura Quinby, CTR. FOR RET. RESEARCH, THE IMPACT OF PUBLIC PENSIONS ON STATE AND LOCAL BUDGETS 1 (2010) (examining “the size of the additional funding relative to state budgets” for public pensions). Some states, however, would have more serious problems. For example, the authors predict that Illinois, a state with severe underfunding of pension plans, may have to devote approximately 17% of its state budget to meet all of its pension obligations. See id. at 6, fig.9.

37. See Carter, supra note 16.
Morrissey points out that the same conservatives who use the low Treasury Bill rate as the expected return on pension fund assets touted privatization of Social Security accounts on the basis of much higher returns in the stock market.38

My sense is that while there may be some exaggeration out there, the pension funding crisis is real. In a detailed review of public pension financing, Jonathan Forman makes a convincing case that there is a funding problem.39 As Forman explains:

Because governments tolerate an 80% funding level and use actuarial valuations instead of market valuations, public pensions are almost guaranteed to be underfunded. Public sector workers tend to get larger pensions as a result, but much of the cost of those larger pensions is pushed onto future generations of taxpayers.40

The 2012 analysis by a group led by Paul Volcker, with distinguished members such as Alice Rivlin, Nicholas Brady, and George Shultz, concludes that unfunded pension and retiree health care liabilities are significant and, absent serious reform, will contribute to future fiscal problems.41 The amount of time and energy being devoted to raising alarms about the fiscal consequences of promises to retirees by responsible groups seems out of proportion if the purpose is to mount an indirect attack on public employee unions and public collective bargaining. While

39. See Jonathan B. Forman, Funding Public Pension Plans, 42 J. MARSHALL L. REV. 837, 837 (2009) (discussing “the major financial, accounting, and legal issues that relate to the funding of state and local government pension plans” and options to ensure public employees will have future retirement benefits).
40. Id. at 860. Forman explains that “bad things happen” when pension funds are fully funded because employees often successfully lobby for increased pension benefits and legislatures reduce payments or take funding “holidays” to use the money for other purposes. Id. at 860–61. Surprisingly, Forman nevertheless calls for full current funding but proposes a more radical restructuring for future government employees that would either eliminate the traditional method of calculating pension payments based on the highest salary or replace benefit plans with contribution plans. Id. at 870–73.
some politicians may have used this pension issue as a basis for attacking public employee unions, there seems to be genuine concern over future pension funding from a diverse array of observers, including the New York Times, which does not generally carry the flag for conservative causes.

The situation with health care promises to retirees may be even worse than the pension problem because fewer state and local government entities have set aside any funds to pay for those expenses. Coupled with serious inflation in the cost of health care and health insurance, the failure to set aside funds to pay for this may prove disastrous as more workers retire.

B. Are Public Pension Promises Excessive or Abusive?

While the point is subject to dispute, let us assume that unfunded promises to current and future retirees constitute a significant fiscal problem for state and local governments. The next set of questions involves whether excessive promises of retirement benefits have been made to public employees and whether public pension plans are subject to abuse.

The defense of defined benefit public pensions often begins by pointing out that the average government employee pension is less than $20,000 per year, which certainly does not sound excessive. It is not clear, however, whether this is a meaningful figure. There are many government pension recipients who worked for the government just long enough to qualify, and who thus receive very small pensions. What really needs to be examined is the pension available to the government employee who makes a career in government service, how that fits into the overall compensation package for government employees, and how public retirement benefits compare to the retirement benefits available to private sector employees.

42. Retirement Income, supra note 11.

43. Another issue pertinent to evaluating the generosity of public pension promises is the age at which public employees can retire. For example, in Wisconsin, which seems to be typical, “[m]ost public-sector employees are able to retire at age fifty-seven with a full pension if they have at least thirty years of services.” Paul M. Secunda, Constitutional Contracts Clause Challenges in Public Pension Litigation, 28 Hofstra Lab. & Emp. L.J. 263, 273–74 (2011). Full pension benefits at age fifty-seven is generous on its own, and also means that
One possibility that should be dismissed is to make a direct comparison between public sector pensions and federal Social Security retirement benefits. One could imagine comparing contributions and benefits and ask whether public pension recipients are receiving overly generous benefits. There are two sets of reasons why this comparison is not apt. First, Social Security taxes pay for aspects of the program that go far beyond retirement benefits. In addition to retirement benefits, the payroll deductions required by the Federal Insurance Contribution Act (FICA) \(^44\) pay for disability benefits, survivor benefits for spouses and children, a small death benefit, and potential benefits for multiple former spouses. \(^45\) Further, Social Security is fully portable between jobs. Second, public sector pensions are part of the state and local employees’ compensation packages from their employer. In principle, the magnitude of their contributions to the fund is irrelevant to whether the pension promises are overly generous. When a person decides whether to accept government employment, and to remain in government employment when other opportunities arise, pensions and other postemployment benefits are undoubtedly part of the calculus. Current salaries may be lower for government employees in the public sector than for workers in the private sector, and the public sector may offer fewer opportunities for advancement, especially for those without political connections. Greater job security, pensions, and retiree health care promises may balance these factors out, so that overall the promises to retirees are not out of line. As a form of deferred compensation, public sector pensions may be perfectly reasonable.

Thus, even if it is true, as one study claims, that public pensions can be 4.5 times higher than Social Security benefits based on the same work history, \(^46\) this may not establish

---


\(^45\) Id.

anything about the fairness of public pensions. This possibility should also be tempered by the fact that Social Security recipients contribute less than many public pension recipients. Before recent stimulus measures, the combined employer-employee contribution to Social Security was 12.4% of the first $110,000 of income, while the combined contribution to public pensions in some jurisdictions may be closer to 20% or even more. There may be states and localities in which employees are required to contribute much less, with the expectation that the government will fund retirement benefits, but again, the real question is whether the pension is reasonable as an element of compensation, not as a direct comparison with Social Security benefits.

This picture is complicated by disagreement over whether public-sector workers truly earn less in current and overall compensation than their private-sector counterparts. In some circles, it is now widely thought that public-sector workers earn

47. It may also be the case that Social Security is underpaying based on contributions. For an argument that Social Security is a bad deal for current workers, see generally Jagadeesh Gokhale & Laurence J. Kotlikoff, Social Security’s Treatment of Postwar Americans: How Bad Can It Get?, in THE DISTRIBUTIONAL EFFECTS OF SOCIAL SECURITY REFORM (Martin Feldstein et al. eds., 2002).


49. One report states that in Missouri, combined teacher and employer contributions have risen to 29% of salary in an attempt to accumulate sufficient equity to support promised pensions. See Robert Costrell, Michael Podgursky & Christian Weller, Fixing Teacher Pensions, EDUCATION NEXT, Fall 2011, at 60–69, http://educationnext.org/files/ednext_20114_forum.pdf. (containing an exchange between Professors Costrell and Podgursky on one side, and Professor Weller on the other). Costrell and Podgursky advocate tying pension payments to contributions, with employers guaranteeing a level of payments in case of investment underperformance. Id. at 64–65.

50. It may be more useful to compare the replacement rate of public pensions with the replacement rate of private pensions. The replacement rate is the percentage of salary replaced by the pension. In 1985, a study calculated that the average worker retiring in 1984 at a $40,000 salary with forty years of service received a pension replacing 32.3% of salary. See DONALD SCHMITT, MONTHLY LABOR REVIEW, TODAY’S PENSION PLANS: HOW MUCH DO THEY PAY? 22 tbl.5 (1985). These retirees would also receive Social Security benefits, which would replace another portion of their salaries. Still, this is likely to be a lower replacement rate than what many public sector employees receive today.
greater total salary and benefits than comparable private-sector workers. For example, the Bureau of Labor Statistics found that in December 2010, private-sector workers earned approximately $28 per hour in total compensation, while their public-sector counterparts at the state and local level earned approximately $40.\textsuperscript{51}

Politicians have noticed this purported fact. Indiana Governor Mitch Daniels has described public sector workers as “a new privileged class in America,”\textsuperscript{52} while former Minnesota Governor Tim Pawlenty stated: “It used to be that public employees were underpaid and over-benefited. Now they are over-benefited and overpaid compared to their private-sector counterparts.”\textsuperscript{53} It is unclear, however, whether this is due to gains by public employees or losses in the private sector, where defined benefit pension plans have virtually disappeared along with many high-paying jobs.

As should be expected, it is also not clear whether the apparent compensation disparity between public and private sector employees is real. Views on this seem to fall along similar political fault lines as to whether the funding crisis is real or imagined. Some studies dispute the disparity theory by claiming that higher pay for government workers is attributable to age, education, and skill level required for the jobs.\textsuperscript{54} When one


\textsuperscript{54} See Sylvia A. Allegretto & Jeffrey Keefe, Berkeley Ctr. on Wages and Empl Dynam., The Truth about Public Employees in California: They Are Neither Overpaid nor Overcompensated 3 (2010) (“A re-estimated regression equation of total compensation (which includes wages and benefits) demonstrates that there is no significant difference in total compensation between full-time state and local employees and private-sector employees.”) (emphasis omitted); Keith A. Bender & John S. Heywood, Nat’l Inst. on Ret. Sec., Out of Balance? Comparing Public and Private Sector Compensation...
accounts for these and similar traits, it is argued that public-sector workers are undercompensated relative to their private-sector counterparts.\footnote{Keefe, \textit{supra} note 54, at 11–12.} One 2010 study, by the Center for Economic and Policy Research, found a 4\% wage “penalty” for public sector workers, taking into account wages and benefits, and controlling for age and education.\footnote{Id. at 5. Keefe’s analysis has been attacked. \textit{See, e.g.}, CTR. FOR UNION FACTS, THE ECONOMIC POLICY INSTITUTE IS WRONG: PUBLIC EMPLOYEES ARE OVERPAID 1, 7–8, http://www.unionfacts.com/downloads/Public_Sector_Unions Brief.pdf (claiming that Keefe’s study is incorrect and public employees are overcompensated).}

There is no question that public employees as a group receive vastly higher defined benefit pension compensation than private employees because most private employers have halted the practice. Many public employees, about one in four, are not in the Social Security system, which means that their state pension is their only source of employer and government support in old age.\footnote{U.S. GOV’T ACCOUNTABILITY OFFICE, GAO-12-322, \textit{STATE AND LOCAL GOVERNMENT PENSION PLANS ECONOMIC DOWNTURN SPURS EFFORTS TO ADDRESS COSTS AND SUSTAINABILITY} 5 (2012).} It would be grossly unfair to state employees if pension reform did not take into account the fact that they do not participate in the federal Social Security system. Comparing the raw numbers between private and public employee pension payments should take Social Security into account, especially because participating employers and employees both contribute to Social Security.
One author reports that in Wisconsin, which he characterizes as the eighth most generous state in terms of income replacement, the average retired worker receives a pension equal to 57% of their preretirement salary. The full pension is paid after thirty-five years at age fifty-seven for retirees other than public safety employees. More comprehensively, a 1997 table reports average replacement rates for public employees without Social Security of about 62%, but this may be lower than the replacement rate for current retirees if reports that governments have sweetened pensions in recent years are true. This rate is more generous for most private employees receiving pensions but not to such a great extent when Social Security payments are included in the comparison.

As in many situations, the view that public-employee pensions are excessive is supported by notorious instances of what is known as pension “spiking,” in which employees take advantage of provisions in pension plans that allow them to increase their pension benefits, often as they prepare to retire. Public employee pensions are usually based on the employee’s pay at the end of the career, often the average of the employee’s last three or five years of government employment. Employees make efforts to increase their pay at the end of their careers to “spike” their pensions. Even if the methods employees use to spike their pensions are within the rules of the pension system, they seem illegitimate for the simple reason that pensions manipulated in this manner are not related to the employee’s needs and legitimate expectations after retirement.

Here are a few examples of pension spiking. One way that pensions can be spiked is to add additional part-time work during the years when salary is used to calculate pension benefits. For example, in some jurisdictions, public high school teachers can teach evening courses at a community college and then count that...
pay in total salary for pension purposes. This apparently common practice among teachers in some areas can boost pension benefits significantly. In Massachusetts (and perhaps in other states), longevity clauses are included in public employees’ collective bargaining agreements.61 The employee informs the employer either one or three years in advance that they plan to retire and under the agreement, their salary is boosted in recognition of their longevity. This also boosts their pension, which is the design of the contract. If the employee changes her mind and decides not to retire, she can simply pay the bonus back to the governmental unit. The amount and length of the bonus (usually either one or three years) is determined in unionized sectors in collective bargaining between the employee union and the governmental unit.

Another legally sanctioned form of pension spiking involves pension “buybacks” for various forms of service outside the pension system. Under a buyback program, an employee is allowed to pay a year’s contribution to the system to purchase a year of service credit toward a state pension. Employee contributions are not sufficient to cover the increased costs to the pension system, so these buybacks are a good deal for the employee but not for taxpayers who will be required to make up the shortfall sometime in the future. For example, in 2002 Massachusetts enacted a provision allowing public school teachers to buy pension credit for years in the Peace Corps.62 Several other bills were proposed in the following years to expand buybacks, in the midst of efforts to eliminate abuses such as counting volunteer service on government boards toward pension service; one day to one year of service provisions (which were used by outgoing legislators to receive an entire year of service credit for the first week of January when their terms expired); and king for a day provisions, which allowed employees to be


62. See MASS. GEN. LAWS ch. 32 § 4(1)(f); see also KEN ARDON, PIONEER INST. FOR PUB. POLICY RESEARCH, PUBLIC PENSIONS: UNFAIR TO STATE EMPLOYEES, UNFAIR TO TAXPAYERS 10–13 (2006) (detailing buyback and similar provisions in the public pension system in Massachusetts).
promoted for one day and then retire at a higher rate. For example, school nurses sought to be allowed to buy pension credit for years in nursing before they entered a school system, and higher education teachers sought to be included in the Peace Corps buyback provision.

One of the most striking examples of legislative largesse in the pension area happened in Rhode Island in the 1980s. Rhode Island public school teachers had been covered by state pensions since 1936. As is generally true of public school teachers in the United States, Rhode Island public school teachers are highly unionized. In the 1980s, they lobbied for inclusion of their union’s employees in the state pension plan despite the fact that they were not government employees. In 1987, the Rhode Island General Assembly obliged, and union employees were allowed to join the teachers’ pension plan, conditioned on payments to buy years of creditable service. As a court later detailed:

Bernard Singleton, for example, became a member of the Retirement System effective January 1, 1990 . . . and promptly purchased roughly 25 years of service credit for his prior union employment at a cost of $25,411.09. On July 28, 1990, several months later, at age 52, he took “early retirement” and immediately began to collect a pension of approximately $53,000 per year, with an expected lifetime benefit of about $750,000.

63. A bill eliminating some of these abusive practices was passed and signed in 2009. See S. 2079, 186th Gen. Court, Reg. Sess. (Mass. 2009). This law eliminated pension credit for volunteer service and the one-day rule, under which one day of work counted for pension purposes as a full year of service, and it prohibited the practice of combining work from multiple government jobs to receive a higher pension. Id.; see also Michael Levenson, Key Measures Passed in Mass., Bos. Globe, Aug. 17, 2010, at 1 (discussing legislation passed in the previous two years).


65. See Ardon, supra note 62, at 12.


67. Id.

68. Id.

69. Id.
The return on investment for these participants was beyond even Bernard Madoff’s wildest dreams. “The district court later calculated the plaintiffs’ total contribution to the Retirement System at $1,995,784, the present value of their projected pension benefits at about $11,430,579, and an average projected rate of return for the individual plaintiffs of approximately 1250 percent.”70 Once the details of this plan became generally known, the Rhode Island General Assembly repealed it and provided that no further benefits would be paid.71 The United States Court of Appeals for the First Circuit upheld this repeal against attacks based on federal constitutional rights to continued benefits.72

There are notorious individual instances of pension spiking under which employees have boosted their pensions in ways that seem illegitimate. The most famous example in Massachusetts is William Bulger, who retired after thirty-five years in Massachusetts government, including seventeen years as President of the State Senate and seven years as President of the University of Massachusetts.73 His retirement salary was approximately $300,000, entitling him to a lifetime pension of $179,000.74 In the last few years of his service as University President, the Board of Trustees added a housing allowance to his compensation, even though Bulger was living at his longtime home that he owned.75 When Bulger retired (under pressure over his relationship with his then-fugitive brother Whitey Bulger), he included the housing allowance as part of his salary for pension purposes, and the Massachusetts Supreme Judicial Court agreed, boosting the pension to $196,000 annually.76

70. Id. at 24–25.
71. Id. Participants were given a refund of their contributions in excess of the amount they had already received in benefits.
72. Id.
75. See Bulger, 856 N.E.2d at 805 (“The trustees were fully aware that Bulger would continue to live in his home in the South Boston section of Boston throughout his tenure as president of the university.”).
76. See id. at 801.
Another Massachusetts example of pension spiking, which provoked the above-mentioned reform efforts, involved a public school teacher who added almost $5,500 per year to her $26,000 pension by including years of volunteer service on the board of her city’s public library.\(^77\) The fact that she counted two years during which she failed to attend a single library board meeting made her case look even weaker than it would have had she been a dedicated volunteer board member.\(^78\) One state representative\(^79\) who spoke out in favor of closing this method of pension spiking later included unpaid service on a local school board as part of his pension-eligible service, provoking cries of hypocrisy in a newspaper editorial.\(^80\) Finally, also in Massachusetts, is the example of an employee working two full-time government jobs and claiming two separate full pensions.\(^81\)

77. See Sean P. Murphy, Ex-Lawmaker’s Wife Got Pension Boost: Credit Given for Lynn Library Job, **BOS. GLOBE**, Apr. 19, 2009, at B1. The article also reports that the teacher’s ex-legislator husband also benefitted from generous pension provisions apparently designed just for him by the Massachusetts legislative leadership. Id. “The carefully tailored provision, which did not mention Bassett by name, permitted him to collect his $41,000-a-year state pension even while working full time as the Essex Regional Retirement Board chairman and executive director, a job that currently pays him an estimated $123,000 a year.” Id. Ex-representative Bassett was fined $10,000 for engaging in private lobbying activity on government time using government facilities. See Paul Leighton, **Bassett Fined $10,000**, **SALEM NEWS** (Oct. 21, 2011), http://www.salemnews.com/local/x2117288138/Bassett-fined-10-000/print (last visited Feb. 2, 2013) (describing Bassett’s illegal activities and penalties) (on file with the Washington and Lee Law Review). He had been fired the prior year for deficient performance “after years of controversy over his high salary, lavish expense accounts, and exorbitant legal and consultant fees.” Id. The pensions of both Bassetts apparently were boosted by legislative action crafted exclusively for them at both the city and state levels. Id.

78. See Sean P. Murphy, Former Essex Retirement Chief Fined, **BOS. GLOBE**, Oct. 21, 2011, at M1 (detailing the facts surrounding the fine of Timothy Bassett).

79. See Edward Mason, Pol OK’d Pension Reform, but then Tried to Cash In, **BOS. HERALD**, Sept. 30, 2009, at 6 (detailing the acts of a representative who wanted credit for his years of service as an unpaid school committee member days before an act banning such credit went into effect). This particular state representative had been in the news for an “arrest in 2004 for drunken driving, gross lewdness and disorderly conduct, and his $17,000 fine in 2007 and $10,000 in 2004 for violating Massachusetts campaign finance law.” Id.

80. See Editorial, Poster Boy for Reform, **BOS. HERALD**, Oct. 1, 2009, at 22 (describing the “unmitigated gall” of the representative as “breathtaking”).

81. See Matt Carroll, Ex-Officer Is Cleared on Fraud Charges, **BOS. GLOBE**, June 15, 2007, at B1 (noting that the employee was collecting a $139,787
Even if pensions to public officials are generally not abusive, examples of abusive practices like those discussed above taint the entire system.

C. Why?

Assuming that there is a funding crisis and that public sector employees have been promised generous, and perhaps excessive and potentially abusive retirement benefits, including pensions and health care, the final question for this part of the discussion is why did this happen. Why would politicians make such promises and underfund them?

To a certain extent, the pathology is typical of deficit spending by government. Incumbents can gain political support by enacting programs favored by constituents without requiring taxpayers to currently pay the full cost of the programs. Taxes can remain low even as services expand. Taxpayers are happy to enjoy the value of current services and reelect politicians that provide them.

Deficit spending is not unambiguously bad. During poor economic times, its use as economic stimulus may help cushion the effects of recession and even spur economic growth. Too often, however, deficit spending seems to be intended more for political stimulus than economic stimulus. After record surpluses at the end of the Clinton administration, tax cuts and increased spending under George W. Bush put the federal budget in deficit, which has continued and been amplified during the Obama administration. Although the argument in favor of tax cuts is that they increase economic activity which leads to more tax

---

pension, based on his average pay for the last three years of his working career, which was the highest in Plymouth County history).

Sullivan [the prosecuting U.S. Attorney], in his report, said the taxpayers of Plymouth County should find “the [employee’s] pension situation to be incredibly offensive,” noting that [the employee] worked only three years for the county but will be paid about $60,000 a year by [county] taxpayers for the rest of his life.

Id.

revenue, it appears that tax increases during the Clinton years contributed to surpluses then, and tax cuts at the outset of the administration of George W. Bush contributed to deficits in every budget he signed. Deficit spending appears to be a powerful political stimulus.

Unfunded pension promises benefit politicians in two ways. First, as in all deficit spending, they allow for current officials to provide services without requiring taxpayers to pay for them until much later, when they may be out of office. Second, pension promises help politicians shore up support among government workers, or at least avoid opposition from government workers, which would be substantial if significant reductions in pension benefits were proposed.

Taxpayers go along with underfunding for several simple reasons. First, each taxpayer’s share of the overall liability is likely to be relatively small, or at least appear to be small, at the time the promises are made. The psychological tendency to discount long-term problems likely reinforces the impression of each taxpayer that the unfunded liability is not a problem for them. Second, information on the extent of unfunded liabilities is not readily available and what information there is may be difficult to interpret. Taxpayers may simply not know that public employees have been promised overly generous pensions or that tax revenues are insufficient to fund them. This problem is aggravated by the use of overly optimistic projected rates of return on pension fund investments, which help obfuscate the financial status of the funds. Third, some taxpayers may conclude that they are unlikely to be affected by the whole mess at the time the obligations come due. Taxpayers move, retire, and die, all of which would minimize or exclude them from the negative effects future taxpayers may suffer due to unfunded pension liabilities.


84. See id. at 691 (noting that lawmakers with a balanced budget requirement can run a de facto deficit when they underfund government workers' pensions).

85. See Robert P. Inman, Public Employee Pensions and the Local Labor Budget, 19 J. PUB. ECON. 49, 50 (1982) (arguing that mobile taxpayers are likely to support deferring payment for current services until later at the expense of
Excessive or abusive pension promises also occur due to the nature of the relationship between government employees, elected officials, and policymakers’ self-interest. Government employees are often among the most ardent supporters of incumbent politicians because such employees depend on politicians for their jobs, levels of pay, and working conditions. In the age of patronage, the relationship between employees and elected officials was quite direct because virtually all government workers owed their jobs to some sort of connection to an elected official. But even in this era in which civil service is the dominant government employment system, patronage still exists at high levels and in various pockets of government. Further, even if only a small percentage of employees are in a close relationship with elected officials, whatever system of pay and benefits is created will normally be designed to cover everyone. In other words, the desire to be generous to “connected” employees contributes to excessive compensation for all employees. Finally, in some situations, officials have the power to shape policies governing their own pensions, which can also result in generous promises that include themselves and other public employees.

In the pension area, the effects of close relationships between politicians and employees can be quite direct. For example, in the case discussed above involving the employee in Plymouth County, Massachusetts, who worked two full-time jobs and claimed two


86. See Skeel, supra note 83, at 691, 711.

87. In Massachusetts, a scandal over patronage hiring at the state probation department has led to federal indictments of several officials including the former head of the department. It has been reported that federal prosecutors are investigating whether state legislators who “recommended” candidates to probation department jobs violated federal law in the process. See Andreas Estes & Thomas Farragher, *Ex-Probation Chief, 2 Aides Indicted in Hiring Scandal: Accused of Rigging Selection Process for Job Applicants*, Bos. GLOBE, Mar. 24, 2012, at A1 (discussing an investigation of the hiring practices of the probation commissioner and others); Andreas Estes & Scott Allen, *DiMasi Facing a Cancer Diagnosis; Ex-Speaker’s Illness Likely to be Treated at Prison Medical Center*, Bos. GLOBE, May 19, 2012, at A1, A12 (suggesting that the investigation includes looking into whether state legislators violated federal law).
separate pensions, one of his employers, an elected county sheriff, sat on the retirement board that approved one of the pensions.\textsuperscript{88} The employee had helped the sheriff’s election campaign.\textsuperscript{89} There are thousands of similar relationships throughout state and local government that undoubtedly influence compensation decisions.\textsuperscript{90} In short, before the recent spotlight that has shined on the pension issue, from one perspective, the entire system may have operated like an enormous conspiracy to capture as much of the taxpayers’ money for retired workers as possible.

We now have two general ways of understanding why government employees might be overcompensated and why an important part of that compensation takes the form of unfunded pension obligations. There are also particulars concerning how unfunded pension promises developed that can illuminate this problem. Political scientists and economists began looking at this issue as long ago as the 1970s. One early view was that as government employment became more professionalized and wages increased, deferred compensation in the form of pensions became very attractive at the same time that taxpayers demanded increased services without really wanting to pay for them.\textsuperscript{91} It also appears that at certain times public employee unions placed a higher priority on current wages than on

\begin{itemize}
\item \textsuperscript{88} See generally Carroll, supra note 81.
\item \textsuperscript{89} See Steve Bailey, \textit{Putting a Face on the Need to Reform}, \textit{Bos. Globe}, June 7, 2006, at D1

The pension system is the way it is because those who oversee it[,] the cops and firefighters who run the retirement boards[,] have it just the way they like it. As the inspector general notes, Lincoln was no accident. Former Plymouth County Sheriff Joseph McDonough, who hired Lincoln for this three-year victory lap at the jail, knew how the system worked. He is on the Plymouth County retirement board. Lincoln, not coincidentally, helped on McDonough’s campaign in 2000.

\item \textsuperscript{90} Another good illustration is the ability of the state teachers’ union in Rhode Island to convince the legislature to allow employees of the union to buy into the state pension system, resulting in a 1,250% return on investment. See Nat’l Educ. Ass’n-R.I. \textit{ex rel. Scigulinsky v. Ret. Bd. of R.I. Emps.’ Ret. Sys.}, 172 F.3d 22, 24–25 (1st Cir. 1999).

\item \textsuperscript{91} See Epple & Schipper, supra note 13, at 170. Interestingly, Epple and Schipper suggest that public pension underfunding should decrease as the school-aged population of baby boomers declines. \textit{Id.}\
\end{itemize}
adequate funding of pension promises, even if this created some risk of nonpayment in the future.92

Two additional historical factors have contributed to the problem of pension funding. One factor is that in good economic times, governments have tended to increase all forms of employee compensation, including pension promises.93 Assuming a general level of underfunding, a higher overall payroll is likely to produce a higher level of underfunding. Another factor is that in tight fiscal times, governments have foregone or reduced pension contributions and used the money to fund other services.94 This is not surprising because constituents’ demand for services may actually increase in periods when funds are tight due to economic downturn. State balanced budget requirements may contribute to this aspect of the problem: Because borrowing to meet operating expenses may not be available, underfunding pension obligations becomes a necessary tool to balance the budget without making drastic cuts to services.95 These two dynamics, increased promises in boom times coupled with decreased funding in tough times, are a recipe for fiscal disaster.

In sum, unfunded pension and health care promises to retirees are, in a sense, the state and local version of the federal deficit. Politicians have twin incentives at work: To defer payment for current services to future generations of taxpayers and to reward loyal supporters in the ranks of government workers with handsome compensation packages, including generous retirement benefits. Even if most government workers are of little concern to politicians, the desire to reward the

---

92. See Olivia S. Mitchell & Robert S. Smith, Pension Funding in the Public Sector, 76 Rev. Econ. & Stat. 278, 282–83 (1994) (testing data on how public pensions tend to be funded over time, allowing for several factors). Public employee unions have challenged underfunding as violating their contractual or constitutional rights, apparently out of concern that if the system is underfunded, their pensions might not be paid in full.


95. See generally id.
connected few (and often themselves) contributes to the phenomenon of all boats rising together. Even legislators themselves may need to establish an attractive pension system for all government workers to justify their own generous postservice compensation. Taxpayers may now be waking up, but as we shall see in the discussion of legal constraints on pension reform, it may be too late to avoid severe fiscal hardship.

III. State Law Constraints on Underfunding Pension Liabilities and Pension Reform

Recent and continuing fiscal difficulties in many state and local government entities have inspired searches for ways to save money. Pensions are an obvious candidate, but even if state legislatures were determined to reduce pension promises, state contract law and state constitutional law designed to protect the legitimate expectations of state and local employees may stand in the way. In this part of the Article, I look at three state law issues concerning pension reform: The effects of state balanced budget requirements on pension funding, state law constraints on underfunding pension contributions, and state contract and constitutional law constraints on reducing pension benefits or promises to workers not yet retired. As we shall see, state law can pose significant impediments to pension reform.

A. State Balanced Budget Requirements and Pension Plan Funding

In debates over fiscal policy, the fact that balanced budget requirements exist in nearly every state is held up as evidence...
that the federal government could and should follow suit and balance its budget. This has been a cornerstone of the Tea Party movement, and during 2011's controversy over increasing the federal government's debt limit, there was a proposal to condition the extension on Congress voting for a balanced budget amendment to the federal Constitution.\textsuperscript{97} As we have seen, however, the magnitude of unfunded state pension and health care promises shows that states are not nearly as constrained as might appear from the existence of balanced budget requirements. This raises questions of whether the failure to fund pension obligations constitutes unlawful deficit spending, and whether such a violation would justify renunciation of some portion of unfunded obligations.

The simple answer is that state failure to fund pension liabilities is not considered a violation of state balanced budget requirements. Further, in some states, competing constitutional requirements prohibiting diminution of pension promises mean that the weight of state constitutional law is more strongly on the side of what is, in effect, deficit spending, than it is on the side of fiscal constraint.

The first thing to understand about state balanced budget requirements is that they are quite diverse and impose varying levels of fiscal discipline. One important fact is that state balanced budget requirements normally affect only state operating budgets, not capital or long-term debt obligations.\textsuperscript{98} government and that the evidence is inconclusive on whether there is a difference in effects between constitutional and statutory balanced budget requirements. \textit{Id.} at 78; see also James M. Poterba, \textit{Balanced Budget Rules and Fiscal Policy: Evidence from the States}, \textit{48 Nat'l Tax J.} 329, 329–34 (1995) (considering state balanced budget requirements and the possibility of a federal balanced budget law).

\textsuperscript{97} See Alan Fram, \textit{Balanced Budget Amendment Injected Into Debt Ceiling Fight}, \textsc{Huffington Post} (July 14, 2011, 6:24 PM), http://www.huffingtonpost.com/2011/07/14/balanced-budget-amendment_n_899301.html (last visited Feb. 2, 2013) (describing efforts to include a Balanced Budget Amendment as part of a deal to raise the debt ceiling) (on file with the Washington and Lee Law Review).

\textsuperscript{98} See \textsc{Nat'l Conference of State Legislatures}, \textit{supra} note 2, at 6 (noting that “[s]tate budget processes focus on balancing state operating budgets with less emphasis on balancing the rest of the state budget”). For a general look at state balanced budget requirements, see U.S. \textsc{Gen. Accounting Office}, GAO-93-58, \textit{Balanced Budget Requirements: State Experiences and Implications for the Federal Government} (1993).
This means that states are free to finance capital projects with long-term debt,99 which is sensible fiscal policy because current taxpayers might be unwilling to fully finance projects with long-term benefits. Interest payments on long-term debt would presumably be included in the operating budget, which must be balanced each year, but there is no prohibition on incurring long-term debt. However, state constitutions often contain stringent limits on the use of debt financing.100 Thus, the exclusion of long-term debt from balanced budget requirements does not necessarily release states from the fiscal constraints under which they would otherwise operate.

The Association of State Budget Officers reports that state balanced budget requirements generally take three forms, with many states operating under two or even all three of the requirements: (1) The governor’s proposed budget must be balanced; (2) The enacted budget must be balanced; and (3) No deficit can be carried forward from one fiscal period into the next.101 Further, some states require that the governor sign a balanced budget.102 State constitutions and statutes do not always explicitly require these steps, but some courts have read them to exist.103

State balanced budget provisions also vary in the availability of enforcement mechanisms.104 In a very few states, mandatory

99. See Nat’l Conference of State Legislatures, supra note 2, at 7 (indicating that state governments generally do not consider debt obligations for capital expenditures to violate a balanced budget “either because those provisions specify a way that general obligation debt may be issued, or because . . . judicial decisions have validated the issuance of other forms of debt”).

100. See id. at 8–9 (indicating that states often place constitutional requirements on budget balancing and that some states have specific constitutional requirements for debt financing).


102. See Nat’l Conference of State Legislatures, supra note 2, at 5 (discussing North Dakota as an example of a state that requires the governor to sign a balanced budget).

103. See id. at 9–10 (noting that some requirements for state budget balancing have emerged from judicial decisions predicated upon constitutional provisions that have little to do with budgetary matters).

104. See id. at 8–9 (discussing the various enforcement mechanisms that state balanced budget provisions contain).
spending reductions are required if expenditures would otherwise exceed revenue. At least one state provides for criminal punishment of officials who authorize deficit spending. In other states, governors monitor expenditures and are required to make cuts during the fiscal year to ensure that the budget remains in balance. Some states may also simply prohibit the paying of bills if funds have run out. Some states are more liberal, allowing borrowing at the end of the fiscal year to satisfy outstanding obligations. The overriding factor may be the political culture of state government. Even in states with uncertain enforcement, operating budgets remain balanced because the political costs of running an illegal deficit would simply be too high.

Ironically, state balanced budget requirements are negatively correlated with pension funding to full actuarial standards. In other words, states with strict balanced budget requirements are less likely than other states to fully fund their projected future pension obligations. The reason for this may be simple: When balanced budget requirements are likely to be strictly enforced, expenditures are moved to areas that do not constitute deficit spending. Because pension promises are an off-budget method of providing compensation to state employees for current services, the larger the share that can be paid in the form of deferred compensation, the more services government can provide out of

105. See id. at 9 (noting that Alabama and Oklahoma “require mandatory reductions in expenditures to keep budgets in balance”).

106. See id. (“The state constitution [in Alabama] allows claims against appropriations to become void at the end of the fiscal year if the treasury lacks money to pay them. A treasurer who violates this provision is subject to a $5,000 fine, two years’ imprisonment in the state penitentiary, or both . . . .”).

107. See id. (indicating that “[a] substantial number of states allow or require governors to reduce state spending when it is likely to exceed available resources”).

108. See id. (noting that in Alabama the state constitution permits the voiding of claims against appropriations if the treasury lacks the funds to pay them).

109. See id. at 7–8 (describing states that allow borrowing from one fiscal period to the next and the possible consequences of prolonged borrowing).

110. See id. (noting that the most important factor for most states’ budget balancing is that a tradition of budget balancing has created intense political pressure to continue balancing the budget).

111. See Chaney et al., supra note 94, at 307.
current revenue. Further, in tight fiscal times, the tendency for state governments to reduce or suspend pension funding for one or more years\textsuperscript{112} to avoid serious cuts to current services can aggravate pension fund deficits during bad economic times when stock market downturns reduce pension fund investment values and state tax revenue declines.\textsuperscript{113}

The relative freedom of states to determine their own discount rates also contributes to the general underfunding of pension obligations. States can tinker with pension growth forecasts and discount rates to make it appear that they are funding future obligations adequately or creating only a relatively small funding gap when they decrease their contributions to bridge budget gaps.\textsuperscript{114} These temporary budget fixes contribute to cumulative problems because later budgets do not make up for the earlier gap in funding. States may also issue pension obligation bonds to meet required annual contribution requirements, but this move passes the cost on to future generations of taxpayers who must pay the bonds and may also need additional funds to make up for underfunding due to inflated discount rates.\textsuperscript{115} Thus, the short-term nature of state budgeting and the inapplicability of “balanced budget”

\textsuperscript{112} See \textsc{Report of the State Budget Crisis Task Force, supra} note 22, at 37–38. The report contains a detailed discussion of state and local underpayment of projected pension liabilities and reform efforts that may make it more difficult in some states for government entities to continue underpaying. This, in turn, would lead to more stress on already tight state and local budgets. \textit{See id.} at 40–41.

\textsuperscript{113} See \textsc{Thad Calabrese, Soc’y of Actuaries, Public Pensions, Public Budgets, and the Risks of Pension Obligation Bonds} 3 (2010) (discussing how pension fund deficits grow during times of economic downturn).

\textsuperscript{114} \textit{See id.} at 4–8 (discussing the way that states can maintain the appearance of adequately meeting pension funding obligations without actually doing so); \textit{see also Josh Barro & Stuart Buck, Manhattan Inst., Underfunded Teacher Pension Plans: It’s Worse Than You Think} (2010), http://www.manhattan-institute.org/pdf/cr_61.pdf (noting that states have more leeway than private entities to alter their discount rates because they generally follow the pension standards set by the Governmental Accounting Standards Board rather than the market-based standards established by the Financial Accounting Standards Board).

\textsuperscript{115} See \textsc{Calabrese, supra} note 113, at 7–11 (discussing the intricacies of how issuing pension obligation bonds to meet annual contribution requirements passes the cost to future taxpayers).
requirements conspire to create a long-term mess of underfunded pension obligations.

This should be discouraging to those who champion balanced budget requirements as devices to bring fiscal constraint to government. Underfunding future pension obligations shares many of the vices of deficit spending and is different from long-term borrowing for capital projects because pension promises are more like operating expenses than capital borrowing. While deficit spending may make sense when economic stimulus is desired, for programs that do not promise to grow the economy for the future, it is a simple intergenerational wealth transfer, with current taxpayers pushing off the expense of providing current government services to future taxpayers. For the most part, pension promises fall into this category. Generous, secure pension promises allow government employers to pay their employees less in current cash compensation. Underfunding pension obligations means that future taxpayers will essentially pay the bill for services provided in the past without any current benefit, such as a building, park, or highway, which is still being used while bond payments are made. An effective state balanced budget requirement would thus include advance funding (under realistic projections and discount rates) of pension and retiree health care promises to public employees as part of the current operating expenses required, under state law, to be part of a balanced budget.

B. State Law Limitations on Pension Reform

In many states, the weight of constitutional law is with state employees rather than the taxpaying public. In a

116. Although state statutory and constitutional protections of public pensions are distinct from federal law, except in states with very specific constitutional protections for pension promises, the considerations state judges use to decide whether to protect pensions under state law are very similar to the considerations they use to determine whether a reform violates the federal Contracts Clause. Generally, once a state court finds that an employee has a contractual right to a feature of a pension plan, the court finds a violation of either state pension provisions or federal constitutional law.

117. For a general discussion of the legal status of public pension reform, see David L. Gregory, The Problematic Status of Employee Compensation and Retiree Pension Security: Resisting the State, Reforming the Corporation, 5 B.U.
comprehensive review of state pension plan protections, Amy Monahan has demonstrated that many states protect pension plan participants from significant modifications to their plans under both constitutional and contract law theories.118 In another article, Monahan reports that “courts in California and the twelve other states that have adopted California’s precedent have held not only that state retirement statutes create contracts, but that they do so as of the first day of employment.”119 Jonathan Forman concludes that state law places serious constraints on pension reform with regard to existing workers: “Through state constitutional provisions and court interpretations of property and contract rights, most states essentially guarantee that their public workers will get the pensions that they were promised when they were hired.”120

Some state constitutions contain provisions that explicitly prohibit the state from reducing pension payments or pension promises to state employees. For example, the New York constitution provides that “[a]fter July first, nineteen hundred forty, membership in any pension or retirement system of the state or of a civil division thereof shall be a contractual relationship, the benefits of which shall not be diminished or impaired.”121 This has been interpreted to protect the level of benefits promised as of the date that the employee became eligible to participate in the pension plan.122 The Illinois constitution contains a very similar provision, which has been

---


119. Amy B. Monahan, Statutes as Contracts? The “California Rule” and its Impact on Public Pension Reform, 97 IOWA L. REV. 1029, 1032 (2012) [hereinafter Monahan, Statutes as Contracts]. Monahan is highly critical of this line of cases, finding it to be inconsistent with more general legal principles concerning flexibility in government regulatory programs. For further discussion, see infra notes 212–13 and accompanying text.

120. Forman, supra note 39, at 866.

121. N.Y. CONST. art. V, § 7.

interpreted to preclude the Illinois legislature from unilaterally cutting pension benefits to current employees. Similarly, the Michigan constitution provides that “[t]he accrued financial benefits of each pension plan and retirement system of the state and its political subdivisions shall be a contractual obligation thereof which shall not be diminished or impaired thereby.”

States with provisions like these may be unable to reduce pension payments or promises to state workers even if the magnitude and nature of pension promises is in serious tension with state balanced budget requirements. It should also be noted that many state courts use the federal Contract Clause to protect pension promises, finding first a contractual relationship under state law, and then protecting employee rights under federal constitutional law.125

Most states recognize that public pension rights vest at some point, after which the state is precluded from amending the contractual promises. The most common point at which rights are solidified under state law is when an employee satisfies the requirements for grant of the pension, commonly referred to as “vesting,” which usually occurs at some point after the onset of employment and before retirement. Some states’ laws are even


125. See, e.g., Or. State Police Officers’ Ass’n v. State, 918 P.2d 765, 768 (Or. 1996) (holding that state constitutional amendments altering employee contribution amounts, prohibiting guaranteed rates of return on pension funds, and prohibiting inclusion of unused sick leave in pension calculations violated Contract Clause rights of employees); Singer v. City of Topeka, 607 P.2d 467, 477 (Kan. 1980) (holding that changes in retirement benefits promises violated contractually protected rights and therefore violated the federal Contract Clause).

126. Many decisions recognize vested rights in dicta while denying claims brought by employees who sue over pension reform before they are actually eligible to retire. See, e.g., Petras v. State Bd. of Pension Trustees, 464 A.2d 894, 896 (Del. 1983) (noting that while “vested contractual rights were held by those employees and former employees who satisfied the eligibility requirements for a
more favorable toward employees, recognizing pension rights from the onset of government employment. Courts in these states reason that "by accepting the job and continuing work, the employee has accepted the State's offer of retirement benefits, and the State may not impair or abrogate that contract without offering consideration and obtaining the consent of the employee."127

Some states take a reliance interest approach to the question of whether an employee has a vested right to pension benefits that is protected under constitutional or contractual principles.128 For example, the West Virginia Supreme Court has reasoned:


128. For an argument that reliance should be the key issue in Contract Clause jurisprudence, see Robert A. Graham, Note, The Constitution, The Legislature, and Unfair Surprise: Toward A Reliance-Based Approach to the
When considering the constitutionality of legislative amendments to pension plans, an employee’s eligibility for a pension does not determine whether he or she has vested contract rights. Instead, the determination of an employee’s vested contract rights concerns whether the employee has sufficient years of service in the system that he or she can be considered to have relied substantially to his or her detriment on the existing pension benefits and contribution schedules.129

With sufficient length of service, reliance is presumed,130 but only on those provisions that are in effect during the lengthy service. This approach to determining whether the state may alter pension benefits requires that the court determine in each case whether the employee has relied on the particular provision that has been altered, especially with regard to provisions that were not in effect during the entire period of employment. For example, in 1988, the West Virginia legislature amended that state’s public pension statute to include lump-sum payments for unused vacation time in retiring employees’ final salary for pension calculations.131 Apparently many employees took early retirement shortly after the amendment passed so they could take advantage of this method of increasing their pension payments.132 Then, in 1989, the legislature repealed the provision.133 When one employee retired in 1996, he sought to have a lump-sum payment for his unused vacation time included in his final salary for pension purposes even though the provision allowing this had been repealed in 1989.134 The trial court dismissed the employee’s claim, but the West Virginia Supreme Court ruled that he was entitled to an opportunity to prove his allegation that in 1988 he made a decision to continue his employment with the State in reliance upon the 1988 version of the retirement statute, expected that he would be able to add his...

---

130. See id. at 184 (concluding that “after 10 years of state service detrimental reliance is presumed”).
132. Id.
133. Id.
accrued but unpaid leave to his final average salary when he retired, and would thereby receive an increased monthly retirement benefit.\footnote{Id. at 201.}

Without specific evidence of reliance, the particular pension benefit would not be vested and the state would be able to eliminate or modify it. For example, when two employees retired in the late 1990s, they also sought to have lump-sum payments for unused vacation time included in their final salaries on the basis that they relied on the 1988 provision by remaining employed by the state for ten years after the 1988 amendment was adopted.\footnote{Myers, 704 S.E.2d at 743–44.} The West Virginia Supreme Court denied the claim, concluding that reliance on a provision that was in effect for only one year cannot be presumed, and

\begin{quote}

neither [plaintiff] presented any specific evidence indicating that they relied to their detriment on this specific provision . . . . [N]either of the Appellees in this case was eligible to retire during the year this benefit was in effect and, thus, . . . neither of the Appellees could have based any retirement decision on the promise contained in the 1988 amendment. Indeed, neither Appellee introduced any evidence to show that he made any decision whatsoever on the basis of that particular promised benefit.\footnote{Id. at 750–51.}
\end{quote}

Although, as Monahan reports, California protects pension promises from the first day of employment,\footnote{Monahan, Statutes as Contracts, supra note 119, at 1032.} some California decisions take a nuanced view of reliance, balancing employees’ interest in pension benefits against the state’s need for flexibility and control.\footnote{See id. at 1058 (discussing the California decisions that balanced employees’ pension rights against the state need for flexibility and control and held it permissible “to eliminate future benefit accruals once a minimum pension had been earned”).} The California Supreme Court has stated that “[t]he employee does not obtain, prior to retirement, any absolute right to fixed or specific benefits, but only to a ‘substantial or reasonable pension.’”\footnote{Betts v. Bd. of Admin., 582 P.2d 614, 617 (Cal. 1968) (quoting Wallace v. City of Fresno, 265 P.2d 884, 886 (Cal. 1954)).} It is unclear, however, how far this apparent flexibility goes because California cases also state that
normally any reduction in pension benefits must be compensated for by other aspects of the reform provisions.\textsuperscript{141} Further, there are California cases that appear to mechanically enforce provisions of pension laws in effect during employment, even when the results may be seen as abusive double increases in benefits.\textsuperscript{142}

In addition to the contract-based protections employees enjoy, labor law may provide another layer of protection. State and local governments may not be able to unilaterally alter pension benefits for employees in bargaining units engaged in collective bargaining. Because retiree benefits are often specified in collective bargaining agreements, any unilateral attempt to alter them may be considered a breach of contract, no matter how weighty the government interest behind the need for reform.\textsuperscript{143} Thus, for unionized sectors, reform may depend on successful collective bargaining.

In some states, the law goes further than protecting benefit levels and also protects funding levels, requiring an actuarially adequate level of annual contributions to pension funds.\textsuperscript{144} For example, in elaborating on state statutes that create contractual guarantees in pension benefits to public employees, the North Carolina Court of Appeals stated, “[I]t is clear that Plaintiffs had a contractual right to the funding of the Retirement System in an actuarially sound manner. Therefore, we hold that the right to have the Retirement System funded in an actuarially sound manner is a term or condition included in Plaintiffs’ retirement

\textsuperscript{141} See Allen v. City of Long Beach, 287 P.2d 765, 767 (Cal. 1955) (requiring that “changes in a pension plan which result in disadvantage to employees should be accompanied by comparable new advantages”). Because of relatively strict application of this requirement, Monahan views the California decisions as much more favorable to employees than the language from Betts might imply. See Monahan, Statutes as Contracts, supra note 119, at 1062–64 (discussing cases strictly applying the requirement that changes resulting in disadvantages should also include new advantages).

\textsuperscript{142} See Betts, 582 P.2d at 619 (noting that petitioner receives a double increase in benefits and concluding that the legislature must have intended such a result for “constitutional officers serving between 1963 and 1974 because it left in effect both of the formulae during that 11-year period”).


\textsuperscript{144} For a discussion of cases involving funding levels, see Simko, supra note 82, at 1065–79.
contracts.

The North Carolina court cited decisions from several other jurisdictions for the proposition that actuarially sound funding can be a contractually protected term of a pension program.

Other states recognize that the legislature should have discretion over funding decisions and protect only the ultimate pension payments and not the funding of pension funds. For example, in Illinois, pension participants and the funds themselves challenged a statute that changed the method of calculating government contributions to pension funds. They argued that the new statute violated the Illinois Constitution's pension protection provision: “Membership in any pension or retirement system of the State, any unit of local government or school district, or any agency or instrumentality thereof, shall be an enforceable contractual relationship, the benefits of which shall not be diminished or impaired.”


148. ILL. CONST. art. XIII, § 5.

149. McNamee, 672 N.E.2d at 1163; see also People ex rel. Ill. Fed'n of Teachers v. Lindberg, 326 N.E.2d 749, 750–52 (Ill. 1975) (noting that the Governor reduced pension appropriation but that “it cannot be said that under the circumstances this constitutional provision affords plaintiffs the right to judicially circumvent the Governor's actions”). The court cited cases in which it had invalidated legislation that reduced pension benefits, but declined to follow a New York case that protected funding levels. See McNamee, 672 N.E.2d at 1165 (citing Felt v. Bd. of Trustees of Judges Ret. Sys., 481 N.E.2d 698 (Ill. 1985); Buddell v. Bd. of Trustees, State Univ. Ret. Sys. of Ill., 514 N.E.2d 184 (Ill. 1987)). The plaintiffs had urged the court to follow McDermott v. Regan, 624 N.E.2d 985 (N.Y. 1993), in which the New York Court of Appeals had invalidated a provision removing the New York comptroller's power to require actuarially adequate contributions to pension funds. See also Jones v. Bd. of Trustees of Ky. Ret. Sys., 910 S.W.2d 710, 714–16 (Ky. 1995) (noting that the
Judicial insistence on adequate funding would prevent some of the most serious missteps that have contributed to the funding crisis.\textsuperscript{150} It would reduce the tendency of states to use pension obligations as a form of deficit spending, pushing off payment for current services onto future taxpayers. Underfunding of pension funds is sometimes systematic, as when states use unrealistic projected rates of return on pension funds to justify underfunding; and sometimes it is episodic, as when states decide to cut pension contributions to balance the state budget during difficult fiscal times.\textsuperscript{151} While legitimate questions can be raised over whether the courts should prevent the government from allocating funds as it sees fit, judicial compulsion in this context may be the least of several potential evils.

It should not be surprising that the law in many states is very protective of public employees’ and retirees’ pension

state legislature had the power to amend method of calculating public employer contribution to retirement fund without unconstitutionally impairing contracts).

\textsuperscript{150} Some full funding requirements may go too far. The United States Postal Service is legally required to fund its pension and retiree health care obligations in advance. This has proven to be a hardship to the Postal Service, and due to its general downturn in business, it failed to make two payments in 2012, totaling $11.1 billion. See Ron Nixon, \textit{Postal Service Reports Loss of $15 Billion}, N.Y. TIMES, Nov. 16, 2012, at A22 (“[B]ecause of revenue losses, the post office was for the first time forced to default on these payments, which were due in August and October.”).

\textsuperscript{151} For example, the challenge in \textit{Stone v. State} resulted from an executive order issued by North Carolina’s Governor diverting pension contributions to balance the budget. See \textit{Stone v. State}, 664 S.E.2d 32, 40 (N.C. App. 2008). It appears common that in difficult fiscal times, pension contributions are reduced. From the perspective of the government employee, using underfunding as a reason for cutting benefits may appear to be manipulative. Legislators promise generous pension benefits knowing they will underfund them and be able to use the underfunding later as an excuse for reform. This conspiracy theory may be far-fetched in the amount of the advance planning it entails, but it may not seem so to the public employee suffering cuts to promised benefits. Zach Carter’s Huffington Post article accuses conservative state governors of creating the pension funding crisis to finance tax cuts and justify pension reductions to state workers. See Carter, supra note 16. Regarding New Jersey, Carter reports that

During the 1990s, under Gov. Christine Todd Whitman (R), the state slashed its annual pension contributions in order to finance a slate of tax cuts, and didn’t begin seriously boosting those contributions until 2007. . . . Last year, Gov. Chris Christie (R) took a page from Whitman’s playbook, forgoing the $3 billion annual state contribution to the pension plan while pushing $1 billion in tax cuts for the state’s wealthiest citizens.
expectations. For the most part, the employees have traditional contract principles on their side, and in the typical case, they have legitimately relied on their employers’ retirement promises. These state courts recognize that it would be grossly unfair to employees if their retirement savings were subject to the political and fiscal winds that might lead state and local legislative bodies to make significant cuts to their pensions. 152

**IV. Federal Constitutional Law Constraints on State Pension Reform**

Assuming that state law allows it, the next issue to explore concerns federal constitutional constraints on state pension reform. The primary federal constitutional provision that restrains states here is the Contract Clause, which prohibits states from passing “any . . . Law impairing the Obligation of Contracts.” 153 Additionally, the Takings Clause 154 may limit states’ ability to reduce pension payments to some state workers.

**A. The Contract Clause and Pension Reform**

There has been a good deal of litigation in both state and federal courts concerning the application of the Contract Clause

---

152. But see Alicia H. Munnell & Laura Quinby, Ctr. for Ret. Research at Bos. Coll., Legal Constraints on Changes in State and Local Pensions 3 (2012), http://crr.bc.edu/wp-content/uploads/2012/08/slp_25.pdf, for a report arguing for a sharp distinction between benefits earned for past service and benefits expected based on future service. Their main argument in favor of flexibility is that public pension benefits should be subject to the same economic considerations as private pensions. *Id.* In general, private companies can reduce pension promises prospectively—while pension promises based on past service may not be reduced, pension promises based on future service can be reduced along with other elements of future compensation. *Id.* The authors of the report recognize that in some states, this would require a constitutional amendment. *Id.* Munnell and Quinby’s treatment is more balanced than that of some analysts who do not seem to recognize the legitimate reliance interests government workers have in their pension benefits.

153. U.S. Const. art. 1, § 10.

154. *Id.* amend. V (“[N]or shall private property be taken for public use, without just compensation.”).
to state pension reform.\textsuperscript{155} It was understood from very early on that the Contract Clause applied both to state laws impairing private contracts and state laws impairing the obligation of the state’s own contracts.\textsuperscript{156} However, in the early cases, the Supreme Court did not view legislative pension promises as contractual in nature and thus refused to protect them under the Due Process Clause\textsuperscript{157} or the Contract Clause.\textsuperscript{158} In neither case, however, did the Court categorically rule out protecting the pension promises. In the later of the cases, which more closely resembles the current approach under the Contract Clause, the Court found no contractual right to pension promises based largely on decisions of the Illinois Supreme Court, which found that the legislation in question was not intended to preclude subsequent revision of the plan involved.\textsuperscript{159}

Although at one time it might have seemed that the primary focus of the Contract Clause was on state regulation of private

\begin{footnotesize}
\begin{enumerate}
\item \textsuperscript{155} Early Supreme Court decisions on this subject are not favorable to pension plan participants’ claims. In 1889, the Court characterized public pensions as gratuities that could be withdrawn at any time. Pennie v. Reis, 132 U.S. 464, 471–72 (1889). Later, the Court held that a new statute reducing payments under a prior statute to those already receiving their pensions did not violate the Contract Clause. See Dodge v. Bd. of Educ. of City of Chi., 302 U.S. 74, 81 (1937). Neither of these cases has been overruled, and in fact Dodge was cited with approval as recently as 1985. See Nat’l R.R. Passenger Corp. v. Atchison, Topeka & Santa Fe Ry. Co., 470 U.S. 451, 466 (1985). However, due to the significant changes to the law governing constitutional protection of state benefits over the last fifty years, it would be unwise to treat the issues addressed in this Article as settled by those decisions. For further discussion, see Public Employee Pensions, supra note 13, at 996.
\item \textsuperscript{156} See Fletcher v. Peck, 10 U.S. 87, 137–39 (1810) (applying Contract Clause to grant of land by the state of Georgia).
\item \textsuperscript{157} See Pennie, 132 U.S. at 471–72 (1889) (stating that the abolition of pension plans and transfers of funds deducted from employees’ paychecks to other purposes does not violate pension plan beneficiaries’ due process rights).
\item \textsuperscript{158} See Dodge, 302 U.S. at 81 (deciding that a legislative pension promise described as an “annuity” within the statute at issue does not merit protection under the Contract Clause).
\item \textsuperscript{159} See id. (basing its decision heavily on the reasoning of the Illinois Supreme Court). Note that this decision predates the provision of the 1970 Illinois constitution that protects pension benefits. See ILL. CONST., art. XIII, § 5; Felt v. Bd. of Trustees, 481 N.E.2d 698, 700 (Ill. 1985) (noting that the 1970 Illinois constitution protects pension benefits by creating a contractual relationship between public employees and the state which the state cannot impair or diminish).
\end{enumerate}
\end{footnotesize}
contracts, more recently, the Supreme Court, recognizing the potential for state and local governments to use their sovereign immunity to take advantage of contractual partners, has stated that the Contract Clause applies more strictly to states’ own contracts than to private contracts.160 The First Circuit has observed that stricter scrutiny of impairments to the state’s own contracts can be attributed to the fact that “the State’s self-interest is at stake.”161

The Contract Clause, however, is not understood today as an absolute bar on laws altering state pension obligations (and other state promises).162 Beginning in the 1930s, the Supreme Court adopted a relatively lenient view of the Contract Clause, allowing states great latitude in passing economic legislation that might have previously been viewed as impairing the obligation of contracts.163 The standard that has developed in the federal courts to decide whether pension reform violates the Contract Clause has two elements: (1) Whether the change in state law

160. See U.S. Trust Co. of N.Y. v. New Jersey, 431 U.S. 1, 25–26 (1977) (noting that unlike determining whether a state may impair a private contract, when determining whether a state may impair a state contract “complete deference to a legislative assessment of reasonableness and necessity is not appropriate because the State’s self-interest is at stake”).


162. See U.S. Trust Co. of N.Y., 431 U.S. at 25 (“The Contract Clause is not an absolute bar to subsequent modification of a State’s own financial obligations. As with laws impairing the obligations of private contracts, an impairment may be constitutional if it is reasonable and necessary to serve an important public purpose.” (citation omitted)). It may be that under the original understanding of the Contract Clause, all retrospective modifications of contractual obligations would be considered unconstitutional. See Douglas W. Kmiec & John O. McGinnis, The Contract Clause: A Return to the Original Understanding, 14 HASTINGS CONST. L.Q. 525, 526 (1987) (“Correctly interpreted, the Contract Clause prohibits all retrospective, redistributive legislation which violates vested contractual rights by transferring all or part of the benefit of the bargain from one contracting party to another.”). However, as the authors point out, the Clause is not so understood by the Supreme Court today. Id.

163. See Veix v. Sixth Ward Bldg. & Loan Ass’n of Newark, 310 U.S. 32, 41 (1940) (upholding state legislation limiting the withdrawal of bank shares against a challenge that it violated the Contract Clause); Home Bldg. & Loan Ass’n v. Blaisdell, 290 U.S. 398 (1934) (holding that certain portions of a state law providing mortgage relief through judicial proceedings did not violate the Contract Clause).
results in a “substantial impairment of a contractual relationship,”\textsuperscript{164} and if so, (2) whether this impairment is justified as “reasonable and necessary to serve an important public purpose.”\textsuperscript{165} Thus, there must be both a contractual relationship and a substantial impairment, and even when that is present, an important public purpose is sufficient to uphold the impairment.\textsuperscript{166}

The second element, allowing impairment to be justified as “reasonable and necessary to serve an important public purpose,” reads like a form of intermediate scrutiny. The state law must be more than merely rationally believed to serve a legitimate purpose, which would be the test under the lowest level of constitutional scrutiny.

The first element, whether there has been a substantial impairment of a contractual relationship, can itself be divided into three separate inquiries: First, whether a contractual relationship exists; second, whether any such relationship has been impaired; and third, whether any impairment is substantial.\textsuperscript{167}

When determining whether a protected contractual relationship exists, courts are very sensitive to states' interest in remaining flexible and retaining their full regulatory authority. This judicial instinct in the United States dates back at least to the famous \textit{Charles River Bridge} case\textsuperscript{168} in which the Supreme Court held that a company operating a toll bridge under a state charter could not prevent the state from chartering another bridge which, when its tolls expired a few years after opening,

\textsuperscript{164} Parker, 123 F.3d at 5 (internal quotation marks omitted) (quoting Gen. Motors Corp. v. Romein, 503 U.S. 181, 186 (1992)).


\textsuperscript{166} For a detailed examination of Contract Clause protection of public pensions, see Monahan, \textit{Public Pension Plan Reform}, supra note 118.


\textsuperscript{168} Proprietors of Charles River Bridge v. Proprietors of Warren Bridge, 36 U.S. 420 (1837).
would drive the first bridge out of business. In the course of determining that the Charles River Bridge operators did not have an exclusive franchise over river crossings in the area, the Court expressed concern that a contrary finding would prevent state governments from acting in the public interest. As Chief Justice Taney stated in his opinion for the Court rejecting an implied intention of the state to create a binding exclusive contract:

[S]till less will it be found, where sovereign rights are concerned, and where the interests of a whole community would be deeply affected by such an implication. It would, indeed, be a strong exertion of judicial power, acting upon its own views of what justice required, and the parties ought to have done, to raise, by a sort of judicial coercion, an implied contract . . . .

Early cases refusing to recognize vested rights in pension payments clearly rested their analysis on the need to preserve regulatory flexibility over pension payments to retired state workers. Just as Congress remains free to adjust the Social Security program by increasing the retirement age, delaying or reducing cost of living allowances, increasing payroll tax deductions, imposing income tax on benefits payments, and even reducing benefits payments, the Supreme Court has recognized state flexibility in pension terms. As the Court stated very clearly in 1985, the presumption against finding a contractual obligation in pension promises

is grounded in the elementary proposition that the principal function of a legislature is not to make contracts, but to make laws that establish the policy of the state. Indiana ex rel. Anderson v. Brand, 303 U.S. 95, 104–105 (1938). Policies, unlike contracts, are inherently subject to revision and repeal, and to construe laws as contracts when the obligation is not clearly and unequivocally expressed would be to limit drastically the essential powers of a legislative body. Indeed, “[t]he continued existence of a government would be of no great value, if by implications and presumptions, it was disarmed of the powers necessary to accomplish the ends of its creation.” Keefe v. Clark, 322 U.S. 393, 397 (1944) (quoting Charles River Bridge v. Warren Bridge, 36 U.S. (11 Pet.) 420, 548 (1837)). Thus, the party asserting the creation of a

169. Id. at 552.
170. Id. at 550.
contract must overcome this well-founded presumption, *Dodge*, *supra*, 302 U.S., at 79, and we proceed cautiously both in identifying a contract within the language of a regulatory statute and in defining the contours of any contractual obligation.\textsuperscript{171}

In light of these concerns, the courts have developed a strong, clear statement for determining whether a contractual relationship with the state exists.\textsuperscript{172} The standard in this area has been referred to as the “unmistakability doctrine,”\textsuperscript{173} requiring that the state’s intent to be contractually bound be “expressed in terms too plain to be mistaken.”\textsuperscript{174} The purposes of the unmistakability doctrine are to preserve state flexibility in the exercise of sovereign power and to avoid the difficult constitutional questions that arise if a contractual obligation is found.\textsuperscript{175} Due to the strong presumption against finding a


\textsuperscript{172} A related doctrine, the “sovereign acts doctrine,” protects similar interests. As explained by Joshua Schwartz:

> These doctrines preserve the government’s ability to respond effectively to changed circumstances that call for a policy response without undue inhibition because of the collateral effects such a response may have upon subsisting government contracts. At the same time, these rules of law should be framed so as to provide appropriate protection to the reliance and expectation interests of the government’s contractual partners. Indeed, the government shares a long-range interest in achieving a legal regime in which the risks borne by its contractors do not stand as a barrier to entry into a competitive market for government contracts. Finally, in striking a balance between governmental and contractors’ interests, the sovereign acts and unmistakability doctrines must also maintain the constitutional separation of powers among the branches of the federal government.


\textsuperscript{173} See Parker v. Wakelin, 123 F.3d 1, 5 (1st Cir. 1997) (noting that the “threshold requirement for the recognition of public contracts has been referred to as the ‘unmistakability doctrine’”).

\textsuperscript{174} Id. at 5 (quoting United States v. Winstar Corp, 518 U.S. 839, 875 (1996)).

\textsuperscript{175} See United States v. Winstar Corp., 518 U.S. 839, 871–81 (1996). Justice Souter’s plurality opinion in *Winstar* relied upon the purposes of the unmistakability doctrine to argue that the strength of the doctrine should be calibrated to reflect the extent to which a particular contract limits sovereign powers. Id. 878–81. Contracts that would limit important powers such as the
contractual obligation, there are no clear standards governing the determination. \(^{176}\) Rather, all of the facts and circumstances surrounding each alleged contract must be closely examined to determine whether the state legislature intended to create a contractual relationship. \(^{177}\)

It is not altogether clear that the analogy between public pension benefits and cases like *Charles River Bridge* and even Social Security reform legislation is apt. Unlike the typical regulatory program, pension benefits are earned through government employment and, especially with regard to past services, are compensation for work already performed. In employment situations, perhaps the presumption should be flipped—it ought to be presumed that promises made based on employment are intended to be contractual. \(^{178}\) Otherwise, state taxing power should be subject to a strict unmistakability doctrine while “humdrum supply contracts” should not. \(^{176}\) See *Parker*, 123 F.3d at 4 (concluding “that a blanket answer to the issue of Contract Clause protection for vested employees is not possible, because... a detailed examination of the particular provisions of a state pension program will be required prior to determining the nature and scope of the unmistakable contractual rights”).

\(^{176}\) The high bar to finding a contractual obligation stands in contrast to the relatively easier time government workers and government benefits recipients have in establishing property interests in their jobs or benefits. Under the test developed under *Board of Regents v. Roth*, 408 U.S. 564 (1972), government benefits and employment are considered property under federal law whenever ascertainable standards govern their award and termination. \(^{176}\) See *Parker*, 123 F.3d at 4 (concluding “that a blanket answer to the issue of Contract Clause protection for vested employees is not possible, because... a detailed examination of the particular provisions of a state pension program will be required prior to determining the nature and scope of the unmistakable contractual rights”).

\(^{177}\) *See Parker*, 123 F.3d at 4 (concluding “that a blanket answer to the issue of Contract Clause protection for vested employees is not possible, because... a detailed examination of the particular provisions of a state pension program will be required prior to determining the nature and scope of the unmistakable contractual rights”).

and local employers would be free to take advantage of employees in exactly the way that the Contract Clause, as applied to the government’s own contracts, is supposed to prevent. Further, allowing state and local governments complete freedom to alter employee benefits retroactively could hamper public employers’ ability to attract high quality employees or reduce employers’ flexibility regarding the timing of pay and benefits if employees refuse to accept insecure promises of deferred compensation. With regard to Social Security, even though benefits are based on contributions, the case for allowing reform is still much stronger than in the government employment situation. People are likely to understand that Social Security is a government benefits program subject to legislative change.

The high bar against finding a contractual obligation in pension contracts is illustrated by the First Circuit’s decision in Parker v. Wakelin, a case involving statutory amendments to Maine’s public employee retirement laws. The amendments, enacted in 1993, made several changes to the pension system that were unfavorable to employees. Some of the changes applied to all employees while others applied only to those employees with less than ten years of creditable service. The changes that affected all employees included an increase in the required employee contribution to the pension plan (from 6.5% to 7.65%), a cap on salary increases that may be used in calculating pension benefits, and a six-month delay in a retiree’s first cost of living increase. For employees with less than ten years of service, the minimum full pension retirement age was increased from 60 to 62, the penalty for retiring early was increased from 2.25% of the pension benefit to 6% of the pension benefit for each year before

179. Parker v. Wakelin, 123 F.3d 1 (1st Cir. 1997).
180. Id. at 2 (“The question presented by this appeal is whether certain legislative amendments to the Maine State Retirement System (MSRS) violate the Contract Clause of the United States Constitution . . . .”).
181. See id. at 3 (discussing the changes made to Maine’s public employment retirement scheme).
182. See id. (indicating that “three changes apply to the pensions of all current teacher-members”).
183. See id. (noting that three changes applied only to those having less than ten years of creditable service).
age 62, and the ability of employees to include unused sick and vacation pay in calculating pension benefits was eliminated.¹⁸⁴

While many state courts treat pension promises as unilateral contracts that are entered into when the employee begins working,¹⁸⁵ the First Circuit explicitly rejected a blanket rule treating all pensions that way.¹⁸⁶ Instead it chose to closely analyze Maine law to determine whether the State of Maine intended to bind itself to the pension promises made to employees as embodied in the statutory provisions as they existed before the amendments.¹⁸⁷ The most significant indication of contractual intent on the part of the Maine legislature was a statute enacted in 1975 which states: “No amendment to this chapter shall cause any reduction in the amount of benefits which would be due to the member based on creditable service, compensation, employee contributions and the provisions of this chapter on the date immediately preceding the effective date of such amendment.”¹⁸⁸

This is a typical provision found in state law to protect public employee pensions. The question is whether it satisfies the unmistakability doctrine’s standard for finding intent to create a binding contract to maintain pension benefits as of the date a public employee was hired, i.e., whether it creates a contractual obligation.


¹⁸⁶. Parker v. Wakelin, 123 F.3d 1, 4 (1st Cir. 1997) (“We now conclude that a blanket answer to the issue of Contract Clause protection for vested employees is not possible . . . .”).

¹⁸⁷. Id. at 8. The Third Circuit has held that even in Pennsylvania where the state courts view pension promises as contractual, no contractual right exists if the pension plan explicitly provides that administrators have the power to make alterations to the plan. See Transp. Workers Union of Am., Local 290 ex. rel. Fabio v. Se. Pa. Transp. Auth., 145 F.3d 619, 624 (3d Cir. 1998).

In *Parker*, the district court had found that no changes could be made to the potential benefits of Maine employees with enough service to retire before the changes took effect, but that the benefits of employees with some service but not enough to retire could be reduced.\(^{189}\) The court of appeals viewed the question as turning on the meaning of the word “due” in the 1975 statute quoted above.\(^ {190}\) If due means what *would* be payable if the employee retired, then the promise was contractual and the state could not alter the terms of the pension plan.\(^ {191}\) If, however, the word due refers to amounts actually due and owing, then only retired employees already receiving pension payments are protected because no amounts are due to an employee who has not already retired.\(^ {192}\)

Based in part on the reasoning of the Maine Supreme Court in an earlier case involving pension reform,\(^{193}\) the First Circuit held that the 1975 statute was not sufficient to create a contractual obligation in favor of any employee who had not yet retired, even if the employee was eligible to retire but had not yet done so.\(^ {194}\) In the earlier case, the Maine Supreme Judicial Court rejected the notion that pension terms become binding contractual promises at the moment of employment.\(^ {195}\) The First Circuit reasoned that this indicates that the word due in the 1975 statute does not refer to pension terms in effect at the time of employment.\(^ {196}\) The court, however, recognized that this does not resolve the question whether pension terms might be due once an employee has sufficient creditable service, and is old enough, to retire.\(^ {197}\) For the First Circuit, in light of the Maine court’s

---

189. *See Parker*, 123 F.3d at 2 (recapitulating the district court’s holding).
190. *See id.* at 8–9.
191. *See id.* at 8.
192. *See id.*
194. *See Parker v. Wakelin*, 123 F.3d 1, 9 (1st Cir. 1997) (“We need not decide whether the statute ever gives rise to a contractual relationship; it is enough to say that it does not clearly do so before a teacher retires, and thus gains an immediate right to the payment of pension benefits.”).
196. *See Parker*, 123 F.3d at 8–9 (1st Cir. 1997) (reasoning that the word due does not imply a contractual relationship at the time of employment).
197. *See id.* at 8 (noting that neither party argued whether contractual obligations arise respecting pension terms upon an employee receiving sufficient
understanding of the word due, the unmistakability doctrine tipped the scales against finding a contractual obligation to employees who were eligible to retire. Thus, all of the 1993 pension reforms could be applied to all nonretired Maine employees without violating the Contract Clause. This is a relatively narrow understanding of the Contract Clause’s protection of government pension promises.

A finding that a contractual right in pension benefits exists does not mean that pension reform measures are automatically unconstitutional. As mentioned above, the Contract Clause prohibits only substantial impairments, and, as discussed below, allows substantial impairments if they are “reasonable and necessary to serve an important public purpose.” There is no clear line in the case law between substantial and insubstantial impairments. The central inquiry appears to be whether the complaining party actually relied on the altered term or terms. As one court put it:

In determining whether an impairment is substantial and so not “permitted under the Constitution,” of greatest concern appears to be the contracting parties’ actual reliance on the abridged contractual term. Specifically, the Supreme Court has examined contracts to determine whether the abridged right is one that was “reasonably relied” on by the complaining

credible service hours).

198. If, in a subsequent case, the Maine Supreme Judicial Court holds that the 1975 statute prohibits pension plan changes that alter the benefits that would be paid to employees already eligible to retire, the First Circuit’s conclusion would be subject to revision. However, a case subsequent to such a determination by the Maine court is unlikely to arise in federal court because the state courts would have already prohibited the changes to the pension plan that might violate the Contract Clause.


201. See Balt. Teachers Union v. Mayor & City Council of Balt., 6 F.3d 1012, 1017 (4th Cir. 1993) (noting that the “Supreme Court has provided little specific guidance as to what constitutes a ‘substantial’ contract impairment”).
party, . . . or one that “substantially induced” that party “to enter into the contract.”\(^{202}\)

For example, in a case involving an alleged impairment of municipal bonds issued by a water utility, the bondholders complained that their Contract Clause rights were violated when the water utility was no longer legally entitled to place a lien on property based on a default by a tenant.\(^{203}\) When the bonds were issued, default by a tenant allowed the utility to place a lien on the land even if the owner had not contracted for service.\(^{204}\) This increased the likelihood of payment after default. The court concluded that a loss of the ability to place a lien on the landlord’s property after default by a tenant was not a substantial impairment of the contract:

> The bond contracts themselves contain express acknowledgements that the parties’ rights were subject to legislative regulation; there was a long established precedent of extensive state regulation of public utilities; the contracts were not abolished but merely modified; and the abridged right is, by its nature, not one central to the parties’ undertaking.\(^{205}\)

Another factor that is relevant to whether there is a substantial impairment of a contract under the Contract Clause is whether the law has provided alternative benefits to the party whose rights have allegedly been impaired. As discussed above, this is also an important factor in some states for satisfying state law restrictions on pension modification. Rather than isolate the individual elements in the contractual agreement, courts holistically ask whether the parties’ overall situation has been made significantly worse. For example, in the case involving the bondholders discussed above, in the year before the bondholders lost the right to place liens on landlords’ property, they gained the right, under state law, to terminate water service for nonpayment.\(^{206}\) The court held that the addition of this very


\(^{203}\) See id. at 388.

\(^{204}\) Id. at 387.

\(^{205}\) Id. at 394.

\(^{206}\) See id. at 394–95.
effective remedy for nonpayment meant that overall there was not a substantial impairment of the bondholders’ contractual rights. In the pension reform area, this flexibility can be important as governments struggle to reduce their costs without harming employees who depend on the benefits.

Amy Monahan concludes from her examination of the case law that, in general, changes to the level of benefits and changes that affect the rights and responsibilities of employers are held to be substantial impairments. In her view, except perhaps in extraordinary circumstances, changing the method for calculating benefits so that lower benefits are paid is likely to be found to be a substantial impairment of the contract. Monahan points out, however, that some states, such as California, allow substantial pension reform as “reasonable and necessary” impairments before retirement because, in their understanding of state law, employees have a right to a “substantial or reasonable pension” but not to a specific level of benefits.

Despite this recognition that California courts have allowed substantial pension reform as reasonable and necessary, Monahan is highly critical of California’s general approach to pension reform, an approach that she recognizes has been followed by at least a dozen more states. Monahan states that the California rule recognizing contractual rights in pension promises from the first day of employment is, for several reasons, “surprising”:

First, it runs contrary to the well-established legal presumption that statutes do not create contractual rights absent clear and unambiguous evidence that the legislature intended to bind itself. Second, courts interpreting the California Rule have held that the contract protects . . . the rate of future accrual. This interpretation is contrary to federal Contract Clause jurisprudence, which holds that prospective changes to a contract should not be considered

207. See id. (reasoning that, because the new remedy to terminate water service more effectively served the aims of bondholders than the previous remedy, no substantial impairment of contractual rights occurred).
209. Id.
210. Id. at 628 (citing Betts v. Bd. of Admin., 582 P.2d 614 (Cal. 1978)).
211. Monahan, Statutes as Contracts, supra note 119, at 1032.
unconstitutional impairments. Third, not only is this interpretation contrary to general contract theory, it also appears to create economic inefficiency, in that it fixes in place one part of an employee’s compensation. . . . California courts have held that even though the state can terminate a worker, lower her salary, or reduce her other benefits, the state cannot decrease the worker’s rate of pension accrual as long as she is employed. This framework can be welfare reducing. Given the option, an employee may prefer to accept lower future pension accruals in return for avoiding termination or a reduction in current compensation, but such deals are hard to accomplish in a system that protects the right to future accruals. It should also be noted that the protections the California Rule appears to offer are illusory, given that it simply forces a state that needs to reduce costs to do so in some area other than pension accruals—for example, through layoffs or salary reductions. Viewed holistically, the California Rule simply does not protect employees’ economic interests, and in some cases the rule may even harm the interests of the very employees it is meant to protect.212

Monahan may be correct that California law is contrary to general legal principles and more protective of employees than federal Contract Clause jurisprudence, but I do not find California law “surprising.” On her first point, there are good reasons to treat statutory promises to government employees different from promises contained in other regulatory statutes. Most people have multiple employment options at the outset and at various stages of their careers. Retirement promises form part of the inducement for individuals to choose and remain in government employment. While businesses may be in a similar situation and may suffer, as did the Charles River Bridge Company, when the regulatory rug is pulled out from under them, individuals have much less ability to diversify regulatory risk than businesses. Employees cannot be expected to save two or three times for retirement or change jobs every so often so their retirement promises come from multiple employers. This recognition helps explain why federal law protects private pensions through the ERISA and the programs administered by the Pension Benefit Guaranty Corporation. That the federal Contracts Clause may be less protective than state law is no

212. Id. at 1032–33.
reason for state law to change. Under familiar understandings of federalism, in many situations, federal law should be lenient with regard to state law, especially when the state’s own operations are involved, stepping in only in extreme cases.

As to Monahan’s claim that protecting pension promises is inefficient because the optimal result may be reduced pension promises rather than layoffs that might be necessary to fund remaining employees’ pensions, this is a dilemma that is familiar to anyone studying labor economics. As wages and benefits increase, employers may hire fewer employees, may fire existing employees, and may replace employees with technology or workers in jurisdictions with lower salaries. Some unions have dealt with this problem by agreeing to lower wages and benefits for new employees while protecting the wages and benefits of incumbents. More fundamentally, although Monahan clearly understands that pension promises are a form of deferred compensation, her argument in favor of greater flexibility virtually ignores the ex ante perspective of the parties. At the time the contract was made, had the employees known that their pension promises were subject to significant revision, they may not have accepted government employment or they may have demanded significantly higher current compensation. Normally, the security of contract enforcement is thought to increase efficiency, and Monahan does not refute that general tendency.

Monahan’s strongest point is that protecting future accrual levels significantly reduces pension flexibility. If she is correct that public employees are “generally at-will employees, with no guaranteed period of employment,” then it would make legal and practical sense to allow prospective changes to the terms of a contract that both parties could simply terminate at any time. At-will employees’ reliance on future benefits may be viewed as unworthy of protection. However, there are reasons to doubt her premise. Government employees are highly unionized and are much more likely than private employees to have job security in the form of contractual or civil service protections. Further,

---

213. *Id.* at 1077.

214. See Chris D. Edwards, *Public Sector-Unions*, TAX & BUDGET BULL. NO. 61 (Cato Institute, D.C.), March 2010 (“In 2009, 39 percent of state and local workers were members of unions, which was more than five times the share in the private sector of 7 percent.”).
advocates of prospective change should recognize that, for example, a twenty-year government employee suddenly faced with significantly lower future accrual of retirement benefits may be seriously damaged economically by the change and may not be in a position to seek alternate employment or take some other action to ameliorate the effects of the change.215

Sometimes, pension reforms are touted as providing benefits to plan participants even if the predominant effect of reform is to reduce pension expenditures. At a basic level, current and future recipients benefit from any reform that brings a fund closer to full funding because fund enhancement makes pension promises more secure. There are, however, two problems with generalizing from this possibility to a principle that any reform that enhances the assets of a pension fund survives Contract Clause scrutiny. First, this reasoning would allow serious detriment to some participants as long as most participants gain. While this might be appropriate in some contexts, for example if a reform reduces pension spiking by those at the high end of the benefits scale, it would not be appropriate to sacrifice lower-end recipients who are heavily dependent on their benefits.216 The financial health of the fund should not be shored up on the backs of those who can least afford it. Second, using the financial health of pension funds as a justification for reforms that otherwise harm plan participants is illogical if the pension promises involved are viewed as contractual obligations in favor of recipients. Recipients gain nothing if under state law the plan must live up to the promises made regardless of the financial health of whatever fund has


216. I do not mean to say that payments to those receiving the smallest pensions should be immune from reduction or other reform, such as reducing cost of living increases. The real question is economic dependency. Some retirees receiving small pensions barely worked for the government and just got over the eligibility bar with questionable creditable service, such as volunteer service on a local government board or commission. Other retirees receiving small pensions are highly dependent on those benefits because they worked at relatively low paying government jobs for long periods and did not participate in federal Social Security during that time. It is thus difficult to design reforms based purely on the size of the pension.
been established to marshal assets to make the payments. Reform under such circumstances benefits only the state budget, not pension plan participants.217

The final issue in a Contract Clause controversy examines the government interest advanced by the challenged reforms. Although the Contract Clause is phrased as an absolute prohibition on state laws impairing contracts, as noted, courts apply what appears to be akin to an intermediate level of constitutional scrutiny in Contract Clause cases, asking whether the challenged government acts are “reasonable and necessary to serve an important public purpose.”218 In the pension area, this standard may save reforms that are designed to combat abusive pension practices.219 The question remains, however, whether a pure desire to save money is sufficient to save a reform measure that operates only to reduce payments to retirees, increase contributions from retirees, or both.220

217. I leave to the side for now the possibility of bankruptcy, which might allow greater reductions. See infra Part V.


219. Courts seem open to reforms that curb abusive pension practices. For example, in Madden v. Contributory Retirement Appeal Board, 729 N.E.2d 1095 (2000), the Massachusetts Supreme Judicial Court approved a decision by the Teachers’ Retirement Board to close a loophole that would have allowed a part-time teacher to receive full-time credit for part-time service. Id. at 1100. However, the court disapproved of application of the new rule to part-time service before the rule was adopted. Id. at 1099. The Court stated that modifications in benefits are allowed if they are “reasonable and bear some material relationship to the theory of a pension system and its successful operation.” Id. at 1098.

220. For an argument that budget difficulties and financial downturns should provide adequate reasons to allow states to modify their pension obligations, see Whitney Cloud, Comment, State Pension Deficits, the Recession, and a Modern View of the Contracts Clause, 120 Yale L.J. 2199 (2011). See also Gavin Reinke, Note, When a Promise Isn’t A Promise: Public Employers’ Ability to Alter Pension Plans of Retired Employees, 64 Vand. L. Rev. 1673, 1689–91 (2011) (arguing that saving money is a legitimate government interest supporting pension reform against substantive due process challenge).
In general, it appears that courts rarely approve substantial impairments as supported by a sufficient government interest.\(^{221}\) In support of pension reform, it might simply be argued that saving money is an important public purpose and thus, especially if obligations to retirees pose a fiscal crisis as some claim, reducing pension obligations is “reasonably necessary” to serve that interest. The problem is that this could be said about virtually any breach of contract by government—the government has decided that it would be better off not living up to its promises because, at a minimum, it saves resources. As the Supreme Court stated in a Contract Clause case not involving public pensions:

\[
\text{Merely because the government actor believes that money can be better spent or should now be conserved does not provide a sufficient interest to impair the obligation of contract. If a State could reduce its financial obligations whenever it wanted to spend the money for what it regarded as an important public purpose, the Contract Clause would provide no protection at all.}^{222}
\]

There should therefore be some additional government interest behind pension reform. Such an interest might be in eliminating fraud or abusive pension practices that detract from equity among workers and result in unjustifiable benefits, that is, benefits with no relation to the retirement income that the employee was relying on as part of government service.

It is unclear whether the government interest in saving money on pension expenses would be more acceptable if it were linked to a history of overly generous promises and abusive practices. The government should be viewed as having an interest in closing loopholes that allow abusive practices. In general, government has an interest in protecting the integrity and fairness of programs it administers.\(^{223}\) Courts should be more

\(^{221}\) See Monahan, [*Public Pension Plan Reform*, supra note 118, at 631 (“The only public pension plan cases identified that found substantial impairments to be reasonable and necessary to serve an important public purpose were cases in which the court first held that no substantial impairment occurred.”).  


\(^{223}\) See, e.g., United States v. Borjesson, 92 F.3d 954, 955–56 (9th Cir. 1996) (recognizing as important the government’s interest in maintaining integrity and the appearance of integrity in government programs); Donovan v.
receptive to reforms that target practices that are regarded as abusive than to reforms that reduce benefits to employees who legitimately relied on them.

The propriety of considering the government’s interest in saving money as the interest behind pension reform is also linked to the structure of the pension plan and state law on whether pension promises are strictly enforceable. If plan participants are legally entitled under state law to their promised payments regardless of whether the state has set aside sufficient funds to meet its obligations, it would seem that the simple interest in saving money should not be sufficient to support pension reform. Under such circumstances, to allow government’s interest in saving money to support reducing benefits would essentially nullify the plan participants’ legal rights without any compensatory benefit.

B. The Takings Clause and Pension Reform

Another possible constitutional constraint on pension reform is the Takings Clause, which prohibits government from taking property for public use without compensation. In litigation involving public pensions, it is common for claims under the Contract Clause and Takings Clause to be made together over the same reform because under current understandings government contractual promises may be considered property for constitutional purposes. With regard to state and local reforms, the Takings Clause is unlikely to add much to claims under the Contract Clause because a participant’s interest in pension promises is unlikely to be property unless it is found to be a contractual promise protected under the Contract Clause or state

Fitzsimmons, 778 F.2d 298, 319 (7th Cir. 1985) (“[A]side from protecting the individual beneficiaries of these pension programs, the government in this case clearly has a separate and unique interest in protecting the very integrity, heart and lifeline of the program itself.”).

224. See U.S. CONST. amend. V (“[N]or shall private property be taken for public use, without just compensation.”).

225. See, e.g., San Diego Police Officers’ Ass’n v. San Diego City Emps.’ Ret. Sys., 568 F.3d 725, 736–41 (9th Cir. 2009) (discussing plaintiffs’ allegation that failure to fund a pension plan adequately violated both the Contracts Clause and Takings Clause).
law pension doctrine.  It is theoretically possible, however, that a reform that does not violate the Contract Clause, because the government’s action is reasonable and necessary to serve an important public purpose, violates the Takings Clause. This is because the government’s justification for a taking is irrelevant—if it takes property even for the most important of purposes, it must pay compensation.

The takings claim is strongest with regard to benefits that have already been paid, and might also be relatively strong with regard to reforms that reduce pension payments to people already receiving them. In National Education Ass’n-Rhode Island ex. rel. Scigulinsky v. Retirement Board of the Rhode Island Employees’ Retirement System, involving “evictions” of participants from a state pension plan, the First Circuit upheld legislation that halted public pension payments to private union employees. The legislation required the state to repay, with interest, these participants’ contributions to the system insofar as they exceeded what the participants had received in payments. The court noted that “[p]ension payments actually made to retirees become their property and are protected against takings, even if and where the payments are unquestionably a gift.” The law is less clear with regard to promises made to people who have already retired. Some courts view such benefits as vested and immune from reduction. Other courts view such benefits as regulatory

226. See, e.g., Picard v. Members of Emp. Ret. Bd. of Providence, 275 F.3d 139, 144 (1st Cir. 2001) (“In evaluating whether a purported contract or property right is entitled to constitutional protection under the Takings Clause, Contract Clause, or Due Process Clause, this Court generally looks to state law as interpreted by the state’s highest court.”); Nat’l Educ. Ass’n-R.I. ex rel. Scigulinsky v. Ret. Bd. of R.I. Emps.’ Ret. Sys., 172 F.3d 22, 30 (1st Cir. 1999) (“It would make nonsense of such rulings—and the clear intent requirement—to conclude that an expectancy insufficient to constitute an enforceable contract against the state could simply be renamed ‘property’ and enforced as a promise through the back door under the Takings Clause.”).


228. See id. at 31 (finding constitutional the Rhode Island Eviction Act, which eliminated retirement benefits to employees of teachers’ unions).

229. See id. at 24–25 (describing the Rhode Island Eviction Act at issue).

230. Id. at 30.

231. See, e.g., Pierce v. State, 910 P.2d 288, 292 (N.M. 1999) (finding that retirement plans create a property right in the amount of benefits promised
promises that are open to change, assuming state law does not clearly immunize them from revision. The same can be said of benefit promises to people eligible to retire at the time reforms are enacted. Some courts treat these as vested and immutable, but again, this depends largely on the terms of state law. Because of this connection to state law, the Takings Clause is likely to follow the Contract Clause in recognizing only those claims that involve unmistakable contractual promises already protected from reduction under state law.

The possibility that a pension reform measure that satisfies Contract Clause scrutiny but nevertheless might require compensation under the Takings Clause implicates the thorny issue of the extent to which regulation under the state’s police power that reduces the value of property can constitute a taking of that property requiring compensation. If each dollar of promised pension benefits is viewed as a separate property interest, then it would seem that any diminution would violate the Takings Clause. But if instead the property interest is viewed as the value of the pension as a whole, then reforms that preserve the bulk of expected benefits should not be problematic. In this Article, I will not attempt to resolve the conceptual difficulties that plague regulatory takings doctrine. It should be noted, however, that the application of regulatory takings analysis is

upon vesting and requiring compensation for their reduction); see also Reinke, supra note 220, at 1694 (discussing the approaches of different courts with respect to promised future benefits).  

232. See Reinke, supra note 220, at 1693 (discussing the approaches of different courts with respect to promised future benefits).  

233. See id. at 1694 (examining the impact of reform laws on retirees).  

234. Regulatory takings doctrine has proven very lenient in terms of allowing changes in government regulation to cause substantial reductions in the value of private property without requiring compensation. See, e.g., Adrus v. Allard, 444 U.S. 51, 66 (1979) (“When we review regulation, a reduction in the value of property is not necessarily a taking.”). However, the law is very strict when government requires the actual physical occupation of private property. See, e.g., Loretto v. Teleprompter Manhattan CATV Corp., 458 U.S. 419, 426 (1982) (holding that “a permanent physical occupation authorized by government is a taking without regard to the public interests that it may serve”). It is difficult to fit reduction in pension benefits into this paradigm. On the one hand, if each dollar of expected benefits is considered a separate piece of property, then taking one away might be considered a taking. On the other hand, if the property interest is in a reasonable pension in light of work performed, then reforms may not appear to be prohibited takings.
highly uncertain in the public pension context because the property rights at issue are contractual and perhaps even regulatory, which makes it difficult to separate the terms of state law from the value of the property allegedly taken.235

The reasons for the relative leniency of regulatory takings law apply in the context of pension reform. Regulatory takings law recognizes that adapting government policy to changed circumstances or new priorities would be impossible if every regulatory diminution in the value of a property interest requires compensation. Flexibility is even more important if it appears that pension promises are overly generous, subject to abuse by legislators and other officials handing out political favors, and by employees using loopholes and tricks to spike their pensions. It is one thing for the government to breach a simple arm’s-length contract with a supplier of goods or services. It is quite another for government to attempt to rein in excessive pension promises made to secure the power of incumbent politicians at the expense of taxpayers. Just as the law does not generally recognize a reliance interest in a static regulatory environment, so too is it unlikely to recognize a reliance interest in a completely static public pension system.236 To the extent that courts apply the Takings Clause to pension reform, they are unlikely to rule against reforms except in the most extreme circumstances.237

As noted, takings analysis is likely to mirror the analysis undertaken pursuant to state law pension protections and the Contract Clause. The Takings Clause may have independent bite in one potentially significant situation—when pension reform is undertaken pursuant to federal law, either because changes are

235. See supra Part IV.A (discussing the property rights in public pensions as contractual).


237. See, e.g., Concrete Pipes, 508 U.S. at 602 (applying the Takings Clause of the Fifth Amendment to the pension system); Connolly, 472 U.S. at 221 (considering whether employer withdrawal liability for public pensions is a compensable taking under the Takings Clause).
being made to federal government pensions or because state pensions are adjusted pursuant to federal law, most notably federal bankruptcy law. The Contract Clause does not apply to the federal government and therefore federal changes to existing contractual relationships are scrutinized under the more lenient minimal scrutiny applied to substantive due process challenges to economic regulation.\textsuperscript{238} If federal law allows or even requires the reduction of pension benefits to federal or state and local employees, the Takings Clause might be the most promising avenue for attacking the reform. In the current context, a key issue is whether a municipality can use federal bankruptcy law to discharge its pension obligations. As discussed below, the answer appears to be yes, and, because the Contract Clause does not apply to the federal government, the principal legal question becomes whether a discharge pursuant to bankruptcy law could be viewed as an uncompensated taking. This is discussed below.\textsuperscript{239}

\textbf{V. Bankruptcy, Reduction of Pension Obligations, and Default}\textsuperscript{240}

Chapter 9 of the Bankruptcy Code\textsuperscript{241} allows for the “adjustment of debts of a municipality.”\textsuperscript{242} In short, local government units can declare bankruptcy and have their debts adjusted under federal law.\textsuperscript{243} Municipalities may not employ

\begin{thebibliography}{99}
\bibitem{238} See Pension Benefit Guar. Corp. v. R.A. Gray & Co., 467 U.S. 717, 730 (1984) (“The retroactive aspects of legislation, as well as the prospective aspects, must meet the test of due process[.] [T]hat burden is met simply by showing that the retroactive application of the legislation is itself justified by a rational legislative purpose.” (internal quotation marks omitted) (citations omitted)).
\bibitem{239} See infra Part V (concluding that takings principles are unlikely to prevent state and local governments from pursuing pension reform through bankruptcy or otherwise).
\bibitem{240} I am indebted to Ted Orson, lawyer for the city of Central Falls, Rhode Island, and the state of Rhode Island in the city’s municipal bankruptcy proceedings for guiding me through Chapter 9 of the Bankruptcy Code and offering his perspective on the subject. For a theoretical overview of municipal bankruptcy, see generally Michael W. McConnell & Randal C. Picker, \textit{When Cities Go Broke: A Conceptual Introduction to Municipal Bankruptcy}, 60 U. Chi. L. Rev. 425 (1993).
\bibitem{242} Id. § 901.
\bibitem{243} See id. (allowing a municipality to declare bankruptcy and develop a
federal bankruptcy law if the law of their state does not allow it. In other words, local governments need state permission to declare bankruptcy. In theory, in states in which municipal bankruptcy is allowed, federal bankruptcy law could be employed by municipal governments to reduce or eliminate their pension obligations.

There are significant differences between municipal bankruptcy and bankruptcy of private entities. Most significantly, there is no provision for liquidation of municipal assets and termination of the existence of the municipality. It is thought that federal liquidation of a municipal government would be too great an intrusion into state authority. Further, bankruptcy may not be used to restructure the municipal government because that too would interfere with state authority over municipalities. Finally, there is no provision in federal law for states themselves to declare bankruptcy, and any such effort would be met with serious constitutional objections.

There are five statutory conditions that must be met for municipalities to use Chapter 9 to adjust their finances. First, the municipality must be authorized under state law to be a debtor under Chapter 9. Second, the debtor must actually be a

---

244. See id. § 109(c)(2) (specifying an entity may be a debtor under Chapter 9 only if specifically authorized by state law).
245. The funded portion of future pension benefits might not be subject to adjustment in bankruptcy, but unfunded obligations might be subject to “discharge at less than full payment.” Skeel, supra note 83, at 692.
246. See id. (explaining the lack of liquidation provisions in Chapter 9).
247. See id. (“Such a liquidation or dissolution would undoubtedly violate the Tenth Amendment to the Constitution and the reservation to the states of sovereignty over their internal affairs.”).
249. See id. § 109(c) (listing the requirements for a municipality to enter Chapter 9 bankruptcy).
municipality. For villages, cities, towns, counties, and such, this is normally not a difficult condition to meet, but status as a municipality may be less clear for other government entities, such as water districts, school districts, and other special purpose agencies. Third, the debtor must be insolvent. "Insolvent" is defined in the Bankruptcy Code to mean either failing to pay debts or "unable to pay its debts as they become due." In the case law, this is interpreted to mean not only that the municipality is running a deficit but also that it will be unable to pay its debts in the current or next fiscal year. Fourth, the municipality must desire to make a plan to reorganize its debts. This precludes involuntary municipal bankruptcy. Fifth, the municiplity must do one of the following: obtain agreement from creditors holding a majority of claims, negotiate in good faith with creditors, show that negotiation would be impracticable, or reasonably believe that a creditor will obtain a preference absent bankruptcy. Usually, this fifth requirement results in negotiations with creditors before the municipality files.

Municipal bankruptcy allows for adjustment of pension liabilities to both retired workers and current workers, at least

---

252. Id. § 109(c)(3).
253. Id. § 101(32).
254. See In re City of Bridgeport, 132 B.R. 85, 88–89 (Bankr. D. Conn. 1991) (requiring the city to show it was unable to pay bills as they came due). Apparently, a high percentage of municipal filings are rejected, as the city of Bridgeport’s was, because the municipality is not legally insolvent. See McConnell & Picker, supra note 240, at 457–60 (describing the gatekeeper effect of the insolvency requirement).
256. Id. § 109(c)(5).
257. See McConnell & Picker, supra note 240, at 460–61 (explaining that most debtors negotiate prepetition).
258. During the legislative process leading to the adoption of the current version of Chapter 9, there were concerns expressed over the effects of municipal bankruptcy on pensions. It is not clear that legislators understood the extent to which municipal pensions would be subject to adjustment in bankruptcy. On the floor of the Senate, New York Senator Jacob Javits expressed the view that in light of the New York constitution’s provision protecting pension rights, due process would prevent pensions already being received from being subject to adjustment in bankruptcy. 122 Cong. Rec. 4377 (1976) (remarks of Senator Javits). In response to Senator Javits’s request for
with regard to assets that are not held by an entity separate and apart from the insolvent municipality.\textsuperscript{259} For current workers, their labor contracts are considered executory contracts under § 365 of the Bankruptcy Code, which is explicitly applicable to municipal bankruptcy.\textsuperscript{260} Debtors are authorized by § 365 to reject their executory contracts.\textsuperscript{261} Thus, in the bankruptcy of Central Falls, Rhode Island, on the day the petition was filed, the

confirmation of his understanding, North Dakota Senator Quentin Burdick stated that with regard to those whose pension rights had vested, "[u]nder New York law it would be, at the very least, a paramount claim on any assets of the bankruptcy." \textit{Id.} (remarks of Senator Burdick). It is unclear what Senator Burdick meant by a "paramount claim" since there is no provision in the Bankruptcy Code establishing priority for pension claims. In the House of Representatives, New York Representative Elizabeth Holtzman was concerned that rejection of collective bargaining agreements might leave retirees in the position of unsecured creditors. \textit{Id.} at 2422 (remarks of Representative Holtzman). New York Representative Herman Badillo raised the concern that pension funds administered by boards of trustees that included municipal officials might be considered arms of the municipality making their assets subject to adjustment in bankruptcy. Representative Badillo stated for the record that his understanding of the intent of Congress was that such trustees are not acting as municipal officials and thus municipalities do not have any claim on the assets of such separately administered pension funds, which would place them beyond the reach of the bankruptcy court. \textit{See} 122 \textit{Cong. Rec.} 2382 (1976) (remarks of Representative Badillo). This discussion appears to be based on an understanding that if pension funds have been placed into a trust fund separate and apart from the bankrupt municipal government, these funds would not be subject to adjustment in bankruptcy. Future municipal payments to such funds might, however, be adjusted.

\textsuperscript{259} When pension assets have been placed into a trust administered for the benefit of employees, it seems that municipal bankruptcy could not affect those assets. Retirement benefits could still be affected, however, if trust assets are inadequate to continue the level of payments. Bankruptcy would presumably allow the municipality to refuse to make future payments, which would leave the trust unable to maintain the level of pension benefits promised. This would make it necessary for the trust to reduce retirees’ pension payments. For current workers, a similar result is likely. Municipal bankruptcy may not affect employees’ claims to a share of assets already in trust, but bankruptcy would allow the municipality to reduce future payments to the trust, thus reducing the employees’ ultimate pension benefits on retirement.


\textsuperscript{261} \textit{See} 11 U.S.C. § 365(a) (2006) ("[T]he trustee, subject to the court’s approval, may assume or reject any executory contract or unexpired lease of the debtor.").
city rejected all of its collective bargaining agreements and imposed new terms of employment, including new provisions relating to pensions. Due to the special nature of collective bargaining agreements, rejection of municipal collective bargaining agreements is allowed only if the balance of equities favors rejection. If this standard is met, municipalities can

262. Unless otherwise noted, all information concerning the Central Falls, Rhode Island bankruptcy is drawn from a conversation with Ted Orson, bankruptcy attorney for the city of Central Falls and the state of Rhode Island and from the Chapter 9 plan for the city filed with the Bankruptcy Court. Interview with Ted Orson, Bankruptcy Attorney, City of Central Falls and State of Rhode Island, in Bos., Mass. (July 26, 2012) [hereinafter Orson Interview]; see also Fourth Amended Bankruptcy Plan, In re City of Central Falls, Rhode Island, 468 B.R. 36 (Bankr. D.R.I. 2012) (No. 11-13105) [hereinafter Central Falls Bankruptcy Plan], available at http://www.rib.uscourts.gov/newhome/central_falls/CF479.asp.

263. NLRB v. Bildisco & Bildisco, 465 U.S. 513, 527–34 (1984). It should be noted that after the Supreme Court decided that it was not an unfair labor practice for a debtor to reject a collective bargaining agreement in Bildisco, Congress enacted special provisions regarding rejection of collective bargaining agreements. See Bankruptcy Amendments and Federal Judgeship Act of 1984 § 541, Pub. L. No. 98-353, 98 Stat. 333 (codified at 11 U.S.C. § 1113 (2006)); Bildisco, 465 U.S. at 527–34. Those provisions are not among those listed by Congress as applying to municipal bankruptcy, which means that rejection of municipal collective bargaining agreements is governed by Bildisco’s balance of equities standard, which the Supreme Court determined was the most accurate reading of Congress’s intent regarding the application of § 365 to collective bargaining agreements. See Bildisco, 465 U.S. at 521–26 (determining that Congress’s intent was likely that the municipality may reject agreements only if the equities balance in favor of rejection); see also Note, Executory Labor Contracts and Municipal Bankruptcy, 85 YALE L.J. 957, 965 (1976) (suggesting that the balance of equities in municipal bankruptcy is likely to point in favor of rejection to preserve the municipality’s ability to provide essential services). Another difference between rejection under § 365 and rejection under § 1113 is that when a contract is rejected under § 365, the creditor has a claim for damages as an unsecured creditor for breach of contract. See 11 U.S.C. § 365(g) (2006) (“[T]he rejection of an executory contract or unexpired lease of the debtor constitutes a breach of such contract or lease.”). By contrast, the dominant view is that when rejection is accomplished under § 1113, the affected employees have no claim for damages because their rights have already been determined under federal bankruptcy law. See, e.g., In re Blue Diamond Coal Co., 147 B.R. 720, 732 (Bankr. E.D. Tenn. 1992) (“[T]he Bankruptcy Code, as presently enacted, does not provide or recognize a remedy for damages resulting from rejection of a collective bargaining agreement under § 1113.”); see also Executory Labor Contracts and Municipal Bankruptcy, supra, 85 YALE L.J. at 968–73 (discussing renegotiation of rejected collective bargaining agreements and not suggesting that unionized employees would have a claim for damages after rejection). For a discussion of the constitutionality of state and local rejection of collective bargaining agreements, see generally Ronald D. Wenkart, Unilateral...
reduce or eliminate pension and health care promises to current workers, and require them to contribute more toward the costs of both.

With regard to retired municipal workers already receiving pension benefits, the situation is simpler as a legal matter but more complicated as an equitable or political matter. Because retired employees have no substantial remaining contractual obligations to the municipality, their pension promises are no longer considered executory contracts. Rather, under bankruptcy law, the obligation to make future pension and health care payments to retired workers is a simple debt of the debtor, and the creditors (retired workers) have only unsecured claims against the municipality. The claims are unsecured because workers do not have separate individual accounts into which their retirement contributions (and the employer’s matching contributions) have been deposited. In fact, for municipalities


264. See Countryman, supra note 260, at 460 (defining executory contract); see also Hannah Heck, Comment, Solving Insolvent Public Pensions: The Limitations of the Current Bankruptcy Option, 8 Emory Bankr. Dev. J. 89, 124 (2011) (concluding that pension obligations to retired workers are not executory contracts because retirees have no continuing contractual obligations). This comment also argues that any state law impediments to implementation of federal bankruptcy law in the public pension context (other than state refusal to allow its municipalities to use Chapter 9) would be preempted by federal law. Id. at 120–21.

265. For example, after the City of Stockton, California, filed a Chapter 9 case in June 2012, the city council adopted a budget that reduced retiree health care benefits. In re City of Stockton, 478 B.R. 8, 14 (E.D. Cal. 2012). The bankruptcy court denied the retirees’ request for an injunction to restore their benefits to prebankruptcy levels, mainly on the ground that the court had neither the power nor the jurisdiction to grant such an injunction. See id. at 30 (finding that 11 U.S.C. § 904 forbids the bankruptcy court from issuing the requested injunction). The bankruptcy court in the Stockton case also observed that the Contract Clause is no impediment to adjustment of municipal contracts pursuant to bankruptcy because the Contract Clause does not apply to federal law. Id. at 15.

266. See Jeffrey B. Ellman & Daniel J. Merrett, Pensions and Chapter 9: Can Municipalities Use Bankruptcy to Solve Their Pension Woes?, 27 Emory Bankr. Dev. J. 365 (2011), 401–02 (“[A] Chapter 9 debtor’s postpetition obligations to its retirees arising out of prepetition contractual (or impliedly contractual) relationships arguably are entitled to nothing more than general unsecured nonpriority status and may be impaired in a plan of adjustment.”).
with severe underfunding, benefits may be paid out of the contributions of current employees, and there is no segregation of the funds contributed by each worker and by the municipality itself on behalf of each worker.267

This means that theoretically, retired workers could see their benefits subjected to severe reduction. Given that many municipal workers have not participated in the federal Social Security system,268 this could cause serious hardship, basically placing retirees without other savings into abject poverty. There have been few municipal bankruptcies and even fewer in which pensions to current retirees are adjusted that there is no real precedent for how retirees ought to be treated. The proposed plan in the Central Falls, Rhode Island bankruptcy, which cited pension and health care obligations to retirees as a major cause of insolvency,269 would reduce most retirees’ pension benefits by 55%, except that no retiree’s pension would be reduced below $10,000 per year.270 While these cuts may seem draconian, the plan treated retirees better than other unsecured creditors. Apparently, there was a strong feeling among those involved in the bankruptcy that it would have been inhumane to reduce retirees’ benefits to the level they would get as unsecured creditors.271

To some, it may still seem cruel to reduce pensions so much. Many of Central Falls’s employees worked for the city for decades, always expecting that their pensions would be paid based on the formula established in their employment contracts.272 They may have relied on those funds in making important life choices such as whether to continue their city

267. See Heck, supra note 264, at 96 (“[A]dditional difficulties result when local governments attempt to make up the shortfalls in pension revenues by drawing down pension reserves or funds from pension trusts and from funding pension obligations by continuing current employee contributions.”).

268. See Biggs, supra note 23, at 2–3 (explaining that municipal retirement plans are often a substitute or supplement for the Social Security system).

269. See Mary Williams Walsh & Katie Zezima, Small City, Big Debt Problems, N.Y. TIMES, Aug. 1, 2011, at B1 (explaining that Central Falls’s financial troubles were largely due to its pensions and health care systems).

270. See Central Falls Bankruptcy Plan, supra note 262 (describing the provisions of the city’s reduction in retiree pension benefits).

271. Orson Interview, supra note 262.

272. Id.
employment, whether to save or spend their salaries, whether to move, and whether to go back to school to train for a different profession. As discussed above, most pensions are not unreasonable when viewed in light of the employees’ total compensation packages. A worker retiring at a $50,000 salary may see a $35,000 pension reduced to under $16,000, and may have increased health care costs. This is a serious hardship to the people involved and may be life changing for many of them. However, this is the pain caused in many situations of insolvency. Just as Bernard Madoff’s clients were led to believe that their investments were worth much more than was true, the city of Central Falls misled its employees. Apparently the city failed for years to make its actuarially required contributions on behalf of its employees. The money to pay retirees’ pensions in full was simply not there. The question is whether the city or the state should be required to increase taxes or employ some other financial device to make good on these promises.

Given the lack of precedent, it remains to be seen whether other unsecured creditors will challenge favorable treatment to retirees in municipal bankruptcy as unfair to them, perhaps arguing that they will receive lower payouts on their claims as a result of the favorable treatment of pension claims. It also remains to be seen how federal bankruptcy courts will react to such claims. Congress could amend the Bankruptcy Code to deal with the problem, but federalism concerns counsel against it. Congress may not want to interfere with the local political considerations that are likely to affect the treatment of retirees in municipal bankruptcy. Some states have enacted legislation allowing state authorities to assume supervision over distressed municipalities. More specifically, in Central Falls, the retirees

---

273. See supra Part II.B (describing pension benefits in the context of total compensation).

274. Orson Interview, supra note 262.


276. See, e.g., 50 ILL. COMP. STAT. 320/4 (2010) (allowing a local government to petition the state for assistance in cases of financial distress); MICH. COMP. LAWS § 141.1515 (2011) (allowing the Governor to make a determination of financial distress and declare a local government in receivership); N.C. GEN. STAT. § 159-3 (2010) (creating the Local Government Commission to take control
negotiated for a five-year transition period during which their pension benefits would be reduced by only 25%.

This was contingent on the state legislature providing funding during the transition period, which it did. Federal standards on the treatment of government retirees in bankruptcy might interfere with these local political efforts.

The Contract Clause of the federal Constitution is no bar to municipal bankruptcy for the simple reason that the Contract Clause does not apply to the federal government. While the first municipal bankruptcy law was found to violate the Contract Clause by allowing municipalities to violate their contracts, this does not appear to be the current understanding. Later, the constitutionality of federal bankruptcy for municipal governments was upheld against challenges based on federal interference with state sovereignty and due process, and it does not seem that a challenge based on the contractual or property rights of municipal creditors would succeed either. Instead of relatively stringent Contract Clause scrutiny, federal interference with the obligation of contracts is judged under the deferential rational basis standard applied to economic regulation generally.

277. Orson Interview, supra note 262.

278. Id.

279. Federal laws affecting the obligation of contracts are evaluated under a less exacting due process standard. For discussion on this point, see Pension Benefit Guar. Corp. v. R.A. Gray & Co., 467 U.S. 717, 733 (1984) (“To the extent that recent decisions of the Court have addressed the issue, we have contrasted the limitations imposed on States by the Contract Clause with the less searching standards imposed on economic legislation by the Due Process Clauses.”).


282. See Pension Benefit Guar. Corp., 467 U.S. at 728–31 (“Provided that the retroactive application of a statute is supported by a legitimate legislative purpose furthered by rational means, judgments about the wisdom of such
Even in these difficult financial times, municipal bankruptcy has been very rare.283 Further, even if municipal bankruptcy became more common, it would have no effect on the large portion of unfunded retirement obligations owed by states to their current and retired workers. As noted, there is no provision for state governments to file for bankruptcy under federal law.284 There is a question of whether it would be constitutional to amend federal law to allow states to file for adjustment of their finances in the same fashion as municipal governments. Professor Skeel notes that advocates of state bankruptcy do not find the constitutional objection to be serious if two conditions that already apply to municipal bankruptcy are met—the filing must be voluntary, and bankruptcy must not interfere with governmental decisionmaking.285 He also notes that these advocates view the constitutional permissibility of municipal bankruptcy as strong precedent for the constitutionality of state bankruptcy.286

It is not absolutely clear that the approval of municipal bankruptcy is precedent for finding no constitutional difficulty with state bankruptcy. The status of municipal governments under federal law is inconsistent to say the least. Long ago, in refusing to intervene in a dispute concerning municipal boundaries and responsibility for municipal debts, the Supreme Court stated as a basic principle that municipal governments exercise state governmental power and are created and organized purely for the convenience of the states: “Municipal corporations are political subdivisions of the state, created as convenient

legislation remain within the exclusive province of the legislative and executive branches.”).

283. See Municipality Bankruptcy, USCourts.gov, http://www.uscourts.gov/FederalCourts/Bankruptcy/BankruptcyBasics/Chapter9.aspx (last visited Feb. 2, 2013) (“In the more than 60 years since Congress established a federal mechanism for the resolution of municipal debts, there have been fewer than 500 municipal bankruptcy petitions filed.”) (on file with the Washington and Lee Law Review).

284. For a discussion of the possibility of authorizing states to employ bankruptcy to restructure their debts, see generally Skeel, supra note 83.

285. Id. at 679–80.

286. See id. at 680 (“[M]unicipal bankruptcy has long been constitutional if it satisfies these criteria and gives states the power to forbid their municipalities from invoking the law.”).
agencies for exercising such of the governmental powers of the state as may be entrusted to them.”287 In this light, if municipalities are simply state agencies, then the constitutionality of municipal bankruptcy should provide a strong precedent for the constitutionality of state bankruptcy.

There are, however, many ways in which municipal governments and state governments are treated differently under federal law. In the civil rights area, state governments, including state agencies, are immune from damages by virtue of the Eleventh Amendment and principles of sovereign immunity, while municipal governments are not.288 States are not “persons” subject to federal civil rights liability in state courts while municipal governments are.289 Given that the Contract Clause is directed explicitly at states—“No State shall . . . pass any . . . Law impairing the Obligation of Contracts”290—perhaps state attempts to reduce their contractually binding pension obligations should be treated differently than similar actions by municipal governments.

The fact that municipalities are subject to state control may provide a basis for treating state and municipal governments differently with regard to the possibility of using bankruptcy law to adjust their debts. Unlike other debtors, states theoretically have the ability to raise whatever funds are necessary to pay their debts through taxation. Municipal governments may not have this ability because they are subject to state control.291 The state legislature could prevent a locality from raising sufficient funds to pay their debts by forbidding increased taxation or limiting revenue sources. Further, a geographically small

287. See Hunter v. City of Pittsburgh, 207 U.S. 161, 178 (1907) (explaining that municipal governments have state governmental power).


289. See Will v. Mich. Dep’t of State Police, 491 U.S. 58, 71 (1989) (“We hold that neither a State nor its officials acting in their official capacities are ‘persons’ under § 1983.”); Monell v. Dep’t of Soc. Servs., 436 U.S. 658, 690 (1978) (“Our analysis . . . compels the conclusion that Congress did intend municipalities and other local government units to be included among those persons to whom § 1983 applies.”).


291. See Hunter, 207 U.S. at 178 (describing municipal governments as agents for the convenience of the state).
municipal government is much more likely to run up against practical limits on its taxing ability than a large municipality or a state.

States, by contrast, lack funds only when the states’ own governments decline to raise them through sufficient taxes and fees. Lack of taxable wealth may limit the ability of states to raise revenue, but this is much more likely to be a local problem than a statewide problem. As a conceptual matter, unless taxation is at such a high level that there is simply no more wealth to tax, from the point of view of a creditor, state bankruptcy looks more like a political decision not to pay debts than a true state of insolvency.²⁹² However, this picture is somewhat incomplete. While it is true that state taxpayers in a state with underfunded pension liabilities are able to push off some of the costs of state services onto future taxpayers, it is difficult to blame those future taxpayers for resisting tax increases to pay for pension liabilities incurred in the past when they may not have been enjoying the benefits of the services provided in exchange for the unfunded pension promises. Similarly, it is not difficult to imagine that future federal taxpayers would resist tax increases to pay the $16 trillion in debt that the federal government has incurred in the last twelve years or so.

If bankruptcy is not available to states, does that mean that they are stuck with their pension and health care obligations to retired workers? Theoretically, the answer seems to be yes, as perhaps it ought to be given the interests of state retirees and employees. Default on pension obligations, or alterations beyond those allowed under Contract Clause jurisprudence, would violate the federal Constitution and would also be contrary to the law in many, if not all, states. However, just because state action violates federal law does not guarantee an effective remedy.

²⁹² Tax increases sufficient to meet all unfunded pension obligations may be economically disastrous and state taxpayers as a whole would be better off if states were allowed to reduce their obligations rather than raise taxes. High taxes can put a damper on economic activity and encourage business to move to lower tax states or countries. However, state bankruptcy to avoid pension obligations would exacerbate the unwillingness of state politicians to raise sufficient funds for pension obligations, which either results in hardship for workers relying on their pensions or imposes the cost of current labor on future generations of taxpayers.
Surprisingly, whether states can be sued in federal court over alleged constitutional violations in pension reform is unclear.

Controversy over federal remedies for state contractual violations goes back to the beginnings of the republic. When the state of Georgia defaulted on its bonds after the Revolutionary War, the Supreme Court ruled in *Chisholm v. Georgia* that Georgia could be sued in federal court by a nonresident for breach of contract. The state legislature reacted by considering and nearly passing a statute imposing the death penalty, “without benefit of clergy,” on anyone attempting to enforce the judgment in the case. The decision also provoked Congress and the states to pass and ratify the Eleventh Amendment to the Constitution, which reversed the jurisdictional ruling in the *Chisholm* case. One hundred years later, the Supreme Court ruled that the Eleventh Amendment and principles of sovereign immunity precluded federal court jurisdiction over a claim brought by a citizen of Louisiana alleging that the state violated the U.S. Constitution by defaulting on bonds issued in 1874.

Thus, it appears that the federal cases establish that states cannot be sued for damages in federal court without their consent, even for actions that violate the Constitution of the United States. However, under well-established principles,

---

293. *Johnson & Young, supra* note 178, contains an excellent overview of the history and development of law relating to state default on debt.
295. *Id.* at 458.
296. See *William A. Fletcher, A Historical Interpretation of the Eleventh Amendment: A Narrow Construction of an Affirmative Grant of Jurisdiction Rather Than a Prohibition Against Jurisdiction*, 35 STAN. L. REV. 1033, 1058 (1983) (discussing the aftermath of *Chisholm*).
297. See U.S. CONST. amend. XI (providing sovereign immunity to the states).
298. See *Hans v. Louisiana*, 134 U.S. 1, 15–17 (1890) (sheltering the state from suit in federal court on a case arising under the Constitution).
299. See *Johnson & Young, supra* note 178, at 136 (“The general structure of American state sovereign immunity law is designed to prevent courts from compelling payment on debts that threaten the financial viability of the states.”). It appears to be an open question whether states can avoid their Takings Clause obligation to pay compensation for takings of private property by interposing a sovereign immunity defense. A decision by the Supreme Court recognizing sovereign immunity from takings claims would be shocking. The Takings Clause appears to be a limit on sovereignty of both the federal government and state governments now that the Takings Clause applies to
when state law violates the federal constitution, state officials can be sued in their official capacities for injunctive relief, even if the injunction requires future payments from the state treasury. This means that a state official could be ordered, on pain of contempt, to make future payments found to be constitutionally required, but the official could probably not be ordered to make past payments wrongfully withheld. Thus, the

them under the Fourteenth Amendment. See Chi., Burlington & Quincy R.R. Co. v. City of Chicago, 166 U.S. 226, 232–37 (1897) (applying Takings Clause principles to the states pursuant to the Fourteenth Amendment’s Due Process Clause). Perhaps Contract Clause claims against states for breaching their own contracts should be thought of the same way. This depends, however, on the expansion of the Contract Clause to cover the state’s own contracts, and state immunity from contract damages is not directly contrary to clear constitutional text the way that immunity from takings claims would be (again assuming the Takings Clause applies to the states through the Fourteenth Amendment).

300. See Ex parte Young, 209 U.S. 123, 146–47 (1908) (allowing state officials to be sued to prevent them from enforcing unconstitutional laws.) The reach of Young is not completely clear. Other decisions from the same era seem to validate sovereign immunity in federal court from contract damages and from suits seeking specific performance of contracts between a state and private parties. See Ex parte Ayers, 123 U.S. 443, 507–08 (1887) (preventing the state from being sued on a case arising under the Constitution); Hagood v. Southern, 117 U.S. 52, 71 (1886) (refusing to compel states to perform a contract); Louisiana ex rel. Elliott v. Jumel, 107 U.S. 711, 748 (1883) (finding a contract unenforceable against state officials carrying out their duties). Hagood was distinguished in Young as a case in which the state was the actual party in interest, which seems to be the case with regard to pension reform as well. See Young, 209 U.S. at 150 (concluding that Hagood applies when the state is a party on the record).

301. See Milliken v. Bradley, 433 U.S. 267, 289–90 (1977) (requiring the state to share future costs of educational components following desegregation); Edelman v. Jordan, 415 U.S. 651, 668 (1974) (holding that state may not be required to make payments for past violations of federal law but noting that injunctions requiring future payments to comply with federal law are permissible).

302. See Quern v. Jordan, 440 U.S. 332, 346 (1979) (disallowing a notice that would have led to a retroactive award in state court requiring the state to make payment of funds from the treasury). Professor Skeel posits that “the officer could evade a mandamus action seeking to compel performance of the contract by simply resigning.” Skeel, supra note 83, at 686. But normally when the case is brought in the officer’s official capacity, the new occupant of the resigned official’s office is substituted, and the case continues without regard to the resignation of the officer. For example, Quern became the defendant in Edelman v. Jordan when he took Edelman’s position in the state of Illinois. See Quern, 440 U.S. at 333 (noting Quern is the sequel to Edelman). No doubt, state and local officials may sometimes succeed in avoiding liability, but it is not likely to be so simple as resigning once the official is ordered to comply with federal law
conventional understanding seems to be that if state pension reform is found to violate the federal Constitution, injunctive relief may be available in federal court to require responsible officials to make future payments and administer the program based on preexisting standards, but they could not be ordered to make up for past reductions in payments or other past violations.\textsuperscript{303}

This conventional understanding of the line between permissible and impermissible federal relief against state officials is more complicated than it seems because it is not entirely clear that injunctions requiring increases in future payments to meet constitutional obligations are allowed. Consider, for example, a recent decision by a federal district court in New Jersey finding that an attack on pension reforms in New Jersey is barred by the Eleventh Amendment.\textsuperscript{304} The state of New Jersey passed legislation increasing employees’ required contributions to state pension funds and suspending cost of living allowances for both current and future retirees.\textsuperscript{305} The plaintiffs challenged these reforms as impairing the obligation of contracts, taking property without just compensation, and as violating their due process rights.\textsuperscript{306} The court, in a thoughtful opinion, found federal jurisdiction barred because rather than challenging an ongoing violation of federal law, the plaintiffs were seeking a remedy for a past violation, namely the passage of the pension reform statute at issue.\textsuperscript{307} Under this reasoning, there is no federal remedy for a

\begin{flushleft}
305. See id. (describing the changes made under the New Jersey law).
306. Id.
307. The court structured the inquiry as follows:

[T]he question to be answered in this case is appropriately framed as determining whether Plaintiffs’ requested relief is retroactive or prospective in nature. Therefore, at the heart of this Court’s Eleventh Amendment analysis is the following question: was the enactment of Chapter 78 a single act that has continuing ill-effects or does the enforcement of Chapter 78 by the Executive Defendants amount to a continuous violation of the Plaintiffs’ constitutional rights? . . . After
\end{flushleft}
state's breach of contract even if the breach violates the Contracts Clause.308

The final substantive issue to be addressed is whether the Takings Clause provides protection against diminution of government pension obligations pursuant to bankruptcy, on the theory that pensions are property that may not be taken without just compensation. Takings analysis turns out not to be a promising avenue of attack for public pension plan participants seeking to avoid costly reform. In short, although the Supreme Court has made it clear that takings principles apply to bankruptcy's effects on property,309 the Takings Clause is unlikely to provide protection for public pension recipients and government employees with accrued service toward pensions because bankruptcy and other reform does not deprive the pension plan participants of an interest in identifiable property.

Takings principles limit the ability of bankruptcy to destroy creditors' property interests including liens and security interests that creditors often hold in debtors' property.310 It does not

examine the nature of Plaintiffs' claims, the Court has determined that the enactment of Chapter 78 was a single act that continues to have negative consequences for the Plaintiffs. As such, any redress sought by the Plaintiffs would be retroactive in nature and is therefore barred by the Eleventh Amendment.

Id. at *4 (citations omitted).

308. See id. at *5 (“Therefore, the relief requested by Plaintiffs is, in both substance and practical effect, a request for specific performance of the alleged pre-Chapter 78 contract existing between Plaintiffs and the State of New Jersey. Under controlling Supreme Court precedent, such relief is not permitted.” (citing Va. Office of Prot. & Advocacy v. Stewart, 131 S. Ct. 1632, 1639 (2011))). This conclusion may be consistent with Johnson & Young's analysis, assuming that a suit seeking an injunction to force payment of promised pension obligations is viewed as a suit to compel payment of "the original debt or obligation." See Johnson & Young, supra note 178, at 136 (concluding that American state sovereign immunity law is designed to prevent states from being forced to pay debts that threaten the state's financial viability).


310. See Sec. Indus. Bank, 459 U.S. at 78 (construing Bankruptcy Code not
THE PUBLIC PENSION CRISIS

83

protect purely contractual rights such as the details of pension promises made by governmental units to past and current employees. The whole point of bankruptcy is to adjust such unsecured obligations among the creditors so that no creditor or class of creditor gains an unfair share of the debtor’s assets. The analysis might be different if employees’ and retirees’ funds were held in segregated accounts for the benefit of each employee or retiree. Absent an identifiable fund “owned” by the pension recipient or the employee, such as perhaps an annuity purchased in the name of the recipient or a brokerage account in the name of the recipient, the fact that state and local pension promises might be considered “property” for due process purposes does not mean that they are protected by the Takings Clause from rejection or reduction in bankruptcy.

However, even if there is no federal remedy available, state constitutional and statutory provisions, discussed above, may authorize destruction of liens to avoid constitutional question of whether destruction would be a taking requiring compensation. The interaction between bankruptcy law and takings principles became an issue during the Great Depression when Congress enacted statutes providing for relief of bankrupt homeowners against mortgage foreclosure. See Louisville Joint Stock Land Bank v. Radford, 295 U.S. 555, 590 (1935) (finding bankruptcy law unconstitutional insofar as it authorized “the taking of substantive rights in specific property acquired by the” creditor, namely a mortgage held by a bank). But see Wright v. Vinton Branch of Mountain Trust Bank of Roanoke, Va., 300 U.S. 440, 462–63 (1937) (upholding amended provisions preserving mortgagees’ interest while imposing a stay on foreclosure proceedings subject to the discretion of the federal court). For a more recent affirmation of the constitutionality of adjusting mortgagees’ rights in bankruptcy, see Travelers Ins. Co. v. Bullington, 878 F.2d 354, 358–60 (11th Cir. 1989).

311. See In re Nolan, 232 F.3d 528, 534 n.10 (6th Cir. 2000)

Every bankruptcy involves a “transfer” of private property from a creditor to a debtor, in the sense that a creditor is involuntarily deprived of a previously-vested, legally-enforceable debtor obligation to return borrowed creditor property. However, mere reconciliation of debts among private entities does not normally constitute taking private property for public use.

In municipal bankruptcy, the “public use” requirement might be met, but the adjustment of claims would still not constitute a “taking.”

312. Even if there were some separable property interest that could be claimed by each public pension plan participant, ordinarily the interest would be protected only to the extent of its value at the time of bankruptcy. See Skeel, supra note 83, at 698 (“It is quite likely that a court would conclude that pension beneficiaries do have a property interest, but only to the extent of the funds the state has set aside for payment.”).
impose substantial impediments to state pension reform.\textsuperscript{313} As noted, many states prohibit diminution of pension benefits for both retired workers and workers currently employed.\textsuperscript{314} In such states, state courts may declare null and void legislation or executive action purporting to reduce benefits. In these states, and in light of the possibility of federal civil rights injunctive relief, federal bankruptcy law may be necessary to bring about meaningful pension reform in some states. However, as noted, most states do not allow their municipalities to employ bankruptcy to adjust their debts.\textsuperscript{315} Whether courts in those states would prevent reform even in dire financial circumstances, remains, perhaps, to be seen.

\textbf{VI. Concluding Observations}

The public pension crisis, in part, is a state and local analog to the spiraling federal debt. Without significant reform, state and local governments will have to devote increasingly large portions of their limited revenues to fulfilling pension promises that may have been made decades before. We have already seen significant reductions in government services in states with high pension costs, such as California. That state, which once boasted of the most comprehensive and inexpensive higher education systems in the nation, is now finding it impossible, for example, to continue to offer sufficient community college slots for all students.\textsuperscript{316} Pension costs are a major contributor to California’s financial difficulties.\textsuperscript{317}

\textsuperscript{313} See supra Part III.B (discussing state law limitations on pension reform).
\textsuperscript{314} See supra Part III.A (noting state prohibitions on diminution of pension benefits).
\textsuperscript{315} See supra notes 244–45 and accompanying text (explaining a municipality may not use federal bankruptcy law if applicable state law does not allow it).
The pension funding crisis is different from other forms of deficit spending because it involves obligations to individuals, specifically current and former government employees. Most references to the “public pension crisis” are to the financial aspects of the problem. This masks the most important crisis, the human crisis. The vast majority of people receiving government pensions are not wealthy. If many pension plans follow the lead of Central Falls, Rhode Island, it would be a crushing financial blow to many pension recipients, especially those who never participated in the federal Social Security system. Most state and local pensions are relatively modest, and the workers and employers involved have contributed to their pensions the way that workers and employers in the private sector pay Social Security taxes and contribute to 401k accounts, often coupled with employer contributions.318 These workers have structured their finances and made career and personal decisions in reliance on their pension expectations. Reforms that involve significant reductions in pension payouts or large increases in employee pension contributions may appear to be unfair to the majority of workers who have not engaged in any significant manipulation of their pension entitlements. Of course, when a private business goes into bankruptcy, many people’s legitimate expectations are upset, even people who cannot afford the losses they are forced to bear.

In a sense, public pension recipients are in a similar position, but on the lower end of the economic scale, to the victims of Bernard Madoff’s Ponzi scheme. In many public pension funds, the level of contributions was established based on the expectation that the funds would earn 8% per year.319 While average returns over the last twenty years or so may be in that

---

318. State and local employers and employees contribute, on average, a total of 18.5% of salary to public pension funds covering employees not participating in Social Security (10.5% for the employer and 8% for the employee). Social Security contributions for other workers total 12.4%, with employers and employees contributing 6.2% each. ALICIA H. MUNNELL & MAURICIO SOTO, CTR. FOR STATE AND LOCAL GOV'T EXCELLENCE, ISSUE BRIEF: STATE AND LOCAL PENSIONS ARE DIFFERENT FROM PRIVATE PLANS 4 (2007), http://crr.bc.edu/wp-content/uploads/2007/11/slp_1.pdf.

319. BAKER, supra note 24, at 5.
range, over the last decade the returns have been closer to 6%,320 with a 3.2% annual return over the past five years.321 In the fiscal year ending June 30th, 2012, the two largest California public pension funds earned 1% and 1.8%, while New York State’s largest fund earned 6% in its fiscal year that ended in March, before significant market losses in the second quarter of 2012.322 If returns remain well below the 8% level usually relied upon, underfunding will only get worse, and pension fund participants’ expectations will become more and more unrealistic.

Another analog to the Madoff scandal is that these workers were likely led to believe that their employers were contributing to the pension fund in amounts sufficient to fund the promises that were being made. Just as Madoff’s victims received fabricated statements indicating investment gains that did not exist, government workers were told what level of benefits they should expect and that money was being set aside each month on their behalf.

The fairness of significant reductions in pension benefits depends on a variety of considerations, including the magnitude of the contributions made by retirees and employees to the retirement system; the degree to which pensions were spiked in ways not related to the true earnings of the employees; the degree to which employees accepted lower current wages in exchange for generous retirement benefits; and the other ways in which employees structured their finances and their personal and professional lives around their pension expectations. Employees may have rejected other employment opportunities such as moving into higher-paying private sector jobs without pension benefits and they may have saved less for retirement in reliance on their pensions. These decisions are irrevocable for older workers and retirees who have insufficient or no remaining time left in the work force to ameliorate the consequences of these decisions.

321. BIGGS, supra note 23, at 12.
322. Mincer, supra note 320.
Reforms may seem less unfair if pension promises were unrealistically generous in light of contributions to pension funds and true rates of return on pension fund investments. Reduction in benefits may not seem unfair if contributions similar to or slightly higher than those made to Social Security resulted in pension promises two, three, or four times higher than Social Security benefits that would have been earned in that program.\footnote{As noted above, public pension participants contribute somewhat more to their pension funds than the amounts required for participation in Social Security. See supra notes 47–49 and accompanying text (discussing the comparison between Social Security contributions and pension contributions). Higher returns for public pension participants may be justified in part because those funds invest in the stock market, while Social Security funds are invested only in federal Treasury bills that earn a relatively low rate of return.}

If workers or their unions understood that their contributions were based on projected returns that were way out of line with the market, it might not seem unfair to make them bear some of the pain of the shortfall that has resulted, especially if government salaries are similar to or even higher than private sector salaries, as some analysts claim.\footnote{For this view, see, for example, Andrew Biggs & Jason Richwine, \textit{Comparing Federal and Private Sector Compensation} (Am. Enter. Inst. for Pub. Policy Research, Working Paper, 2011), http://www.aei.org/files/2011/06/08/AEI-Working-Paper-on-Federal-Pay-May-2011.pdf (arguing that government compensation is higher than private sector compensation).} However, this ignores the inducement aspect of pension promises, that state and local workers were induced to accept and remain in their jobs in part based on the pension promises that were continually made during their employment.

Reform may seem even less unfair when it is directed at activities that seem to fall into the general category of pension spiking. Insofar as pension benefit calculations are inflated by including overtime, secondary jobs, longevity pay, and artificial promotions, reducing benefits may seem perfectly fair. Public pensions should compensate employees fairly and provide economic security, not provide an opportunity to game the system. Of course, rules in many areas of law are subject to manipulation, but it is generally not viewed as unfair when reforms are directed at issues properly characterized as “loopholes.”
The Massachusetts Supreme Judicial Court’s statement that changes to pension plans are constitutional if they are “reasonable and bear some material relationship to the theory of a pension system and its successful operation”\textsuperscript{325} can help provide the basis for a general understanding of how the contractual underpinnings of contemporary pensions should be tempered to allow for reforms to abusive pension practices. Government pensions are designed to provide financial security, incentives to faithful long-term government service, and to perhaps make up for the lower salaries of government employees, while providing for the reduced economic needs for retired workers as compared to people still active in the work force. Under traditional, straightforward contract principles, employees can make a persuasive case that they should be able to take advantage of all of the features of the pension system in place during their employment. These could include provisions that enable pensions to be spiked based on second, part-time jobs, volunteer service, and longevity bonuses, designed simply to increase pensions. The Massachusetts court’s comment encourages viewing pension reform from the perspective of the goals and nature of a pension system rather than as a simple contractual arrangement. Amounts earned through “gaming” the system are inconsistent with the theory of a pension system. No worker should have a legitimate expectation of a pension boosted by part-time work, end-of-career promotions, and longevity pay earned simply by informing the government employer that retirement is a year or more away.

The simple contractual view is inconsistent with contemporary application of the Contract Clause. Rather than simply disallow all retrospective modifications of the terms of both private and government contracts, the Supreme Court allows even substantial impairments of government contracts if they are supported by an important government interest. This contemporary standard rejects a simplistic contractual view of government-citizen contractual relations in favor of a more realistic view, imbued with policy and analysis of the legitimacy of private expectations.

Reforms targeting abuses should be allowed under any theory. Government employees may recognize that there are contractually-protected loopholes and devices that allow them to spike their pensions. They also probably understand, however, that they are taking advantage of technicalities that go beyond the spirit of the government pension program. A purely contractual view would not take this into account. The core of pensions based on a person’s true long-term service and economic reliance should be protected, but contractual formalities should not prevent the closing of loopholes and the elimination of methods that allow pension spiking.

Fairness aside, if the financial situation of government pension funds does not improve, many state workers and retirees may suffer severe reductions in their pension benefits as public entities find it economically or politically impossible to meet their obligations to retired workers. Municipalities may reduce pension benefits through bankruptcy and states may unilaterally reduce benefits and use their unique positions as sovereign states to resist judicial remedies based on state or federal law. These possibilities may give pension plan participants strong incentives to negotiate over their pension benefits, perhaps resulting in the acceptance of significant reductions that are less painful than what would have otherwise occurred.

What might the future hold for the public pension systems? While reflecting on the relative impecunity of many government pension recipients and their legitimate expectations based on years of contributions and service, it is worth considering whether public pensions should be bailed out the way that financial institutions have been bailed out in the past. According to the website propublica.com, 928 institutions have received more than $600 billion in federal bailout funds during the recent financial crisis.\footnote{Bailout Recipients, ProPublica.COM (Oct. 12, 2012), http://projects.propublica.org/bailout/list (last visited Feb. 2, 2013) (on file with the Washington and Lee Law Review).} This includes nearly $200 billion to the quasi-governmental Fannie Mae and Freddie Mac, and nearly $70 billion to the insurance company AIG.\footnote{Id.} Other large institutions receiving billions of dollars in bailout funds include General
Motors, Bank of America, and Citigroup. There have been additional government bailouts in the United States, including the rescue of New York City in 1975, Chrysler in 1980, and the savings and loan industry in the late 1980s and early 1990s. Perhaps the federal government should step in, in a cooperative plan with the states, and provide funds, loans, and other financing to bail out underfunded public pension funds. If the government is willing to provide funds for mismanaged banks and insurance companies, why not for pension funds? In fact, this would not be the first time that the federal government provided financial assistance to distressed states, although the tendency is for the federal government to stand by while states default on their debts rather than bail them out.

One modest proposed bailout of state and local pension funds is for the federal government to guarantee pension obligation bonds issued by states. Additional proposals in the same vein would provide federal guarantees or favorable tax treatment for such bonds on the condition that the state adopt certain austerity measures such as moving to defined contribution pension plans for new employees and fully funding existing defined benefit plans. These proposals are designed to relieve some of the fiscal pressure on state and local governments while preserving employees’ pensions.

328. Id.
329. Id.
330. See Johnson & Young, supra note 178, at 137–38 (describing federal assumption of state Revolutionary War debts in exchange for allowing the establishment of the national capital on the banks of the Potomac as well as federal aid to states in several recessions since 1973).
331. See id. (describing general practice of federal government allowing states to default, especially in the 1840s and late nineteenth century).
333. See id. at 276–77 (describing the measures taken by many states).
334. Johnson & Young discuss in detail conditions the federal government might impose on states receiving federal bailout funds. See Johnson & Young, supra note 178, at 139–42. There are constitutional limits on the conditions the federal government may place on the receipt of federal funds. See South Dakota v. Dole, 483 U.S. 203 (1987) (upholding the federal government’s conditioning of
There are many practical reasons to be cautious about bailing out public pension funds. The most obvious is that it would be very expensive: in the trillions of dollars, especially if health care promises to retirees are included. It should be noted that according to the *New York Times*, the government’s total commitment of loans and other investments in the recent financial bailout may total more than $12 trillion, but still, in present circumstances, any request to spend trillions more would be greeted with great skepticism to say the least. Further, bailing out hundreds of public pension funds would be a difficult and complex undertaking with enormous moral hazard implications. Each of the hundreds of underfunded pension funds is underfunded to a different degree and got there in its own way. Some funds were abusive, with extravagant promises and minimal contributions, while others simply suffered from lackluster investment performance perhaps owing to unrealistic, but good faith, predictions. Some large bureaucracy, like the Resolution Trust Corporation of the savings and loan crisis, but much larger, would have to be created, and standards would have to be developed to guide the treatment of the funds based on numerous variables.

The moral hazard problem is also significant. In some states and localities, corruption has contributed significantly to extravagant pension promises. Unless serious consequences are

---

highway grants on lowering the minimum drinking age, but noting that these conditions are subject to several restrictions, including being in pursuit of the general welfare). In the pension area, they speculate that a federal bailout “might also require the states to alter some of their obligations to public-sector unions, pension holders and the like.” Johnson & Young, supra note 178, at 143. They speculate that the federal government may not have the power to require states to violate the Contract Clause as a condition for receiving federal funds. *Id.* at 143. As in the case of municipal bankruptcy, this may not be a real problem when federal law dictates changes to state pension plans. There would be no Contract Clause violation since the federal government is not subject to the Contract Clause. *Id.* at 143. Rather, federal legislation must meet the much more lenient constitutional standard governing retroactive legislation. *Id.* at 144–46 (discussing Eastern Enterprises v. Apfel, 524 U.S. 498 (1998), which struck down as unconstitutional the Coal Act, which retroactively imposed liability on coal companies for their employees’ health care costs).

attached to abusive behavior and effective controls are put in place, losses may continue after bailouts.\textsuperscript{336} We have seen this in what seems, at this point, to be a regularly occurring cycle of bailouts directed at financial institutions. If state and local pension funds are too big to fail, their managers can continue with their untoward behavior, assured that the federal government will be there to pick up the pieces when things fall apart.

Despite all of this, the human case in favor of a bailout is compelling. The Rhode Island legislature recognized this when it appropriated funds to cushion the blow suffered by Central Falls retirees. The possibility of large numbers of retirees without sufficient pensions to stay out of poverty may not threaten to bring down the entire financial system, but it is a prospect that is contrary to the ideals established by the Social Security system, that the elderly should have sufficient resources to live out their remaining years with dignity. Of course, there are competing demands for every government dollar, and in an era with no appetite for tax increases, spending on the elderly may come out of funds that might have been devoted to education or health care for children and the poor. There are obviously no easy answers, but the possibility of a large-scale bailout should at least be part of the conversation. Retirees are entitled to at least as much consideration as financial institutions and government bondholders.

Looking at the more distant future, steps ought to be taken to avoid the possibility of this happening again. Investment volatility and political considerations are likely to continue to threaten the financial viability of pension funds if they continue as currently structured. As of yet, there has been no large-scale movement in government away from benefit plans toward defined contribution 401k-style plans. This may be due to a combination of worker resistance and a perceived financial difficulty of making the transition when underfunded pension plans need continued contributions to move toward actuarial soundness.\textsuperscript{337}

\footnotesize
\textsuperscript{336} For example, consider the multi-billion dollar trading losses suffered by JPMorgan Chase after it received (and paid back) $25 billion in federal bailout funds. \textit{See} Dan Fitzpatrick, \textit{J.P. Morgan Hits Executive Reset Button}, \textit{WALL ST. J.}, July 2, 2012, at B1 (describing JPMorgan’s trading losses).

\textsuperscript{337} One respected expert, a zealous advocate of requiring that pensions be
In Massachusetts, for example, one element of pension reform is a long-term schedule for eliminating municipal pension underfunding by requiring higher municipal contributions until full funding is achieved in 2025.\footnote{338. M A S S. G E N. L A W S ch. 32, § 22c (2008). It appears that the Massachusetts legislature altered the schedule in reaction to the stock market and general financial downturn of 2008.} Assuming that no significant movement is made away from defined benefit plans toward contribution plans, reform is likely to include further attacks on pension spiking and a combination of reduced benefits and increased contributions from workers. More states may require their workers to join the federal Social Security program and then scale down the size of pensions accordingly. Health care benefits are likely to be cut by requiring greater contributions from retirees toward premiums, and by increasing co-pays and deductibles.

One final thought. The recent controversy over collective bargaining rights in Wisconsin and related events may lead some to believe that the public pension crisis is less about the problem of chronic underfunding of pensions, and more about the slow but steady elimination of economic security for middle class workers in the United States. Public employment is the last bastion of unionized labor in the United States. Unionized workers tend to earn higher salaries and benefits and enjoy greater job security than their nonunionized counterparts. Perhaps because of this, many unionized jobs in the United States' private sector have disappeared, with manufacturing leading the way. Until now, relatively low-level state and local employees have been able to remain in the middle class and have enjoyed economic security in retirement. Pension reform and elimination of collective bargaining rights could signal the end of that.\footnote{339. Pension reform advocate Robert Costrell blames collective bargaining for the high cost of teacher fringe benefits, including health care expenses and pension promises not based on teacher contributions to the pension fund. Robert M. Costrell, “GASB Won’t Let Me”: A False Objection to Pension Reform, LJAF POLICY PERSPECTIVE (May 2012), http://www.arnoldfoundation.org/sites/default/files/pdf/A9R4D8C.pdf (last visited Feb. 2, 2013) (on file with the Washington and Lee Law Review). Costrell does not believe that there is an actual transition problem, and he advocates linking pension benefits to actual contributions to the pension plan. Id.} It may be only a
matter of time before the twentieth century is viewed by public workers as the good old days.

M. Costrell, *Oh, to be a Teacher in Wisconsin: How Can Fringe Benefits Cost Nearly as Much as a Worker's Salary? Answer: Collective Bargaining*, WALL ST. J. (Feb. 25, 2011), http://online.wsj.com/article/SB10001424052748703408604576164298717724956.html (last visted Jan. 19, 2013) (on file with the Washington and Lee Law Review). Costrell points out that in Wisconsin, some collective bargaining agreements, including Milwaukee's, provided that the school district pays both the employer and employee contributions to the pension system. *Id.* This is also true with regard to health care premiums, so that in fiscal year 2011, teachers in Milwaukee contributed nothing to their health care premiums, which amounted to 50.9 cents on top of every dollar paid in wages. *Id.* These practices have been altered by reform legislation in Wisconsin. *Id.*
Why *Strickland* is the Wrong Test for Violations of the Right to Testify

Daniel J. Capra*
Joseph Tartakovsky**

Abstract

A criminal accused has a constitutional right to testify in his own defense. The right has an undisputed place alongside the most important “personal” rights, like the right to remain silent or the right to represent oneself. But in the 1990s, courts began to apply the ineffective-assistance test of *Strickland v. Washington* to evaluate claims by a defendant that his right to testify was abridged. In practice this nullifies the right. Moreover, the Strickland test is inapposite because it focuses on counsel and not the defendant’s right to testify. This Article proposes a new test to better secure and enforce the right, without subjecting courts to burdensome post-trial motions.

Table of Contents

I. What Is the Constitutional Right to Testify? .................... 96
   A. Origins and Meaning Today........................................ 96
   B. The Right to Testify as Part of the Autonomy Line.................................................................103

II. Why *Strickland* Should Not Be Applied in Right-to-Testify Cases........................................106
   A. The Results of Applying *Strickland* to Right-to-Testify Claims...........................................106

* Reed Professor of Law, Fordham Law School.
** James Wilson Fellow in Constitutional Law at the Claremont Institute for the Study of Statesmanship and Political Philosophy.
I. What Is the Constitutional Right to Testify?

A. Origins and Meaning Today

Anyone born since the invention of television knows that the “Fifth” allows you to keep silent in the face of a prosecution.1 But what about the converse right—the right, during that prosecution, to speak up, to give your side, to explain the mistakes of witnesses, to justify yourself? This is the right to testify, and like the right against compelled self-incrimination, it is a right that defendants decide whether to exercise in every criminal trial. Yet the accused’s right to choose to testify is often violated with impunity by lawyers.

1. See U.S. Const. amend. V ("No person . . . shall be compelled in any criminal case to be a witness against himself . . . .").
A century and a half ago, American states followed the common law in disabling criminal defendants from testifying. The notion behind these so-called “incompetency” statutes was that the self-serving testimony of a defendant was inherently untrustworthy (if not perjurious) and that incompetency statutes actually did the accused a favor—these laws protected the right against self-incrimination by relieving defendants of the negative inference that would be drawn if they were permitted to take the stand but refused. Yet according to Justice Brennan in Ferguson

2. See Ferguson v. Georgia, 365 U.S. 570, 570 (1961) (describing “the common-law rule that a person charged with a criminal offense is incompetent to testify on his own behalf at his trial”).

3. See James Fitzgerald Stephen, A General View of the Criminal Law of England 202 (1863) (“[I]t is not in human nature to speak the truth under such a pressure as would be brought to bear on the prisoner, and it is not a light thing to institute a system which would almost enforce perjury on every occasion.”). For a good state case on this subject, see State v. Wilcox, 175 S.E. 122, 123 (N.C. 1934), in which the Supreme Court of North Carolina reversed a trial judge who instructed that “the law presumes when a man is being tried for crime that he is naturally laboring under a temptation to testify to whatever he thinks may be necessary to clear himself of the charge.” But see McVeigh v. United States, 78 U.S. 259, 267 (1870) (rejecting the rationale that a defendant should be disqualified from testifying because he has the incentive to testify falsely). The Court stated:

If assailed there, he could defend there. The liability and the right are inseparable. A different result would be blot upon our jurisprudence and civilization. We cannot hesitate or doubt on the subject. It would be contrary to the first principle of the social compact and of the right administration of justice.

Id.

4. See Ferguson, 365 U.S. at 571–76, 581–82 (describing the historical justifications for incompetency statutes). For this reason, the federal competency statute, first enacted in 1878, but provided in its present form in 1948, reads as follows:

In trial of all persons charged with the commission of offenses against the United States and in all proceedings in courts martial and courts of inquiry in any State, District, Possession or Territory, the person charged shall, at his own request, be a competent witness. His failure to make such request shall not create any presumption against him.

v. Georgia,\(^5\) decided in 1961, experience made us wiser.\(^6\) Maine acted first, in 1864, enacting a general competency act for criminal defendants;\(^7\) by the end of the nineteenth century, every state but Georgia followed.\(^8\) “In a vast number of instances,” the Supreme Court explained in Wilson v. United States,\(^9\) in 1893, discussing the effects of the federal competency statute enacted fifteen years earlier, “the innocence of the defendant of the charge with which he was confronted has been established.”\(^10\)

Lawyers came to believe that the “shutting out of his sworn evidence could be positively hurtful to the accused, and that innocence was in fact aided, not prejudiced, by the opportunity of the accused to testify under oath.”\(^11\) The defendant, wrote the

---

6. See id. at 580–81 (“Experience under the American competency statutes was to change the minds of many who had opposed them. It was seen that the shutting out of his sworn evidence could be positively hurtful to the accused . . . .”).
7. See Act of Mar. 25, 1864, ch. 280, 1864 Me. Laws 214 (codified as amended at ME. REV. STAT. tit. 15, § 1315 (2011)) (providing that defendants are competent witnesses in prosecutions for all crimes). Prior to enacting the 1864 law, Maine enacted a competency law in 1859. See Act of Apr. 2, 1859, ch. 104, 1859 Me. Laws 97 (providing that defendants are competent witnesses in prosecutions for certain minor crimes); see also Ferguson, 365 U.S. at 577 (describing Maine’s adoption of competency statutes in 1859 and 1864 and similar actions by other states).
10. Id. at 66; see also Rosen v. United States, 245 U.S. 467, 471 (1918)
[T]he truth is more likely to be arrived at by hearing the testimony of all persons of competent understanding who may seem to have knowledge of the facts involved in a case, leaving the credit and weight of such testimony to be determined by the jury or by the court, rather than by rejecting witnesses as incompetent, with the result that this principle has come to be widely, almost universally, accepted in this country and in Great Britain.
11. Ferguson v. Georgia, 365 U.S. 570, 580–81 (1961). The Court made a similar statement in a case decided after the common law rule was abrogated:
This rule, while affording great protection to the accused against
WHY STRICKLAND IS THE WRONG TEST

Court, “above all others may be in a position to meet the prosecution’s case.”\textsuperscript{12} \textit{Ferguson} invalidated a Georgia statute that limited a defendant’s presentation to unsworn statements without questioning from counsel.\textsuperscript{13} The Court rested on the Assistance of Counsel Clause.\textsuperscript{14} It implied, but did not recognize, a right to testify under the Constitution.\textsuperscript{15}

That task was left to \textit{Rock v. Arkansas}\textsuperscript{16} in 1987. “At this point in the development of our adversary system,” wrote Justice

unfounded accusation, in many cases deprived him from explaining circumstances tending to create conclusions of his guilt which he could readily have removed if permitted to testify. To relieve him from this embarrassment the law was passed. In mercy to him, he is by the act in question permitted upon his request to testify in his own behalf in the case. In a vast number of instances the innocence of the defendant of the charge with which he was confronted has been established.


\textsuperscript{12} \textit{Ferguson}, 365 U.S. at 582.

\textsuperscript{13} \textit{See id.} at 596 (“We therefore hold that, in effectuating the provisions of § 38-415, Georgia, consistently with the Fourteenth Amendment, could not, in the context of § 38-416, deny appellant the right to have his counsel question him to elicit his statement.”).

\textsuperscript{14} \textit{Id.}; U.S. \textit{CONST.} amend. VI (“In all criminal prosecutions, the accused shall enjoy the right to . . . have the Assistance of Counsel for his defense.”).

\textsuperscript{15} \textit{See id.} at 572 n.1, 596 (crafting a narrow holding). Justice Clark, however, concurring, thought this was \textit{precisely} what the Court was doing, as a matter of logical necessity, and that it was needlessly formulaic and foolish not to acknowledge this. \textit{See id.} at 602 (Clark, J., concurring) (“Reaching the basic issue of incompetency . . . I do not hesitate to state that in my view § 38-416 does not meet the requirements of due process and that, as an unsatisfactory remnant of an age gone by, it must fall as surely as does its palliative, § 38-415.”). In \textit{Fowle v. United States}, the Ninth Circuit relied on Justice Clark’s concurrence to find a constitutional right, apparently because all the judges agreed that such a right existed. \textit{See Fowle v. United States}, 410 F.2d 48, 53 (9th Cir. 1969) (“The Government says that if [the defendant] had wished to avoid all adverse inferences which might be drawn from his original silence in reliance on his constitutional rights, he should have sacrificed his \textit{constitutional right to testify in his own defense.” (emphasis added)).

\textsuperscript{16} \textit{Rock v. Arkansas}, 483 U.S. 44 (1987) (“The right to testify on one’s own behalf at a criminal trial has sources in several provisions of the Constitution. It is one of the rights that [is] essential to due process of law in a fair adversary process.” (citation omitted)). For a typical pre-\textit{Rock} discussion of the right in a U.S. Court of Appeals, see \textit{Sims v. Lane}, 411 F.2d 661, 664 (7th Cir. 1969), in which the Seventh Circuit stated:

In the federal courts, the privilege of an accused to testify in his own defense is merely statutory, abrogating the common law rule of incompetence . . . . No case has been brought to our attention to
support petitioner’s contention that the Fourteenth Amendment accords a defendant in a state court a federal constitutional right to testify. To the contrary, the federal rule seems to be that the exercise of this right is subject to the determination of competent trial counsel and varies with the facts of each case.

On the other hand, United States v. Von Roeder seemed to suggest that if a trial judge went too far in dissuading a willing defendant from testifying (say, by pointing out the dangers of cross-examination or perjury), this might violate a constitutional right or result in some other impropriety. See United States v. Von Roeder, 435 F.2d 1004, 1008–09 (10th Cir. 1970) (analyzing a judge’s role in notifying a defendant of the possible effects of testifying at trial). Winters v. Cook stated that there exist certain “inherently personal fundamental right[s]” that “can be waived only by the defendant and not by his attorney,” which include the “right to testify personally.” Winters v. Cook, 489 F.2d 174, 179 (5th Cir. 1973). In United States v. Poe, the D.C. Circuit reversed a jury verdict, in accord with the ruling of the district court, which felt the trial was unfair because defense counsel gave legally incorrect advice about impeachment. See United States v. Poe, 352 F.2d 639, 640–41 (D.C. Cir. 1965) (“The trial judge found that appellant was deprived of a fair trial because he was misinformed as to the consequences of taking the stand to deny the charges against him.”). But otherwise lawyers are “free to keep defendants from testifying whenever counsel see fit. Any suggestion to the contrary is chimerical.” Id.

Courts of decades past were also in the habit of holding that if a defendant retained his choice of counsel, no error by that attorney, however grave, could constitute “a denial of due process chargeable to the state.” United States ex rel. Darcy v. Handy, 203 F.2d 407, 425–26 (3d Cir. 1953); see also Hudspeth v. McDonald, 120 F.2d 962, 967–68 (10th Cir. 1941) (stating that counsel was drunk for most of trial and finding no error).

Yet, in the 1970s, a number of states held that the right to testify existed. See, e.g., State v. Noble, 514 P.2d 460, 461 (Ariz. 1973) (“It is well established that in criminal prosecutions an accused has the right to testify in his own behalf.” (citations omitted)); People v. Robles, 466 P.2d 710, 716 (Cal. 1970) (“We are satisfied that the right to testify in one’s own behalf is of such fundamental importance that a defendant who timely demands to take the stand contrary to the advice given by his counsel has the right to give an exposition of his defense before a jury.”). The Supreme Court of Alaska, in Hughes v. State, stated:

[I]t is preferable that a defendant be permitted to testify if he so requests. The right to testify in one’s own behalf is often of vital importance in a trial. No defendant requesting to testify should be deprived of exercising that right and conveying his version of the facts to the court or jury, regardless of competent counsel’s advice to the contrary.

Hughes v. State, 513 P.2d 1115, 1119 (Alaska 1973). But it did not explicitly ground this in any constitution, state or federal, and found waiver of the right in any event. See id. at 1120 (“In the instant case, based on all of the testimony we conclude that Hughes knowingly and intelligently waived his right to testify.”).

Other states held that a criminal defendant had no constitutional right to testify on his own behalf. See, e.g., State v. Hutchinson, 458 S.W.2d 553, 554 (Mo. 1970) (referencing Ferguson and stating that “the assumption that the
Blackmun, “it cannot be doubted that a defendant in a criminal case has the right to take the witness stand and to testify in his or her own defense.”17 The right had its “source” in three provisions.18 The Due Process Clause19 secures “an opportunity to be heard . . . a right to [a] day in court.”20 The Compulsory Process Clause21 guarantees an accused’s right to call witnesses in his favor—and the “most important witness . . . in many criminal cases is the defendant himself.”22 Finally, the right to testify was seen as a “necessary corollary” of the Self-
Incrimination Clause, because every defendant is “privileged to testify in his own defense, or to refuse to do so.”

Rock relied heavily on Faretta v. California, the case from 1975 that established the constitutional right of self-representation. Faretta read the Sixth Amendment to secure a “personal” right to represent oneself: “It is the accused, not counsel, who must be ‘informed of the nature and cause of the accusation,’ who must be ‘confronted with the witnesses against him,’ and who must be accorded ‘compulsory process for obtaining witnesses in his favor.’” It is the “defendant, and not his lawyer or the State, [who] will bear the personal consequences of a conviction.” And though it may be “undeniable” that most defendants could “better defend with counsel’s guidance than by their own unskilled efforts,” wrote Justice Stewart, “his choice must be honored out of ‘that respect for the individual which is the lifeblood of the law.’” The right to self-representation was thus grounded in personal autonomy; the Court in Rock emphatically recognized the right to testify in that light. Rock not only drew on Faretta, but it explicitly stated: “Even more fundamental to a personal defense than the right of self-representation . . . is an accused’s right to present his own version of events in his own words.”

---

23. U.S. Const. amend. V (“No person shall be . . . compelled in any criminal case to be a witness against himself . . . .”).


25. Faretta v. California, 422 U.S. 806 (1975); see also Rock, 483 U.S. at 51–52 (discussing Faretta).

26. See Faretta, 422 U.S. at 807 (ruling that a state may not “constitutionally hale a person into its criminal courts and there force a lawyer upon him, even when he insists that he wants to conduct his own defense”).

27. U.S. Const. amend. VI (“In all criminal prosecutions, the accused shall enjoy the right to . . . be confronted with the witnesses against him; to have compulsory process for obtaining witnesses in his favor, and to have the Assistance of Counsel for his defense.”).

28. Faretta, 422 U.S. at 819.

29. Id. at 834.


If the accused’s right to testify is unquestionably a fundamental constitutional right, grounded in personal autonomy, what is the remedy for violations of it? This Article considers the virtually uniform assessment by federal courts that the denial of the right to testify is a product of faulty lawyering and as such the remedy lies in Strickland v. Washington\(^{32}\) and its progeny—under which the defendant gets a retrial only if he can show that had he testified, there is a “reasonable probability” that the outcome of his trial would have been different.\(^{33}\) This Strickland “prejudice” standard has been nearly impossible for defendants to meet—resulting in a constitutional right without a remedy.

**B. The Right to Testify as Part of the Autonomy Line**

When the Supreme Court in *Rock* relied on *Faretta*, it placed the right to testify in what one might call the “free choice” line of cases.\(^{34}\) The principle behind these holdings is that certain rights are personal to the defendant—a matter of “dignity” and “autonomy” rather than “strategy” or “tactics”—which only the defendant, not his lawyer, can waive, and then only if knowingly and voluntarily.\(^{35}\) *Faretta* was one instance of such “[f]reedom of explicit by federal statute” and is “implicit in the Constitution”); see also Riggins v. Nevada, 504 U.S. 127, 144 (1992) (Kennedy, J., concurring) (“It is well established that the defendant has the right to testify on his own behalf, a right we have found essential to our adversary system.”). One circuit called *Rock’s* holding a “dictum” and assumed without deciding that the right to testify was “fundamental.” See Lema v. United States, 987 F.2d 48, 52 n.3 (1st Cir. 1993) (listing circuit court cases recognizing the right as fundamental).

33. *Id.* at 694.
34. See, e.g., United States v. Hung Thien Ly, 646 F.3d 1307, 1313 (11th Cir. 2011)

Although often framed as a right to testify, it is more properly framed as a right to choose whether to testify. The “choice” concept reflects the competing considerations that make up this right; while the Fifth Amendment gives the accused the right to remain silent, courts have recognized that the accused also has the absolute right to break his silence and to testify.

35. See *id.* (“This right to choose is personal as well as fundamental, and the defendant must make this decision himself.”).
choice." So, too, is the “profound’ choice” about whether to stand trial or plead guilty (Cooper v. Oklahoma) or the personal “choice of the petitioner” in deciding whether to appeal (Fay v. Noia). Other cases in this line include Brookhart v. Janis (a lawyer cannot “overrid[e] his client’s desire” to exercise his “personal” right to confront witnesses); McKaskle v. Wiggins (right to self-representation “affirm[s] the dignity and autonomy of the accused” and can be undermined by intrusive standby counsel); Adams v. United States (defendant can insist on bench trial as matter of “free choice by a self-determining individual,” which to deny is to

36. See Faretta v. California, 422 U.S. 806, 834 n.45 (1975) (citing the constitutional protection afforded to a defendant’s free choice to testify in a criminal proceeding). For a good illustration of how the right is grounded in autonomy and not trial result, see Johnstone v. Kelly, 812 F.2d 821 (2d Cir. 1987). See also infra notes 185–89 and accompanying text.


38. Fay v. Noia, 372 U.S. 391, 439–40 (1963); see also Wainwright v. Sykes, 433 U.S. 72, 92 (1977) (Burger, C.J., concurring) (noting that in Fay the “critical procedural decision—whether to take a criminal appeal—was entrusted to a convicted defendant” and that the case’s touchstone was “the exercise of volition by the defendant himself with respect to his own federal constitutional rights”); Anders v. California, 386 U.S. 738, 744 (1967) (stating that counsel must support the defendant’s appeal to the best of counsel’s ability).

The Supreme Court said as recently as 1983 that “[t]here is, of course, no constitutional right of appeal,” Jones v. Barnes, 463 U.S. 745, 751 (1983), though it is hard to imagine that if such a question arose today, in a case in which a substantial liberty or property interest was at stake, a majority of the Court would care to say so again. For an illuminating discussion of this proposition, see Shifflett v. Virginia, 447 F.2d 50 (4th Cir. 1971). The issue was the extent to which lawyers must explain the rights of appeal to a client in order to enable the client to make an informed decision about whether to exercise the right. See id. at 53–54 (“To assure that the decision to take or forego an appeal would depend only on the defendant’s own informed choice, we required in Nelson that he be given complete information, by his lawyer or by the court, about his right to appeal . . . .”).


40. Id. at 7. The Court seemed to be relying on the defendant’s rights to cross-examine and confront witnesses against him, opportunities that a guilty plea (or “prima facie” trial—apparently an Ohio state procedure of convenience) would of course foreclose.


42. Id. at 176–77.

“imprison a man in his privilege and call it the Constitution”); 44 and United States v. Gonzalez-Lopez 45 (defendant usually has a “right to counsel of choice”). 46 These defendant-only decisions are not among the multitude entrusted to counsel—even if the accused’s choice is inimical to his best interests in view of the bench and bar. 47 Thus the Tenth Circuit wrote that a lawyer “lacks authority to prevent a defendant from testifying in his own defense, even when doing so is suicidal trial strategy.” 48 “If a defendant insists on testifying,” said the Seventh Circuit, “however irrational that insistence might be from a tactical viewpoint, counsel must accede.” 49

44. Id. at 280–81; see also Patton v. United States, 281 U.S. 276, 298 (1930) (affirming a defendant’s choice to waive his right to a trial by jury composed of twelve members).
46. See id. at 147 (“The right to select counsel of one’s choice, by contrast, has never been derived from the Sixth Amendment’s purpose of ensuring a fair trial. It has been regarded as the root meaning of the constitutional guarantee.”). But see infra note 206 and accompanying text (discussing Gonzalez-Lopez and Wheat).
47. See Jones v. Barnes, 463 U.S. 745, 751 (1983)
   It is also recognized that the accused has the ultimate authority to make certain fundamental decisions regarding the case, as to whether to plead guilty, waive a jury, testify in his or her own behalf, or take an appeal, . . . . [though an] indigent defendant [does not] ha[ve] a constitutional right to compel appointed counsel to press nonfrivolous points requested by the client, if counsel, as a matter of professional judgment, decides not to present those points. See also Faretta v. California, 422 U.S. 806, 834 (1975) (“And although he may conduct his own defense ultimately to his own detriment, his choice must be honored . . . .”).
48. Cannon v. Mullin, 383 F.3d 1152, 1171 (10th Cir. 2004); see also United States v. Teague, 953 F.2d 1525, 1533 (11th Cir. 1992) (“The wisdom or unwisdom of the defendant’s choice does not diminish his right to make it.” (quoting Wright v. Estelle, 572 F.2d 1071, 1079 (5th Cir. 1978) (Godbold, J., dissenting))); Ortega v. O’Leary, 843 F.2d 258, 261 (7th Cir. 1988) (“If a defendant insists on testifying, no matter how unwise such a decision, the attorney must comply with the request.”). Such a “suicidal” strategy may have been undertaken in People v. Robles, where it appeared that the defendant was perfectly indifferent to the success of his defense. See People v. Robles, 466 P.2d 710, 716–18 (Cal. 1970) (describing how the defendant insisted on testifying in his own defense, was impeached by prior convictions, made damaging admissions, engaged in offensive outbursts, and told jurors that they were “just a bunch of fools”).
49. United States v. Curtis, 742 F.2d 1070, 1076 (7th Cir. 1984).
II. Why Strickland Should Not Be Applied in Right-to-Testify Cases

A. The Results of Applying Strickland to Right-to-Testify Claims

Allegations by a defendant that his right to testify was coercively abridged are common. The culprit is usually a lawyer who said something like, “I make all decisions concerning this case and I say you’re not going to testify.”50 At other times the lawyer is alleged to have told the accused that he would call him and then rested the case before doing so.51 Almost all such allegations are brought on habeas petitions under the Antiterrorism and Effective Death Penalty Act52—with its forbidding “doubly deferential” standards53—because discussions between counsel and client are usually outside the record. They occur in whispers at counsel table, in prison chambers, hallways, or lawyers’ offices.54

50. Cannon v. Mullin, 383 F.3d 1152, 1171 (10th Cir. 2004).
51. See United States v. Tavares, 100 F.3d 995, 996–97 (D.C. Cir. 1996) (stating that defendant by all appearances planned to testify, fell ill on the day he was to do so, and was persuaded by counsel not to testify). The defendant argued that counsel might have sought a continuance. Id. at 997.
53. See Morris v. Sec’y, Dep’t of Corr., 677 F.3d 1117, 1126 (11th Cir. 2012) (stating that “as a federal habeas court we are not applying Strickland de novo, but rather through the additional prism of AEDPA deference” and determining that “under this doubly deferential standard, [t]he pivotal question is whether the state court’s application of the Strickland standard was unreasonable” (citations omitted)). For the phrase “doubly deferential,” see Knowles v. Mirzavance, 556 U.S. 111, 123 (2009) (citing Yarborough v. Gentry, 540 U.S. 1, 5–6 (2003) (per curiam)).
54. See Massaro v. United States, 538 U.S. 500, 504–05 (2003) (“When an ineffective-assistance claim is brought on direct appeal, appellate counsel and the court must proceed on a trial record not developed precisely for the object of litigating or preserving the claim and thus often incomplete or inadequate for this purpose.”); see also Guinan v. United States, 6 F.3d 468, 476 (7th Cir. 1993) (Easterbrook, J., concurring) (“Lawyers who raise ineffective-assistance claims on direct appeal do their clients a grave disservice, because the inevitable loss will prevent the accused from raising the same claim later, when factual development would permit accurate resolution.”). The Supreme Court in Massaro adopted this reasoning. See Massaro, 538 U.S. at 504 (“In light of the way our system has developed, in most cases a motion brought under § 2255 is preferable to direct appeal for deciding claims of ineffective assistance.”).
Federal courts are loath to entertain these claims, usually because the facts are that the defendant was advised to keep quiet, wisely heeded his lawyers, and now regrets it. But courts do look askance at instances in which a lawyer really seems to have silenced a defendant against his will. The inquiry is conducted under *Strickland v. Washington*, which states the test for whether a lawyer was constitutionally adequate under the Assistance of Counsel Clause.55 Why *Strickland*? The most-cited case applying a *Strickland* analysis to deprivation of the right to testify is *United States v. Teague*, an Eleventh Circuit case from 1992.57 But that circuit (and the Fifth Circuit, from which it was

55. See *Strickland v. Washington*, 466 U.S. 668, 687 (1984) (stating that a convicted defendant must show that his counsel’s performance was deficient and that this deficiency prejudiced the defendant in some way).

56. United States v. Teague, 953 F.2d 1525 (11th Cir. 1992) (en banc).

57. See id. at 1534 (“[W]e believe the appropriate vehicle for claims that the defendant’s right to testify was violated by defense counsel is a claim of ineffective assistance of counsel under *Strickland* . . . .”). As it happened, the original panel that heard the case applied harmless-error review. See United States v. Teague, 908 F.2d 752, 757 (11th Cir. 1990), vacated, 932 F.2d 899 (11th Cir. 1991) (“Thus, the court concluded, because Teague’s testimony would have been largely duplicative, any error in not allowing him to testify was harmless.”). Before *Teague*, but after *Strickland*, that court reviewed such claims under *Chapman v. California* as a sort of independent constitutional violation of the right to testify, independent of any ineffectiveness claim. See *Chapman v. California*, 386 U.S. 18, 24 (1967) (holding “that before a federal constitutional error can be held harmless, the court must be able to declare a belief that it was harmless beyond a reasonable doubt”); see also Nichols v. Butler, 917 F.2d 518, 521 n.2 (11th Cir. 1990), vacated, 932 F.2d 900 (11th Cir. 1991) (finding that a *Strickland* analysis was unnecessary because “such conduct by defense counsel amounts to a violation of the right to testify regardless of whether it also amount to ineffectiveness of counsel”). Because this case was vacated after an en banc poll and redecided after *Teague*, *Strickland* now applied. See Nichols v. Butler, 953 F.2d 1550, 1552–54 (11th Cir. 1992) (en banc) (applying *Strickland*).

The earliest case we find to connect *Strickland* and the right to testify is *United States v. Curtis*, decided a few months after *Strickland*. See United States v. Curtis, 742 F.2d 1070, 1074–76 (7th Cir. 1984) (citing the *Strickland* presumption that a decision not to testify might be considered sound trial strategy). The first post- *Rock* case we find specifically applying *Strickland* to a right-to-testify claim is *Isom v. Lockhart*. See *Isom v. Lockhart*, 847 F.2d 484, 486 (8th Cir. 1988) (applying the two-prong *Strickland* test, which requires objectively unreasonable representation by counsel and a reasonable likelihood that the result of the proceeding would have been different but for this representation). It was a habeas case in which the court applied *Strickland* without much discussion. *Id.* The other claim involved the lawyer’s failure to dismiss the venire after three veniremen—asked by the defendant’s attorney
split) skirmished for years over how, judicially, to enforce the right to testify, reaching a high point in Wright v. Estelle, a brilliant en banc battle royale in 1978, and as eloquent a discussion on the subject as can be found in the Federal Reporter.

In that case Archie Wright, on trial for murder and robbery, was told by his lawyer that if he, Wright, testified, the lawyer would withdraw; the attorney felt Wright’s version did not match up with other accounts and that if the jury thought he, Wright, was lying, he would get a death sentence. The circuit judges split into two camps and put the case strikingly, worth relating at length because they so capably capture the considerations surrounding protection of the right to testify. One group, led by Judge Thornberry, acknowledged that a defendant of course has whether they understood that her client had the right both to testify and to refuse to do so—said things like “I understand the law but I think if I was in it I would want to testify . . . . I think I would be man enough to want to tell my side.” Id. at 485. This latter claim is a classic Strickland question; perhaps the court simply did not consider whether a different standard might apply to the former claim.

The Ninth Circuit appears to have been second: in United States v. Hood, it applied Strickland because the “right to testify certainly may implicate the sixth amendment”—in a case in which the defendant argued that bad advice dissuaded him from testifying to his advantage. United States v. Hood, No. 88-4046, 1989 WL 102017, at *4 (9th Cir. Aug. 25, 1989). Galowski v. Murphy provides another early statement, but it appears that the petitioner also tried to assert the claim as an “independent constitutional violation,” but the claim was barred for not having been asserted before. Galowski v. Murphy, 891 F.2d 629, 636 n.12 (7th Cir. 1989).

The district court in United States v. Butts, an impassioned decision given shortly before Rock, rejected both parties’ argument that right-to-testify claims were Strickland questions. See United States v. Butts, 630 F. Supp. 1145, 1148 (D. Me. 1986) (“The Court does not find Strickland applicable in this case.”). Judge Carter said this was because the attorney’s conduct was “troublesome not just for its possible impact on the reliability of the verdict, i.e., for its Sixth Amendment implications,” but because of concerns for a fair trial under the Fifth Amendment and the right to meet and deny accusations against a defendant under the Sixth Amendment. Id. He also thought “a defendant’s right to testify in a criminal proceeding against him [is] so basic to a fair trial that its infraction can never be treated as harmless error, which is in essence the inquiry required to be made by the second prong—prejudice to the defense—of Strickland.” Id.

58. Wright v. Estelle, 572 F.2d 1071 (5th Cir. 1978) (en banc).
59. Wright v. Estelle, 572 F.2d 971, 972–74 (5th Cir. 1977), aff’d en banc, 572 F.2d 1071 (5th Cir. 1978).
the right to reject or accept a lawyer to conduct his defense. But once the defendant chooses to entrust his life and liberty to a person versed in the law, that delegation includes the decision as to whether the client testifies:

Trial attorneys are professional artisans working in a highly competitive arena that requires all the skills which education, training, and experience have given them. Criminal defendants are entitled to no less. A defendant has a right to necessary surgery, but he does not have the right to require the surgeon to perform an operation contrary to accepted medical practice. If, despite his counsel’s advice, a defendant continues to believe that his testimony is more important than the continued services of an attorney who insists he should not take the stand, the conflict must be resolved by the court. Only in this way may the right to testify be reconciled with the right to effective assistance of counsel . . . .

While *Faretta* allows a defendant to have a fool for a client . . . , there is nothing in its logic that commands that the defendant may also have a fool for an attorney . . . . [T]he decision whether to testify is properly allocated to the defendant’s attorney and not to the defendant. An attorney is not necessarily ineffective if he determines not to allow his client to testify, even though he should give great deference to a defendant’s desire to testify[;] however, we are here concerned with constitutional requirements and there is no constitutional requirement that a court-appointed attorney must walk his client to the electric chair.

Judge Godbold answered for himself and two others, relying on a personal autonomy argument:

The rationale of *Faretta v. California* and its precursors, relating to the right of the accused to defend himself, leads to the conclusion that the right to testify is a fundamental right reserved to the defendant for decision. In making the choice on whether to testify, just as the choice on whether to represent himself, the defendant elects whether to become an active participant in the proceeding that affects his life and liberty and to inject his own action, voice and personality into the process to the extent the system permits.

---

60. See Wright, 572 F.2d at 1073 (Thornberry, J., specially concurring) ("The defendant, of course, has the authority in the first instance to accept or reject court-appointed representation.").

61. Id. at 1073–74 (citation omitted).
In the narrow world of the courtroom the defendant may have faith, even if mistaken, in his own ability to persuasively tell his story to the jury. He may desire to face his accusers and the jury, state his position, and submit to examination. His interest may extend beyond content to the hope that he will have a personalized impact upon the jury or gain advantage from having taken the stand rather than to seek the shelter of the Fifth Amendment. Or, without regard to impact upon the jury, his desire to tell “his side” in a public forum may be of overriding importance to him. Indeed, in some circumstances the defendant, without regard to the risks, may wish to speak from the stand, over the head of judge and jury, to a larger audience. It is not for his attorney to muzzle him.

Indeed, our history is replete with trials of defendants who faced the court, determined to speak before their fate was pronounced: Socrates, who condemned Athenian justice heedless of the cup of hemlock; Charles I, who challenged the jurisdiction of the Cromwellians over a divine monarch; Susan B. Anthony, who argued for the female ballot; and Sacco and Vanzetti, who revealed the flaws of their tribunal. To deny a defendant the right to tell his story from the stand dehumanizes the administration of justice. I cannot accept a decision that allows a jury to condemn to death or imprisonment a defendant who desires to speak, without ever having heard the sound of his voice.62

Fourteen years later, in 1992, the Eleventh Circuit, sitting en banc in Teague, revisited the question. Without any considered discussion, the court simply settled on this:

Because it is primarily the responsibility of defense counsel to advise the defendant of his right to testify and thereby to ensure that the right is protected, we believe that the appropriate vehicle for claims that the defendant’s right to testify was violated by defense counsel is a claim of ineffective assistance of counsel under Strickland . . . .63

No authority was cited. Rock, decided five years earlier, certainly never invoked the Assistance of Counsel Clause in support of the right to testify.64 Nonetheless, other circuits, largely relying on

62. Id. at 1077–78 (Godbold, J., dissenting).
63. United States v. Teague, 953 F.2d 1525, 1534 (11th Cir. 1992) (en banc).
64. See Rock v. Arkansas, 483 U.S. 44, 51–53 (1987) (indicating that the
WHY STRICKLAND IS THE WRONG TEST

Teague, were persuaded to adopt Strickland because the choice of whether to testify is usually made in consultation with the attorney.\textsuperscript{65} Today every circuit uses Strickland to evaluate right-to-testify claims, which can range from allegations of a mere right to testify on one's own behalf in a criminal trial is grounded in the Constitution's Compulsory Process Cause, Due Process Clause, and Self-Incrimination Clause).

\textsuperscript{65} See United States v. Espinoza, 392 F. App'x 666, 668 (10th Cir. 2010) (“Because the choice whether to testify is often made in consultation with an attorney, violations of the right to testify are 'best treated' as ineffective assistance of counsel claims.” (citing Cannon v. Mullin, 383 F.3d 1152, 1170 (10th Cir. 2004))). The panel in United States v. Tavares seemed to apply a sort of heightened Strickland test:

A more reasonable approach, and one in keeping with Strickland's two-part test, is to continue to assign special significance to the defendant's precluded right to testify and at the same time to inquire whether it is reasonably probable that the defendant's testimony would have changed the outcome of the trial in his favor.

United States v. Tavares, 100 F.3d 995, 998 (D.C. Cir. 1996).

Yet as late as 1998, the Eighth Circuit was willing to assert that, when the error is defense counsel's, “it is unclear if harmless error analysis applies to the denial of a defendant's right to testify.” Frey v. Schuetzle, 151 F.3d 893, 898 n.3 (8th Cir. 1998); see also United States v. Leggett, 162 F.3d 237, 248 n.11 (3d Cir. 1998) (emphasizing that it is a matter of judicial discretion to determine whether a harmless-error analysis should be applied to actions committed by the district court). These musings have been abandoned. See infra note 66.

In a few published (and still cited) opinions, the Seventh Circuit applied a harmless-error test when the allegation of abridgment of the right was directed at defense counsel. In Ortega v. O'Leary, the panel wrote that it "has previously ruled that the Chapman standard applies when a petitioner has been denied the right to testify." Ortega v. O'Leary, 843 F.2d 258, 262 (7th Cir. 1988) (citing Alicea v. Gagnon, 675 F.2d 913, 923 (7th Cir. 1982) (per curiam)). Curiously, years later, in Ward v. Sternes, that Circuit again applied Chapman's harmless-error standard (as had the district court under review) and found that:

[H]ad the jury been given the opportunity to observe Ward testify while sober yet still exhibiting these signs of his mental deficiencies, it is conceivable that the jury would have given more credence to the expert psychiatric testimony and particularly Dr. Traugott's opinion that Ward's brain injury alone, regardless of his intoxication, would have rendered him incapable of conforming his actions to the law. On this close question, the inability to hear Ward testify was not harmless error.

Ward v. Sternes, 334 F.3d 696, 708 (7th Cir. 2003). But the discussion in Alexander v. United States suggested that in most cases the circuit applies Strickland. See Alexander v. United States, 219 F. App'x 520, 523 (7th Cir. 2007) (applying Strickland).
failure to adequately inform a defendant about his right, all the way to instances of genuine coercion that prevent its exercise.\textsuperscript{66}

What has been the result? \textit{Strickland} requires a petitioner (usually proceeding pro se) to prove deficient performance by the lawyer and prejudice to the client.\textsuperscript{67} The first prong usually involves a dispute over what the defendant told his lawyer about his desire to testify. Far more important is the “prejudice” prong. It is here that the right is rendered a nullity. Under \textit{Strickland}’s second prong, the petitioner must show a “reasonable probability

\textsuperscript{66} See Owens v. United States, 483 F.3d 48, 57–59 (1st Cir. 2007) (applying \textit{Strickland} to a defendant’s ineffective assistance of counsel claim relating to counsel’s alleged failure to inform defendant of his right to testify at trial); Brown v. Artuz, 124 F.3d 73, 74 (2d Cir. 1997)

We conclude that the decision whether a defendant should testify at trial is for the defendant to make, that trial counsel’s duty of effective assistance includes the responsibility to advise the defendant concerning the exercise of this constitutional right, and that the two-part test established in \textit{Strickland v. Washington}, should be used to assess a defendant’s claim that defense counsel rendered ineffective assistance by preventing him from testifying or at least failing to advise him concerning his right to testify.

\textit{See also} Palmer v. Hendricks, 592 F.3d 386, 397 (3d Cir. 2010) (“[E]very authority we are aware of that has addressed the matter of counsel’s failure to advise a client of the right to testify has done so under \textit{Strickland}’s two-prong framework.”); Sayre v. Anderson, 238 F.3d 631, 634 (5th Cir. 2001) (“Sayre contends only that his attorney interfered with his right to testify, not that the state trial court (or prosecutor) did so.”); Sexton v. French, 163 F.3d 874, 882 (4th Cir. 1998) (“[A] criminal defendant’s claim that his trial counsel was constitutionally ineffective because trial counsel failed to inform him of his right to testify or because trial counsel forced him to testify must satisfy the two-prong test established in \textit{Strickland v. Washington} . . . .”). In \textit{United States v. Brown}, the Fifth Circuit recently held that “[t]he appropriate vehicle for such claims is a claim of ineffective assistance of counsel.” United States v. Brown, 217 F.3d 247, 258–59 (5th Cir. 2000) (citation omitted); see also United States v. Hubbard, 638 F.3d 866, 870 (6th Cir. 2011) (applying \textit{Strickland}); Matylinsky v. Budge, 577 F.3d 1083, 1097 (9th Cir. 2009) (“The \textit{Strickland} standard is applicable when a petitioner claims his attorney was ineffective by denying him his constitutional right to testify.”); Cannon v. Mullin, 383 F.3d 1152, 1170 (10th Cir. 2004) (“We agree [with other circuits] that Mr. Cannon’s claim is best treated as an ineffective-assistance-of-counsel claim and analyze it as such.”); Rodriguez v. United States, 286 F.3d 972, 983 (7th Cir. 2002) (en banc) (applying \textit{Strickland}); United States v. Webber, 208 F.3d 545, 551 (6th Cir. 2000) (same); United States v. Teague, 953 F.2d 1525, 1534 (11th Cir. 1992) (citing \textit{Strickland}).

\textsuperscript{67} See \textit{Strickland v. Washington}, 466 U.S. 668, 687–88 (1984) (“A convicted defendant’s claim that counsel’s assistance was so defective as to require reversal of a conviction or death sentence has two components.”).
WHY STRICKLAND IS THE WRONG TEST

that, but for counsel's unprofessional errors, the result of the proceeding would have been different. 68 When the claim is deprivation of the right to testify truthfully, the defendant loses almost every time. The court usually finds some or all of the following: the accused would not have been found credible; 69 his testimony would have been cumulative; 70 he would have been exposed to impeachment with prior convictions; 71 the evidence

68. Id. at 694.

69. See, e.g., Morris v. Sec'y, Dep't of Corr., 677 F.3d 1117, 1130 (11th Cir. 2012) (finding that petitioner, a first-degree murderer of an elderly lady, “would not have been credible”); Alexander v. United States, 219 F. App’x 520, 524 (7th Cir. 2007) (“It is not reasonably likely that Alexander’s testimony, given his diminished credibility as a convicted felon, would have swayed the jury’s verdict in any way.”); United States v. Mullins, 315 F.3d 449, 456 (5th Cir. 2002) (“The difficulty is that a denial by Mullins from the stand would come at a high price. It would juxtapose a police officer whose account is supported by Mullins’s signed statement with a felon with a large incentive to lie.”).

70. See, e.g., Morris, 677 F.3d at 1130 (finding that the defendant “would not have been credible in reasserting his innocence and that his proposed testimony would have been cumulative,” although the record was “unclear” about whether counsel told him of his right to testify); Washington v. Kemna, 16 F. App’x 528, 530 (8th Cir. 2001) (“Washington’s testimony at trial would have merely reiterated the alibi defense already provided through the trial testimony of his mother, Patricia Washington.”); United States v. Tavares, 100 F.3d 995, 998 (D.C. Cir. 1996) (assuming that defendant would have testified “absent his counsel’s actions” but finding that his “proposed testimony would have been largely cumulative and, to the extent it was not cumulative, largely peripheral”).

71. See, e.g., Dows v. Wood, 211 F.3d 480, 487 (9th Cir. 2000)

Moreover, [attorney] Egger had very good reason for suggesting that Dows not testify. Dows had three prior convictions for robbery and assault, which, in all likelihood, would have been admitted to impeach Dows on the stand if he had testified . . . . As noted with some irony by the trial court, if Egger had allowed Dows to testify under this scenario, that would be definite proof he was suffering from Alzheimer’s disease. Dows cannot prove deficient performance or prejudice based upon his failure to testify at trial.

See also Medley v. Runnels, 506 F.3d 857, 861–62 (9th Cir. 2007) (“Medley’s lawyer recommended that Medley not testify because he would have been impeached by his prior convictions and statements he made during a lengthy interview he gave police, which apparently were inconsistent with what Medley intended to testify.”); Rodriguez v. United States, 286 F.3d 972, 983 (7th Cir. 2002) (en banc) (“First, Rodriguez cannot show that his counsel’s advice concerning the impeachment value of his prior crime was unreasonable.”).

For an example of the analysis under the old, pre-Rock state of the law, see Hudgins v. United States, 340 F.2d 391, 396 (3d Cir. 1965), in which the Third Circuit stated:

Any statements made by him in his testimony at the hearing upon
against him was “overwhelming”;\textsuperscript{72} or his testimony was weak and would not have helped.\textsuperscript{73} Often a court simply passes over the

\textsuperscript{72} See, e.g., United States v. Ailemen, 473 F. App'x 754, 755 (9th Cir. 2012) ("[E]ven if we assume that Ailemen was prevented from testifying by his attorneys, he has failed to show that he could have overcome the overwhelming evidence against him."); Battle v. Sirmons, 304 F. App'x 688, 693 (10th Cir. 2008) ("To counter the overwhelming weight of the evidence presented against him, Battle argues that his testimony would have explained away all of the state's evidence. Even if we entertain Battle’s arguments presented in brief, which reflect the 20/20 wisdom of hindsight, we cannot find prejudice."); Franklin v. United States, 227 F. App'x 856, 857 (11th Cir. 2007) ("Franklin's proposed testimony . . . would not have been credible nor would it have refuted the overwhelming evidence of his guilt. Franklin failed to establish prejudice because there was no reasonable probability that his self-serving testimony would have convinced the jury to reject the evidence and acquit him."); Donato v. United States, No. 98-2991, 208 F.3d 202, at *1–2 (2d Cir. 2000)

Donato wrote the court a letter dated June 16, 1998 stating that “the defendant never informed defense counsel, that the defendant did not wish to testify” and that “this decision not to testify was made by counsel Cohen, and the defendant did not have any decision input on this matter.” In its decision, the district court was apparently unaware that Donato had sent this letter, as it found that “Donato has not answered” the court's request for his response to Cohen's letter. Therefore, when the district judge made a finding of fact that Cohen's account was credible, he may have assumed that the issue was uncontested. Nonetheless, even assuming that there was a deficient performance by counsel denying Donato his right to testify, unless Donato's potential testimony could have established a defense, he cannot demonstrate prejudice.

\textsuperscript{73} See, e.g., United States v. Willis, 273 F.3d 592, 599 (5th Cir. 2001) (finding that the defendant’s “ineffective assistance of counsel claim [was] without merit”). The court, considering Strickland deficiency, concluded without discussing the allegation that “counsel interfered with his right to testify,” that the defendant had “not convincingly argued that his testimony would have assisted him at either the pretrial hearing or at trial[,]” and that the defendant did “not even address the viability of the countervailing tactical reasons that his counsel might have had for declining to call him to the stand.” Id. at 598. This, in our view, is not supposed to be part of the test.
WHY STRICKLAND IS THE WRONG TEST

A typical case might explain that even though counsel should not have rested after defendant expressed a wish to “contribute[] something new and substantive” to his defense, the evidence of his guilt in “possessing the 8,440 doses of LSD found in the record album in his house was so strong that there is no reasonable probability that his testimony would have altered the outcome of the trial.”

Dozens of cases have sentences like: “We need not address whether Lee’s counsel was deficient for failing to call Lee to testify, because Lee cannot show that he was prejudiced under Strickland,” or “[E]ven if Liriano could establish that her counsel’s conduct was deficient, she has made no showing of prejudice.” Frequently it is left entirely mysterious what the actual allegations against the attorney were. Was there

---

74. See Strickland v. Washington, 466 U.S. 668, 697 (1984) (“In particular, a court need not determine whether counsel’s performance was deficient before examining the prejudice suffered by the defendant as a result of the alleged deficiencies.”).

75. See, e.g., United States v. Ailemen, 473 F. App’x 754, 756 (9th Cir. 2012) (“Further, ‘[i]f it is easier to dispose of an ineffectiveness claim on the ground of lack of sufficient prejudice... that course should be followed.’” (quoting Strickland, 466 U.S. at 697)); United States v. Tavares, 100 F.3d 995, 997 (D.C. Cir. 1996) (“Although it does not expressly concede that Werdig’s performance was inadequate, the government does not contest Tavares’s arguments on this point. The only question before us is whether Tavares was prejudiced by his counsel’s actions—specifically by Werdig’s failure to ensure that Tavares had an opportunity to testify.”).

76. Tavares, 100 F.3d at 999.

77. Lee v. Culliver, 300 F. App’x 689, 690 (11th Cir. 2008).

78. United States v. Betancur, 84 F. App’x 131, 135 (2d Cir. 2004). See, e.g., Brown v. Artuz, 124 F.3d 73, 80 (2d Cir. 1997) (“We need not remand for a finding on this point because, even if Brown’s conclusory allegation raised an issue on the performance prong of Strickland... Brown cannot satisfy the prejudice prong of the Strickland test.”).

79. See, e.g., Hester v. United States, 335 F. App’x 949, 951–52 (11th Cir. 2009) (per curiam) (containing no discussion of what happened); Battle v. Sirmons, 304 F. App’x 688, 693 (10th Cir. 2008) (stating “[i]f Battle were prevented from testifying” but containing no discussion of what happened). It may be that no hearing or inquiry into the matter was conducted at the district court level and that these panels simply did not have any specifics in their record.
coercion? A threat by the lawyer to withdraw if the client testified? An unfulfilled promise to call the defendant?

But under Strickland’s prejudice prong, it does not matter that the accused was left in ignorance of his rights, misled, lied to, or ignored. It does not even matter that a defendant was coerced, threatened, cajoled, or otherwise improperly influenced into relinquishing his right. Such acts go to the deficient-performance prong; they are irrelevant to whether the accused suffered prejudice. Thus, in United States v. Mullins, a lawyer admitted that her client “repeatedly requested to testify, and that she ‘prevented’ him from doing so against his wishes.” Yet the court found no prejudice because the defendant’s testimony “would [have] come at high price” by opening him to impeachment and because it was doubtful that a felon would be credited over an officer. In Gross v. Knight, a lawyer testified, “I’m sure I just told him it wasn’t going to happen . . . . [It was] a kind of discussion I’d have with my 9-year-old about whether he’s going to clean his room.”

80. See, e.g., Nichols v. Butler, 953 F.2d 1550, 1553 (11th Cir. 1992) (“[T]o coerce his client into remaining silent by threatening to abandon him mid-trial goes beyond the bounds of proper advocacy.”).
81. See, e.g., United States v. Scott, 909 F.2d 488, 492 (11th Cir. 1990) (stating that defense “counsel moved to withdraw shortly after the prosecution rested its case” and that the trial court assumed that counsel “made the motion because she discovered that [the defendant] intended to commit perjury”).
82. This was the specific allegation in United States v. Burnell, but what actually happened between the client and his attorney was left undetermined because the court found that no prejudice was demonstrated. See United States v. Burnell, No. 11-8100, 2012 WL 1664124, at *1 (10th Cir. May 14, 2012) (“Because Mr. Burnell has not made an adequate showing of prejudice, the district court has not abused its discretion in denying him an evidentiary hearing on his ineffectiveness claim concerning the right to testify.”).
83. See, e.g., Foster v. Delo, 39 F.3d 873, 883–84 (8th Cir. 1994) (stating that trial lawyers admitted that they could not recall whether they informed the defendant that he was permitted to testify at the penalty phase of his capital trial).
84. See, e.g., Nichols, 953 F.2d at 1553 (stating that the defendant’s attorney threatened to abandon him mid-trial).
85. United States v. Mullins, 315 F.3d 449 (5th Cir. 2002).
86. Id. at 455.
87. Id. at 456.
88. Gross v. Knight, 560 F.3d 668 (7th Cir. 2008).
89. Id. at 670.
WHY STRICKLAND IS THE WRONG TEST

than negligible chance his testimony [as to remorse] would have resulted in a different outcome,” largely because it would only have added to a “vast amount of [other] mitigating evidence.”90

Defendants prevail on ineffectiveness claims only in those exceedingly rare instances where the government’s case is so weak that the accused’s failure to testify is found actually to render a result unreliable.91 Nichols v. Butler92 was a “very close” robbery case in which the only evidence against the defendant was an eyewitness who glimpsed him briefly.93 The court found that permitting the defendant to present his account “in his own words” would have allowed the jury to “weigh his credibility” against that of the witness.94 In Cannon v. Mullin,95 the “power of a face-to-face appeal” might have convinced a jury that what the prosecutor called murder was in fact manslaughter in self-defense.96 And in Canaan v. McBride,97 an attorney testified that it never “crossed [his] mind” to call his capital defendant during the penalty phase even though the testimony “would have been the only mitigating evidence the jury heard” and “may have

90. Id. at 673.
91. See, e.g., United States v. Lore, 26 F. Supp. 2d 729, 739–40 (D.N.J. 1998) (finding prejudice in counsel’s refusal to permit defendant to testify, reasoning that the evidence against defendant was weaker than against codefendants and that defendant could have provided a noncriminal explanation for the government’s alleged extortionate activity); Campos v. United States, 930 F. Supp. 787, 793–94 (E.D.N.Y. 1996) (finding prejudice in lawyer’s refusal to allow defendant to testify, noting that testimony could have made a difference because the government’s case turned almost entirely on the testimony of a DEA agent); see also United States v. Walker, 772 F.2d 1172, 1178–79 (5th Cir. 1985) (relying on the notion that the accused, with his knowledge of facts, his testimony, and his demeanor, is of “prime importance” in a trial in which the “very point” is to determine guilt); United States v. Irvin, 450 F.2d 968, 971 (9th Cir. 1971) (Kilkenny, J., dissenting) (“The facial expressions of a witness may convey much more to the trier of the facts than do the spoken words.”).
93. See id. at 1551 (“A store employee testified that he glanced at this man for ‘not even a second.’”).
94. Id. at 1554.
95. Cannon v. Mullin, 383 F.3d 1152 (10th Cir. 2004).
96. See id. at 1152 (finding that live testimony from the defendant may have been more persuasive to the jury than the recorded statement by the defendant that was offered at trial).
persuaded the jury to be lenient.\textsuperscript{98} To summarize the state of the law, one might say one of two things. At best, defendants have a right with a very unpromising remedy; at worst, the right to testify does not exist unless its denial renders the trial unfair.

Defense counsel causes most of these errors. But a trial court, a statute, or a prosecutor can also infringe the right. Rock itself involved an Arkansas rule of evidence that excluded hypnotically refreshed testimony.\textsuperscript{99} In such circumstances Strickland makes absolutely no sense.

If a source other than defense counsel causes a violation of the right to testify, reviewing courts apply a harmless-error standard,\textsuperscript{100} or if the defendant failed to object, plain-error review.\textsuperscript{101} Trial-court errors include instances where, for instance, a judge failed to correct an obvious misapprehension on the part of a pro se defendant who did not know that he had a right to testify in narrative form.\textsuperscript{102} Another court erred when it failed to make inquiry after an attorney stated that his client wished to remain silent, yet the defendant interrupted, calling the lawyer a liar and insisting, “I want to take the stand.”\textsuperscript{103} A court also erred when it circumscribed a defendant’s planned testimony on motive by ruling, incorrectly, that aspects of it were irrelevant.\textsuperscript{104}

\begin{itemize}
  \item \textsuperscript{98} Id. at 385–87.
  \item \textsuperscript{99} See Rock v. Arkansas, 483 U.S. 44, 54–56 (1987) (describing the rule of evidence that affected defendant’s right to testify and providing a balancing test to be applied in cases of conflict between a statute and the right to testify).
  \item \textsuperscript{100} See, e.g., United States v. Leggett, 162 F.3d 237, 248, 248 n.11 (3d Cir. 1998) (reviewing for harmlessness where the district court discouraged the defendant from testifying by advising him “if my son were on trial here, I would tell him to follow his lawyer’s advice[,]” and “[i]f I were on trial, I would follow my lawyer’s advice”).
  \item \textsuperscript{101} See, e.g., United States v. Lechner, 341 F. App’x 443, 447–48 (10th Cir. 2009) (using a plain-error standard to review the appropriateness of the trial court telling a defendant who did not object at trial “your failure to testify will not have any impact on my decision making whatsoever”).
  \item \textsuperscript{102} See United States v. Hung Thien Ly, 646 F.3d 1307, 1317 (11th Cir. 2011) (“By not informing Ly that he could testify in narrative form, the district court denied his right to choose whether to testify ‘knowingly and intelligently.’”).
  \item \textsuperscript{103} See Ortega v. O’Leary, 843 F.2d 258, 259–60, 259 n.1 (7th Cir. 1988) (describing the defendant’s attempts to testify and the trial judge’s failure to permit the testimony).
  \item \textsuperscript{104} See United States v. Leo, 941 F.2d 181, 194–95 (3d Cir. 1991) (finding that the court’s failure to provide the defendant with an “opportunity to explain
Finally, a common court error arises when a judge, concerned with perjury, discourages a defendant who indicated interest in testifying. A United States Attorney could infringe the right by engaging in misconduct like threatening a defendant with a perjury indictment without any basis to suspect an intent to lie. These cases show that using a Strickland analysis in evaluating a deprivation of the right to testify is misguided—the right denied is not the right to effective counsel, but the right to testify. And that right should have the same remedy as the denial of other constitutional rights.

105. See, e.g., United States v. Scott, 909 F.2d 488, 492 (11th Cir. 1990) (determining that the trial court erred in telling the defendant that he could keep his current lawyer or proceed pro se, forcing him to choose between the right to testify and the right to counsel). The court forced the defendant to make that choice after assuming that the defense lawyer’s motion to withdraw was occasioned by the concern that the defendant would commit perjury. Id. at 492 n.3.

106. See United States v. Davis, 974 F.2d 182, 185 (D.C. Cir. 1992) (describing a prosecutor’s request that the court “instruct Davis about the possibilities and penalties of a perjury charge were he to take the stand and lie”). And this, too, incidentally, is not structural error. See United States v. Simmons, 670 F.2d 365, 372 n.4 (D.C. Cir. 1982) (per curiam) (noting that the defendant must prove “substantial prejudice” to obtain a reversal of conviction on the grounds that the prosecutor deprived him of defense testimony by threatening a witness).
B. The Inconsistency Between the Autonomy-Based Right to Testify and Strickland’s Purpose of Ensuring “Reliability” of Trial Results

The problem is that Strickland’s concern is with whether a lawyer’s incompetence was so egregious that the “trial cannot be relied on as having produced a just result.” By contrast, “personal” rights, like taking the stand, involve a defendant’s autonomy, his day in court, the right to meet his accusers himself, the chance to participate in settling his own legal fate, and the notion (as Justice Frankfurter wrote) that the “most persuasive counsel may not be able to speak for a defendant as the defendant might, with halting eloquence, speak for himself.” These considerations, under this “free choice” line, hold even if a defendant’s election—like the choice to proceed prose—is an ill-advised one that makes an unreliable conviction more likely. Courts consider an adverse result to be fair because the accused, so long as his lawyer has reasoned with him about the perils, alone bears the consequences of his choice. It is an irony that a right in which outcome is irrelevant is reviewed under a test that focuses on outcome.

Strickland’s inquiry asks whether a lawyer fails his client, yet suppressing a clear constitutional right seems ipso facto failure. Strickland requires us to “indulge a strong

109. See Faretta v. California, 422 U.S. 806, 834 (1975) (“The defendant, and not his lawyer or the State, will bear the personal consequences of a conviction” and so his choices regarding his defense, even if “ultimately to his own detriment[,] . . . must be honored out of “that respect for the individual which is the lifeblood of the law”).
110. Theoretically, Strickland might remain appropriate in true ineffectiveness cases: say, a defendant claims counsel was deficient in failing to inform him of his right to testify. See Sexton v. French, 163 F.3d 874, 882 (4th Cir. 1998) (“Because the burden of ensuring that a criminal defendant is informed of the nature and existence of the right to testify rests upon trial counsel, the burden shouldered by trial counsel is a component of effective assistance of counsel.”); see also Palmer v. Hendricks, 592 F.3d 386, 397 (3d Cir. 2010) (“[E]very authority we are aware of that has addressed the matter of counsel’s failure to advise a client of the right to testify has done so under
presumption that counsel’s conduct falls within the wide range of reasonable professional assistance,”111 but can a lawyer make a valid choice to violate his client’s prerogative?112 Strickland demands that the defendant “overcome the presumption that, under the circumstances, the challenged action ‘might be considered sound trial strategy,’”113 yet what of pronouncements that the decision to testify is never merely a matter of strategy but always a matter of personal right?114 Clearly there is tension

Strickland’s two-prong framework.”). Or when a defendant claims counsel was deficient in telling him that the court will not let him testify. See United States v. Hubbard, 638 F.3d 866, 868–70 (8th Cir. 2011) (evaluating an ineffectiveness claim where defense counsel told defendant that the district court would not permit defendant to testify in his own defense). Or when a defendant claims counsel was deficient in failing to seek a stay to allow an ill defendant to testify upon recovery. See United States v. Tavares, 100 F.3d 995, 996–97 (D.C. Cir. 1996) (“Tavares argues that his counsel’s failure to take appropriate measures in light of his health problems effectively deprived him of his right to testify.”). Or when a defendant claims counsel was deficient in advising him to testify but neglecting to prepare him adequately or misjudging its scope. See United States v. Smith, 421 F. App’x 889, 898–99 (10th Cir. 2011) (“Together, Smith contends, these shortcomings contributed to her decision to give limited testimony at trial, led to the government’s allegedly damaging cross-examination, and prejudiced her defense.”). Or when a defendant claims counsel was deficient in failing to explain that whether to testify is ultimately the defendant’s decision. See United States v. Herschberger, No. 90-3237, 1991 WL 136337, at *3 (10th Cir. July 24, 1991) (stating that the defendant asserted an ineffective assistance of counsel claim “that his counsel failed to advise him he had the right to decide whether he would testify”). But our view is that, for purposes of doctrinal clarity and consistency, such allegations of error ought to be considered together with actual coercion cases and with purported violations by a court, a prosecutor, or a statute, under an independent, freestanding right-to-testify analysis.

112. In the words of Judge J. Skelly Wright, sitting as a district judge in Poe v. United States:

The right to testify is personal to the accused. He must make the ultimate decision on whether or not to take the stand. In this regard it is unlike other decisions, which are often called “trial decisions,” where it is counsel who decides whether to cross examine a particular witness or introduce a particular document. Here it is the accused who must decide and it is the duty of counsel to present to him the relevant information on which he may make an intelligent decision.

113. Strickland, 466 U.S. at 689 (citation omitted).
114. See, e.g., United States v. Curtis, 742 F.2d 1070, 1076 (7th Cir. 1984) (“We hold that a defendant’s personal constitutional right to testify truthfully in his own behalf may not be waived by counsel as a matter of trial strategy.”).
between concessions that a defendant may choose to testify, “however irrational that insistence might be,” and denial of relief because that testimony “would [have] come at a high price.” Strickland is concerned with ensuring an “adversarial process,” but our scenario involves relations between attorney and client, not attorney and attorney. Finally, there is a poor fit between Strickland’s concern with the “justice of the finding of guilt” and the Faretta–Rock concern with the propriety of the mode by which guilt is found. The right to choose to testify is akin to the right to choose counsel, which under Gonzalez-Lopez endures regardless of how effective—or ineffective—one’s preferred lawyer may prove.

Powell v. Alabama, the famous Scottsboro Boys case from 1932, in which the lawyer did not interview his nine capital defendants until hours before the trial, is the patron saint of Assistance of Counsel Clause decisions. Justice Sutherland wrote a stirring tribute to the “guiding hand” of a defense lawyer. But the problem in coercive right-to-testify cases—the reason it is not a Strickland matter—is not too little but too much counsel, a guiding hand that has become overmastering. When this happens, as in Faretta, “in a very real sense, it is not his defense.” The right to self-representation, unlike the right to testify, is exercised only while unrepresented, but a client-thwarting lawyer can still become, like an unwanted lawyer, “an

115. Id.
117. See United States v. Cronic, 466 U.S. 648, 657 n.21 (asserting that when evaluating Sixth Amendment claims, “appropriate inquiry focuses on the adversarial process, not on the accused’s relationship with his lawyer as such”).
118. For the phrase “justice of the finding of guilt” see United States v. Agurs, 427 U.S. 97, 112 (1976), the case from which Strickland purported to take its “prejudice” test. Strickland v. Washington, 466 U.S. 668, 694 (1984); see also infra note 267.
119. See United States v. Gonzalez-Lopez, 548 U.S. 140, 148 (2006) (finding that the defendant was erroneously deprived of his right to choice of counsel when the court prevented him from “being represented by the lawyer he want[ed], regardless of the quality of the representation he received”).
121. See id. at 52–58 (describing the factual circumstances relating to the appointment of counsel for the defendants).
122. See id. at 68–69 (describing the importance of the right to counsel).
organ of the State interposed between an unwilling defendant and his right to defend himself personally.”

Thus, we have a constitutional right without a remedy. If a defendant is intimidated, pressured, or tricked by his lawyer into remaining seated—a typical threat is to withdraw midtrial—the defendant is all the same denied relief unless he can show that his testimony would have had decisive, but-for evidentiary effect, even though the right is not supposed to turn on evidence but autonomy. To that extent, he is denied his full opportunity to be heard, which the Supreme Court has never hesitated to call the “fundamental requisite of due process.”

C. The Ethical Consideration

A lawyer is required to exercise most rights on the client’s behalf. A lawyer is also required to stand down if the client decides, contrary to his advice, to testify. Model Rule of Professional Conduct Rule 1.2(a) says: “In a criminal case, the lawyer shall abide by the client’s decision, after consultation with the lawyer, as to a plea to be entered, whether to waive jury trial and whether the client will testify.” To be sure, the Supreme Court reminds us that the “Constitution does not codify the

124. Id. at 820.


126. MODEL RULES OF PROF’L CONDUCT R. 1.2(a) (2012).

127. Id. (emphasis added). See ABA Comm. on Ethics & Prof’l Responsibility, Formal Op. 87-353 (1987) (“If the lawyer does not offer the client’s testimony and, on inquiry by the court into whether the client has been fully advised [of his] right to testify, the client states a desire to testify . . . the lawyer may have no other choice than to disclose . . . the client’s intention to testify falsely.”); see also ABA STANDARDS FOR CRIMINAL JUSTICE § 4-5.2(a) (3d ed. 1993) (listing “whether to testify in his or her own behalf” among the “decisions relating to the conduct of the case . . . ultimately for the accused”).
ABA’s Model Rules,” but, with equal sureness, it tells us that the judiciary has an “independent interest in ensuring that criminal trials are conducted within the ethical standards of the profession.” Mild reproof is often the sole consequence of a lawyer’s right-to-testify violation. Counsel must advise on exercise of the right—and the line between advice and coercion is a question of fact on which many right-to-testify cases turn.

III. Is a Violation of the Right to Testify a “Structural” Error?

Arizona v. Fulminante reaffirmed the rule that “most constitutional errors can be harmless.” The varieties of harmless error are legion, and they even include errors that

130. See, e.g., Jiles v. United States, 72 F. App’x 493, 493 (7th Cir. 2003) (“If counsel actually said this, it is inexcusable behavior that likewise has the potential to establish prejudice for purposes of Strickland.”).
131. See, e.g., Lema v. United States, 987 F.2d 48, 52–53 (1st Cir. 1993) (describing the principles guiding courts in drawing the “difficult line” between “earnest counseling and over coercion”).
132. In Part III we discuss four categories largely of our devising. There is probably a fifth category for judgments rendered by tribunals without jurisdiction, which is irrelevant here. See Nguyen v. United States, 539 U.S. 69, 71 (2003) (determining that a Ninth Circuit panel comprised of two Article III judges and one Article IV judge from the Northern Mariana Islands did not have the authority to decide the appeal); Wingo v. Wedding, 418 U.S. 461, 472 (1974) (concluding that the habeas corpus statute requires district courts, not magistrate judges alone, to conduct evidentiary habeas hearings); United States v. Amer.-Foreign S.S. Corp., 363 U.S. 685, 685–86 (1960) (concluding that retired circuit judges were ineligible to participate in the decision of a case on rehearing en banc).
134. Id. at 306.
cripple the presentation of a defense. For example, in *Crane v. Kentucky*, a defendant sought to introduce testimony attacking the credibility of his confession, a move the judge erroneously blocked as an attempt to relitigate a suppression motion. In *Delaware v. Van Arsdall*, a judge improperly prevented the accused from impeaching a government witness over his alleged dealmaking with the prosecution. These errors, the Court wrote in *Fulminante*, were subject to harmless-error analysis because they involved “trial error” that “occurred during the presentation of the case to the jury” and so could be “quantitatively assessed in the context of other evidence presented in order to determine whether its admission [was] harmless beyond a reasonable doubt.” Under that definition of trial error, the Court in *Fulminante* used harmless-error review in evaluating the admission of a coerced confession. By contrast, said the Court, “structural” errors occur when the “entire conduct of the trial from beginning to end is obviously affected” by the errors and so “defy analysis by ‘harmless-error’ standards.”

The inquiry into whether an error is structural really turns on the nature of the right and the effect of the error. The main complication is that the right to testify has two natures: first, as recognition of the defendant’s dignity, and second, as a means for him to strive for an acquittal. With respect to dignity, suppression is the error and only reversal can remedy it. But with respect to acquittal, the denial of the right is a reversible error only if the denial diminished his chance of acquittal; if it would not have helped in the end, there is nothing for a reviewing court

---

137. See id. at 690–91 (“Under these principles, the Kentucky courts erred in foreclosing petitioner’s efforts to introduce testimony about the environment in which the police secured his confession.”).
139. See id. at 675–77 (describing the court’s failure to permit defense counsel to impeach a government witness during cross-examination).
141. See id. at 307–09 (analyzing precedent and determining that harmless-error review was appropriate).
142. Id. at 309–10.
to do. No circuit court has held that a right-to-testify error is structural, yet some seem to consider the question open.

A. Structural Because Unassessable?

Errors are structural for different reasons. Probably the most common reason is the difficulty of assessing the error’s effect. In United States v. Gonzalez-Lopez a defendant was denied counsel of choice. The Court found the harmless-error and Strickland tests inapposite because they concern “identifiable mistakes” that judges can “assess” as they bear on the “outcome.” But to assess a wrongful denial of choice of counsel, judges would need to speculate on the probable acts of the

143. See, e.g., United States v. Tavares, 100 F.3d 995, 997–98 (D.C. Cir. 1996) (describing the standard for reviewing defense counsel’s failure to permit defendant to testify and citing cases that support the court’s choice of standard).

144. See, e.g., United States v. Flores-Martinez, 677 F.3d 699, 712 n.8 (5th Cir. 2012) (arguing that “the right to testify is a constitutional right so basic to a fair trial that its infraction can never be treated as harmless error” yet stating that the Court “need not decide this question” (emphasis added)); see also Barrow v. Uchtman, 398 F.3d 597, 608 n.12 (7th Cir. 2005) (per curiam)

It is an interesting question whether defendant’s forfeiture of his constitutional right to testify, standing alone, is sufficiently “prejudicial” to warrant reversal of a conviction, or whether the decision not to testify—even when based on erroneous legal advice—is not prejudicial unless it actually affects the outcome of the trial. Seventh Circuit precedent seems to support the latter view that defendants who allege they waived their right to testify still must show that this waiver was prejudicial, i.e., that the failure to testify affected the outcome of the trial.

145. We cite some better-known cases below, but there are others. In Stirone v. United States, a judge in a Hobbs Act case permitted a defendant to be tried on a charge not made in the indictment. Stirone v. United States, 361 U.S. 212, 214 (1960). The Court stated that this error could never be harmless because “we cannot know whether the grand jury would have included [the added charge] in its indictment,” and yet “this might have been the basis upon which the trial jury convicted petitioner.” Id. at 219. In Ballard v. United States, violation of a statutory scheme resulted in an all-male jury panel. Ballard v. United States, 329 U.S. 187, 193 (1946). The Court considered the “subtle interplay of influence[s]” that women have on men, and men on women, to be “among the imponderables,” and reversed without any examination of prejudice. Id. at 193, 195–96.


147. Id. at 150–51.
rejected counsel—from his relations with prosecutors to cross-examination questions to courtroom style.\footnote{148. Id.} In Holloway v. Arkansas,\footnote{149. Holloway v. Arkansas, 435 U.S. 475 (1978).} the Court found that unconstitutional multiple representation was not subject to harmless-error analysis because “to assess the impact of a conflict of interests on the attorney’s options, tactics, and decisions in plea negotiations would be virtually impossible.”\footnote{150. Id. at 491.} In Waller v. Georgia,\footnote{151. Waller v. Georgia, 467 U.S. 39 (1984).} the Court held that the violation of the public-trial guarantee was not reviewable for harmless error\footnote{152. Id. at 49.} because the “benefits of a public trial are frequently intangible, difficult to prove, or a matter of chance.”\footnote{153. Id. at 49 n.9.} In Vasquez v. Hillery,\footnote{154. Vasquez v. Hillery, 474 U.S. 254 (1986).} the Court held that when black citizens are “systemically excluded” from grand-jury service, the error is not “amenable” to harmless-error review because the “effect of the violation cannot be ascertained.”\footnote{155. Id. at 263–64; see also Batson v. Kentucky, 476 U.S. 79, 86–87 (1986) (reviewing the effects of racial discrimination in jury selection not only on the accused but also on society, the public’s impression of the court’s competence, and the excluded juror).} In Price v. Georgia,\footnote{156. Price v. Georgia, 398 U.S. 323 (1970).} a state court reversed a manslaughter conviction because of an erroneous jury instruction,\footnote{157. Id. at 324.} but the Court prohibited retrial on the original murder charge because it could not “determine whether or not the murder charge against petitioner induced the jury to find him guilty of the less serious offense of voluntary manslaughter rather than to continue to debate his innocence.”\footnote{158. Id. at 331.} In Gray v. Mississippi,\footnote{159. Gray v. Mississippi, 481 U.S. 648 (1987).} the Court rejected the argument that an improper for-cause exclusion of a prospective juror reluctant to impose the death penalty could be harmless, even when a prosecutor would otherwise have exercised a peremptory challenge, for “we cannot know whether
in fact he would have had this peremptory challenge left to use. And in Sullivan v. Louisiana, the Court found that an erroneous instruction on reasonable doubt could never be harmless because the consequences were “necessarily unquantifiable and indeterminable.”

Does the unassessability rationale apply in the right-to-testify context? Courts do identify the “special significance” that a defendant’s testimony can have, considering the “power of a face-to-face appeal” and “inherent significance” of his word. The dissenters in Foster v. Delo, for instance, felt that the

160. Id. at 664–65. See Witherspoon v. Illinois, 391 U.S. 510, 522 (holding that a death sentence “cannot be carried out if the jury that imposed or recommended it was chosen by excluding veniremen for cause simply because they voiced general objections to the death penalty or expressed conscientious or religious scruples against its infliction”).


162. Id. at 281–82.

163. Other mainstays of due process jurisprudence certainly partake of this unassessability rationale but are better discussed under different subheadings below. See, e.g., Gideon v. Wainwright, 372 U.S. 335 (1963) (right to counsel in criminal cases); Tumey v. Ohio, 273 U.S. 510 (1927) (biased judge).

164. See, e.g., Cannon v. Mullin, 383 F.3d 1152, 1172 (10th Cir. 2004) (“We are also cognizant of the power of a face-to-face appeal . . . . Mr. Cannon’s testimony, and his demeanor while testifying, could have special significance to the jury on this matter.”). In addition, a defendant’s testimony allows the jury to assess the defendant’s physiognomy and demeanor. See Riggins v. Nevada, 504 U.S. 127, 137–38 (1992) (reversing a conviction after the accused had been forcibly medicated with antipsychotics during trial—in part because medication could have “effects” on his “outward appearance” that “cannot be shown from a trial transcript”). In Sell v. United States, the Court noted that involuntary medications can “diminish the ability to express emotions” at trial, which can “undermine” its fairness. Sell v. United States, 539 U.S. 519, 185–86 (2003). There are also cases teaching that elements of appearance in the form of clothing—like, presumably, elements of appearance in facial expressions—can have consequences not capturable by a transcript. See, e.g., Estelle v. Williams, 425 U.S. 501, 504–05 (1976) (forced wearing of prison clothing); Illinois v. Allen, 397 U.S. 337, 344 (1970) (binding and gagging accused during trial).

165. See, e.g., Nichols v. Butler, 953 F.2d 1550, 1553 (11th Cir. 1992) (“The testimony of a criminal defendant at his own trial is unique and inherently significant.”); United States v. Walker, 772 F.2d 1172, 1178–79 (5th Cir. 1985) (stating that the accused, with his knowledge of facts, his testimony, and his demeanor, is of “prime importance” at trial); United States v. Irvin, 450 F.2d 968, 971 (9th Cir. 1971) (Kilkenny, J., dissenting) (“The facial expressions of a witness may convey much more to the trier of the facts than do the spoken words.”).

166. Foster v. Delo, 39 F.3d 873 (8th Cir. 1994).
accused’s testimony could have been decisive against a prosecutor who “dehumanized” the defendant by labeling him a “that”—as in “that (indicating) is no man.”167 There is a sense in which the force of ungiven testimony is immeasurable.168 It goes beyond the substance of his testimony, but rather is a matter of a willingness to speak directly to those judging you, to refuse to hide behind the Fifth Amendment, to put on display any emotion that testifying arouses, whether a tremble in the voice of one falsely accused or the manufactured confidence of one seeking to lie his way out of guilt.

The elusive effect of all this is hardly captured by words on a page. This sort of evidence is of a quality different from other types that a defendant might present. Testimony from the individual at the heart of events is uncommonly probative. Confessions—another form of defendant testimony—may be the “most probative and damaging evidence that can be admitted against him.”169 They “come from the actor himself, the most knowledgeable and unimpeachable source of information about his past conduct” and have such a “profound impact on the jury”

167. Id. at 885–86 (Bright, J., dissenting).
168. See Wright v. Estelle, 572 F.2d 1071, 1081 (5th Cir. 1978) (Godbold, J., dissenting)

[T]he right to testify resembles other rights recognized as requiring automatic reversal because it is impossible, and perhaps improper, to attempt to judge the effect that the defendant’s appearance on the stand would have had on the jury . . . . Judges can, with a reasonable degree of assurance, identify and sort out merely trivial or cumulative evidence and form a reasoned judgment on possible impact upon the jury of what it erroneously heard or failed to hear. There is a degree of speculation, but the risk is acceptable. Where the error is in keeping the defendant from the stand the judge can consider the content of what the defendant might have said the same as for a nonparty witness. But he cannot weigh the possible impact upon the jury of factors such as the defendant’s willingness to mount the stand rather than avail himself of the shelter of the Fifth Amendment, his candor and courtesy (or lack of them), his persuasiveness, his respect for court processes. These are elusive and subjective factors, even among persons who might perceive and hear the defendant, but more significantly, they are matters neither communicated to an appellate judge nor susceptible of communication to him. Appellate attempts to appraise impact upon the jury of such unknown and unknowable matters is purely speculative.

that we “may justifiably doubt its ability to put them out of mind even if told to do so.”\textsuperscript{170} If a defendant’s testimony can doom him in compelled-confession cases, why might not it save him in right-to-testify cases?

On the other hand, the unadmitted testimony can almost always be described later in hearings and affidavits (e.g., “I would have said this, to establish that”) and at that point it can be weighed alongside the rest of the record.\textsuperscript{171} If \textit{Fulminante} holds that the erroneous admission of a defendant’s testimonial statement can be harmless,\textsuperscript{172} why should the erroneous exclusion of testimony require reversal? Compelled testimony, too, is a matter of the gravest dignity, volition, and fairness. \textit{Fulminante} suggests that, if the criterion is assessability of the error’s effect, the denial of the right to testify is a trial error.\textsuperscript{173} The accused in

\begin{itemize}
  \item \textsuperscript{170} \textit{Id.}
  \item \textsuperscript{171} But this is not permitted in all circumstances. In \textit{Luce v. United States}, the Court held that to preserve a claim of improper impeachment with a prior conviction under Fed. R. Evid. 609(a), a defendant must testify—otherwise a court cannot assess harmlessness. \textit{Luce v. United States}, 469 U.S. 38, 43 n.5 (1984). Thus, a defendant could not use an offer of proof on appeal as a way to allow an evaluation of harmlessness.
  \item \textsuperscript{172} See \textit{Arizona v. Fulminante}, 499 U.S. 279, 310 (1991) (“When reviewing the erroneous admission of an involuntary confession, the appellate court, as it does with the admission of other forms of improperly admitted evidence, simply reviews the remainder of the evidence against the defendant to determine whether the admission of the confession was harmless beyond a reasonable doubt.”).
  \item \textsuperscript{173} One could argue that \textit{Gonzalez-Lopez} supports the proposition that right-to-testify errors are structural. Under its rule, the dissent observed, a “defendant who is erroneously required to go to trial with a second-choice attorney is automatically entitled to a new trial even if this attorney performed brilliantly.” \textit{United States v. Gonzalez-Lopez}, 548 U.S. 140, 160 (2006) (Alito, J., dissenting). The majority required the stringent remedy because of the unassessability of what might have been, see \textit{id.} at 150 (majority opinion) (“We have little trouble concluding that erroneous deprivation of the right to counsel of choice . . . unquestionably qualifies as ‘structural error. . . .’ It is impossible to know what different choices the rejected counsel would have made, and then to quantify the impact of those different choices on the outcome of the proceedings.”), but why does it even matter that a defendant gets his counsel of choice? It seems linked to autonomy. Certainly Justice Scalia made clear it was not merely about a fair trial. \textit{See id.} at 146 (“In sum, the right at stake here is the right to counsel of choice, not the right to a fair trial; and that right was violated because the deprivation of counsel was erroneous.”). Nonetheless, because the Court did not invoke that rationale, and in fact labeled it a different “criterion” for finding structural error, \textit{id.} at 149 n.4, our reliance is hesitant.
\end{itemize}
WHY STRICKLAND IS THE WRONG TEST

the right-to-testify context, unlike a Faretta defendant,\textsuperscript{174} does not get an entirely different proceeding as a result of the error, but rather a trial that is simply missing a piece of evidence—an important piece, but one that can be ascertained post-trial. Gonzalez-Lopez explained that we cannot know what the lawyer-who-might-have-been would have done at trial;\textsuperscript{175} similarly, we cannot know how a Faretta defendant might have acquitted himself. But where the defendant is prohibited from testifying, no alternate universe of possibility has been cut off. And this, under this particular line of structural-error precedent, seems to make all the difference.

\textbf{B. Structural Because Risk of Prejudice Too Great?}

A second rationale for structural error, especially in prejudicial-publicity cases,\textsuperscript{176} appears when there is “such a probability that prejudice will result that it is deemed inherently lacking in due process.”\textsuperscript{177} The concern is that a pervasive, insinuating press is likely (even if undetectably) to erode juror objectivity.\textsuperscript{178} Or, in \textit{Tumey v. Ohio},\textsuperscript{179} the Court invalidated a procedure whereby a judge, the village mayor, received fees and

\textsuperscript{174} See Faretta v. California, 422 U.S. 806, 835–36 (1975) (vacating the judgment and remanding Faretta’s case for new proceedings after Faretta was forced to accept appointed counsel despite requests to represent himself).

\textsuperscript{175} See United States v. Gonzalez-Lopez, 548 U.S. 140, 150 (2006) (describing the Court’s inability to determine how a different lawyer would have handled a case and the impact those decisions might have had on the outcome of the proceedings).

\textsuperscript{176} See generally Estes v. Texas, 381 U.S. 532 (1965) (concluding that the defendant was deprived of due process by the televising of his criminal trial); Irwin v. Dowd, 366 U.S. 717 (1961) (finding that defendant was denied right to an impartial jury due to extensive media coverage of the case in months prior to the trial); see also Sheppard v. Maxwell, 384 U.S. 333, 353–58 (1966) (describing the courtroom as a “carnival atmosphere” and the press’ impact on jurors who became “celebrities,” subject to months of skewed press coverage urging death penalty).

\textsuperscript{177} \textit{Estes}, 381 U.S. at 542–43.

\textsuperscript{178} See \textit{id.} at 545 (noting that it is “highly probable” that invasive press coverage “will have a direct bearing on [a juror’s] vote as to guilt or innocence” due to the juror feeling “the pressures of knowing that friends and neighbors have their eyes upon them”).

\textsuperscript{179} Tumey v. Ohio, 273 U.S. 510 (1927).
costs upon convictions but not acquittals;\textsuperscript{180} this would “offer a possible temptation to the average man as a judge to forget the burden of proof.”\textsuperscript{181} But the impartiality concerns that justify structural-error analysis in these cases do not relate to the autonomy concerns underlying the right to testify.

\textit{C. Structural Because Harm at Trial Irrelevant?}

A third rationale is the “irrelevance of harmlessness.”\textsuperscript{182} This is the reason why \textit{Faretta} errors are structural. In \textit{McKaskle v. Wiggins},\textsuperscript{183} the Supreme Court said: “Since the right of self-representation is a right that when exercised usually increases the likelihood of a trial outcome unfavorable to the defendant, its denial is not amenable to ‘harmless error’ analysis. The right is either respected or denied; its deprivation cannot be harmless.”\textsuperscript{184} In \textit{Johnstone v. Kelly},\textsuperscript{185} defendant Johnstone wanted to represent himself, yet the trial court appointed counsel.\textsuperscript{186} This ruling was reversed.\textsuperscript{187} On retrial, the judge told Johnstone he had to go pro se—as he had demanded.\textsuperscript{188} But the circuit court rejected the argument that Johnstone could not make a fresh election this time around: the trial court’s error was that it “denied him the \textit{choice} whether to have counsel or proceed pro se. It is that \textit{choice} that must be accorded at a retrial.”\textsuperscript{189}

If \textit{Rock v. Arkansas} declared the right to testify “[e]ven more fundamental to a personal defense” than the right of self-representation,\textsuperscript{190} would not the right to testify, logically, be in

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{180} \textit{Id.} at 520.
\item \textsuperscript{181} \textit{Id.} at 532.
\item \textsuperscript{182} United States v. Gonzalez-Lopez, 548 U.S. 140, 148–49, 149 n.4. (2006).
\item \textsuperscript{184} \textit{Id.} at 177 n.8.
\item \textsuperscript{185} Johnstone v. Kelly, 808 F.2d. 214 (2d Cir. 1986).
\item \textsuperscript{186} \textit{Id.} at 215–16.
\item \textsuperscript{187} \textit{Id.} at 218.
\item \textsuperscript{188} \textit{See Johnstone v. Kelly, 812 F.2d} 821, 821 (2d Cir. 1987) (discussing the State’s argument that it could “satisfy its obligation to [the defendant] by affording him a retrial at which he would be \textit{required} to represent himself” (emphasis added)).
\item \textsuperscript{189} \textit{Id.} at 822 (emphasis added).
\item \textsuperscript{190} Rock v. Arkansas, 483 U.S. 44, 52 (1987).
\end{itemize}
\end{footnotesize}
WHY STRICKLAND IS THE WRONG TEST

the structural class? Judge Godbold made the argument in his dissent in Wright v. Estelle: “[w]hen personal rights are involved, the harmless error rule does not apply because we are not concerned with the ‘ultimate consequences’ of trial, but with preventing the individual from being overcome by the criminal process.”¹⁹¹ Is this right?¹⁹² If a defendant was silenced, but would have testified only on some entirely peripheral point, must this require retrial? In one sense, the error would not “affect” his “substantial rights,” which obligates courts to ignore it.¹⁹³ On the other hand, if the right is simply to be able to choose to testify truthfully, no matter how significant the testimony might be as a piece of evidence, a violation eliminates the right. The answer hinges on what you believe to be the particular harm being reviewed for harmlessness: if the wrong is the denial of choice, error is never harmless; if the wrong is impairing a defense as it affects a jury’s verdict, error here could very well be harmless.

¹⁹¹. Wright v. Estelle, 572 F.2d 1071, 1081 (5th Cir. 1978) (Godbold, J., dissenting) (explaining that the harmless-error rule is inapplicable to double-jeopardy analysis (citing Price v. Georgia, 398 U.S. 323, 331 (1970))). Judge Reinhardt made the argument in his dissent in United States v. Martinez. See generally United States v. Martinez, 883 F.2d 750 (9th Cir. 1989) (Reinhardt, J., dissenting), vacated, 928 F.2d 1470 (9th Cir. 1991). He argued: (1) because the Chapman inquiry “involves delicate judgments about fact specific situations, errors that have an indeterminate impact upon the appellate record cannot be harmless” and (2) “harmless error analysis, designed to insure correct outcomes, is essentially irrelevant to a panoply of constitutional rights that protect individual dignity.” Id. at 770 n.23. Many a petitioner has echoed this argument; all have been met with rejection. See, e.g., Skeens v. Haskins, 4 F. App’x 236, 238 (6th Cir. 2001) (“Skeens's argument that his championed error was 'structural error' is without merit. Most errors do not automatically render a trial unfair and thus, can be harmless.” (citing Arizona v. Fulminante, 499 U.S. 279, 306 (1991))).


¹⁹³. See 28 U.S.C. § 2111 (2006) (“On the hearing of any appeal or writ of certiorari in any case, the court shall give judgment after an examination of the record without regard to errors or defects which do not affect the substantial rights of the parties.”); Fed. R. Crim. P. 52(a) (defining harmless error as “[a]ny error, defect, irregularity, or variance that does not affect substantial rights must be disregarded”).
D. Structural Because Procedurally Intolerable?

Sometimes the Court refuses to consider harmlessness because the right is “fundamental and essential to a fair trial.” These include the rights to counsel in criminal cases (Gideon v. Wainwright); to trial in serious criminal cases (Duncan v. Louisiana); and to have appointed counsel prosecute certain meritorious appeals (Anders v. California). These holdings invoke some of the rationales already described, but the thrust is that in such cases the problem is not merely a reliable conviction but “fair procedure” and preventing an individual from simply being hustled through the system. It is not just an evidentiary matter or a concern about just outcomes but a problem with means. It is no rhetorical excess to say that we adhere to forms of procedure, for their own sake, almost religiously, because those procedures work to minimize discretion and safeguard justice and because that is just how we do things in this country.

The rationale of procedural impermissibility may apply here in that a silenced defendant has had a right withdrawn from him—regardless of whether he faced overwhelming evidence or wished to say things that would have angered jurors and provoked a more severe sentence. Like a man who is forced to plead guilty, he is wronged in a way independent of concerns about accuracy of result. He is wronged like a convict who is denied a right to allocute, though he could hardly have swayed a judge. Or like a man who sought jury trial but got a bench trial, though clearly he would have been convicted under either fact-}

198. See id. at 741 (“We have concluded that California’s action does not comport with fair procedure and lacks that equality that is required by the Fourteenth Amendment.”).
199. See Blackburn v. Alabama, 361 U.S. 199, 206–07 (1960) (explaining that the accused’s guilt or innocence is not the only consideration in a proceeding and noting that “important human values are sacrificed” when the government is permitted to use unfair procedures to secure a conviction); see also Spano v. New York, 360 U.S. 315, 320–21 (1959) (noting that “in the end life and liberty can be as much endangered from illegal methods used to convict those thought to be criminals as from the actual criminals themselves”).
WHY STRICKLAND IS THE WRONG TEST

finder. These are simply choices a man facing the State has. On this theory, the question that most federal courts ask when a defendant claims his right to testify was abridged—Did it matter in the end?—is the wrong one. The proper question is: Was this an invalid proceeding? Is gagging a man at this dramatic crossroad in his life very different from trying him in absentia?

It comes down to how one reads this sentence from Fulminante:

[T]he harmless error doctrine is essential to preserve the “principle that the central purpose of a criminal trial is to decide the factual question of the defendant’s guilt or innocence, and promotes public respect for the criminal process by focusing on the underlying fairness of the trial rather than on the virtually inevitable presence of immaterial error.”

So are we talking here about the factual question of guilt? Or underlying fairness? Or both? Or autonomy? We know trial fairness is not the sum of things. In Lafler v. Cooper, the Supreme Court said that when pleas are mishandled by defense counsel, the issue was not the “fairness or reliability of the trial but the fairness and regularity of the processes that preceded it.” Even if the trial reaches an accurate result, there remains a sense that a wrong was done which requires its own remedy. This is why, under Hamilton v. Alabama, we “do not stop to determine whether prejudice resulted” to a defendant left to face arraignment without a lawyer—even if it can be shown that the lack of counsel had no effect whatsoever on the trial. In Gonzalez-Lopez, the Court similarly declared that the “right to

202. Id. at 1388.
204. Id. at 55.
205. See id. ("When one pleads to a capital charge without benefit of counsel, we do not stop to determine whether prejudice resulted. . . . [T]he degree of prejudice can never be known. Only the presence of counsel could have enabled this accused to know all the defenses available to him and to plead intelligently.").
select counsel of one’s choice” has “never been derived from the Sixth Amendment’s purpose of ensuring a fair trial.”

This is a difficult question. On the one hand, there is doubtless precedent to support the notion that the right to testify truthfully, bound up with the notion of not treating a man as a voiceless object to be disposed of by lawyers, demands reversal to remedy error. It is a question of process, not evidentiary weight. At the same time, other precedent suggests that this right, despite its uncontested importance, is still at bottom a trial

206. United States v. Gonzalez-Lopez, 548 U.S. 140, 147 (2006). There is something of a tension, however, with the choice rationale and Wheat v. United States, 486 U.S. 153 (1988). In Wheat, Chief Justice Rehnquist (who dissented in Faretta) wrote that the right to counsel exists to ensure that a man does not get railroaded by the State. See Wheat, 486 U.S. at 158–59 (asserting that the right to counsel exists to ensure that criminal defendants receive a fair trial and noting that “an unaided layman may have little skill in arguing the law or in coping with an intricate procedural system”). Once that guarantee is enforced, the defendant has some latitude in choice, but a fair trial is always the main thing. Id. A bigger problem is that Wheat rejects the argument that a man should have his lawyer of choice and that if his lawyer has a conflict of interest, it is simply the client’s choice to waive. See id. at 164 (“The District Court must recognize a presumption in favor of petitioner’s counsel of choice, but that presumption may be overcome not only by a demonstration of actual conflict but by a showing of a serious potential for conflict.”). Chief Justice Rehnquist said that courts have an independent interest in ethical standards and in trials with the appearance of integrity. See id. at 160 (“Federal courts have an independent interest in ensuring that criminal trials are conducted within the ethical standards of the profession and that legal proceedings appear fair to all who observe them.”). Perhaps this represents a shift away from autonomy, but in any event our situation is still nearer to Faretta than Wheat. Faretta, like the right to testify, concerns what a man whose liberty is at stake may do for himself. Wheat involves what one is allowed to have another man do on one’s behalf. Perhaps Gonzalez-Lopez shifted us back toward choice. Justice Scalia, speaking of Wheat, wrote: “It is one thing to conclude that the right to counsel of choice may be limited by the need for fair trial, but quite another to say that the right does not exist unless its denial renders the trial unfair.” Gonzalez-Lopez, 548 U.S. at 147 n.3. Besides, if courts have an interest in ethical standards, one worthy way to uphold these standards would be to properly remedy the unethical misconduct of defense lawyers who prevent their clients from testifying.

207. See, e.g., Wright v. Estelle, 572 F.2d 1071, 1081 (5th Cir. 1978) (en banc) (Godbold, J., dissenting) (stating that “the right to testify resembles other rights recognized as requiring automatic reversal”).

208. See id. at 1075 (stating the belief “that the federal constitution now requires state and federal courts to allow a defendant to testify” and that “[m]ost often the right is treated as part of due process”).
right. It allows an accused to proffer evidence to attempt to exonerate himself. A chance at trial victory is why the common law disability was lifted. It is why defendants speak. It is why they complain about denials of the right. And this right can only be exercised at trial. If testimony cannot under any circumstances have secured a defendant an acquittal, a new trial would be futile.

But we reason as follows: the right to testify derives in part from the Fifth Amendment. It mirrors the right against compelled self-incrimination; so said the Court in *Rock* when it called the right to testify the “necessary corollary” of the right to remain silent. And there seems no logical reason to protect the right against forced speech more than the right against forced silence: compulsion to testify to your detriment is just as bad as compulsion against testifying to your benefit. Yet *Fulminante* commands that a Fifth Amendment violation, even one involving

209. See, e.g., United States v. Hung Thien Ly, 646 F.3d 1307, 1313 (11th Cir. 2011) (characterizing the right to testify as a “fundamental trial right”).


211. Of course, the exercise of one right waives the other. If anything, it seems safer to presume that the right to silence is the default preference. See *Harvey v. Shillinger*, 76 F.3d 1528, 1535 (10th Cir. 1996) (“A defendant’s choice to exercise his right to allocution, like the choice to exercise the right to testify, is entirely his own . . . . Once a defendant chooses to testify . . . he waives his privilege against compelled self-incrimination.”); United States v. Pennycooke, 65 F.3d 9, 11 (3d. Cir. 1995) (noting that “[e]xercise of either the right to testify or the right not to testify necessarily would waive the other right” and cautioning that a court’s advice regarding the right to testify could inappropriately influence a defendant to waive his or her right to remain silent).

Another interesting problem is how a pro se defendant testifies. See generally *United States v. Nivica*, 887 F.2d 1110 (1st Cir. 1989). In *Nivica*, the examination began and ended thus:

MR. WELLINGTON: The question is: Does Mark Pedley Wellington, a/k/a Jack Williams, have anything to hide?

The answer is No.

[PROSECUTOR]: Objection.

THE COURT: Sustained. Please strike the answer. Please wait until an objection is made, if any is made, before you answer.

MR. WELLINGTON: Well, I guess I can’t ask myself any more questions then.

THE COURT: Thank you. You are excused.

*Id.* at 1120.
actual coercion, does not mean automatic reversal.\textsuperscript{212} A violation of the right to testify, even one involving suppression, likewise should not bring on instant reversal.\textsuperscript{213} Both errors may involve egregious trampolings on fundamental rights, but both errors occur within a larger record that allows a court to reconstruct the probable result had no mistakes been made. The right to testify emanates from the Compulsory Process Clause\textsuperscript{214}: a defendant may almost always call witnesses, including himself.\textsuperscript{215} Washington v. Texas\textsuperscript{216} declared that the “right to present the defendant’s version of the facts” is a “fundamental element of due process.”\textsuperscript{217} But like with the Fifth Amendment, errors under this Clause have not been declared, without more, to be structural.\textsuperscript{218}

\textsuperscript{212.} See Arizona v. Fulminante, 499 U.S. 279, 295–302 (applying harmless-error analysis to a coerced confession).

\textsuperscript{213.} Really the only lingering question is whether Fulminante would control if the right against compelled self-incrimination was flouted flagrantly enough in court to require a finding of mistrial. The Court in Brecht v. Abrahamson observed that certain “deliberate and especially egregious” errors could destroy the integrity of proceedings even if the error was normally of the “trial type.” Brecht v. Abrahamson, 507 U.S. 619, 638 n.9 (1993). If so, might the mirror-image analogue be a total suppression of the right to testify? It is interesting to recall that Chapman v. California said that “there are some constitutional rights so basic to a fair trial that their infraction can never be treated as harmless error.” Chapman v. California, 386 U.S. 18, 23 (1967). Among these rights, the Court cites the defendant’s right, under Payne v. Arkansas, to not have coerced confessions produced against him. See id. at 25–26 (stating that a prosecutor commenting on a defendant’s failure to testify “can no more be considered harmless than the introduction against a defendant of a coerced confession”); Payne v. Arkansas, 356 U.S. 560, 568 (1958) (reversing a judgment because the court admitted a coerced confession into evidence). But Payne ceased being the law on the day Fulminante was announced. See Arizona v. Fulminante, 499 U.S. at 288 (stating that the decision “abandons” the Court’s previous rulings in cases such as Payne).

\textsuperscript{214.} U.S. Const. amend. VI (“In all criminal prosecutions, the accused shall enjoy the right . . . to have compulsory process for obtaining witnesses in his favor . . . .”).

\textsuperscript{215.} Id.

\textsuperscript{216.} Washington v. Texas, 388 U.S. 14 (1967).

\textsuperscript{217.} Id. at 18.

\textsuperscript{218.} See Pennsylvania v. Ritchie, 480 U.S. 39, 58 (1987)

Ritchie is entitled to have the CYS file reviewed by the trial court to determine whether it contains information that probably would have changed the outcome of his trial. If it does, he must be given a new trial. If the records maintained by CYS contain no such information, or if the nondisclosure was harmless beyond a reasonable doubt, the lower court will be free to reinstate the prior conviction.
WHY STRICKLAND IS THE WRONG TEST

Our conclusion is the only practical one. If right-to-testify errors are declared to be structural, courts will never find them, especially when the error can be asserted in most cases in which the defendant did not testify. It makes sense not to call it structural error and instead to seek other ways to invigorate the right.

IV. The Proper Test: Harmless Error Under Chapman/Kotteakos

If structural error and Strickland are not proper standards for assessing abridgements of the right to testify, that leaves us with the doctrine of harmless error. The standard, on direct review, is set out in Chapman v. California: “before a federal constitutional error can be held harmless, the court must be able to declare a belief that it was harmless beyond a reasonable doubt.” In collateral attacks, courts apply the supposedly more

Compulsory Process errors are not structural, but they are also almost never harmless. See Holmes v. South Carolina, 547 U.S. 319, 331 (2006) (holding that the South Carolina rule limiting the defendant’s evidence of third-party guilt to facts that are inconsistent with his own guilt violates the Compulsory Process Clause); Chambers v. Mississippi, 410 U.S. 284, 298 (1973) (holding that a state “voucher” rule violates due process).

The right to testify, as noted above, also finds a source in the Due Process Clause, but that Clause comprehends too many rights in too many circumstances to make any informative generalization about structural error.

219. We realize that, generally, under the Federal Rules of Criminal Procedure, a failure to object to a supposed constitutional error entitles the complainant only to plain error review, where he has the burden. See FED. R. CRIM. P. 52(b) (“A plain error that affects substantial rights may be considered even though it was not brought to the court’s attention.”). But the nature of the denial of the right to testify is such that, in the instances most concerning to us, the defending attorney would have to object to himself. It would also be unfair to require the defendant to object (where? how?) to his counsel’s effort to silence him. In any event plain-error review is more appropriate than Strickland review (same burdens, but no presumptions; and it does not turn on the wrong clause). In order to make our point we will leave this particular question—plain error or harmless error—aside.


221. Id. at 24. Chapman also involved defendant testimony. Id. In that case, a prosecutor commented on petitioner’s failure to testify. Id. at 19. This was the very concern that prompted the common-law ban on an accused’s testimony, which Rock v. Arkansas finally did away with. See Rock v. Arkansas, 483 U.S. 44, 62 (1987) (rejecting Arkansas’s limitation on the defendant’s right to testify on his own behalf). Chapman speaks in terms of affecting the “result of the
error-forgiving standard of *Kotteakos v. United States*; error requires reversal only if it had a “substantial and injurious effect or influence in determining the jury’s verdict.” The principle behind all harmless-error tests—that courts will not bother with futile exercises—is expressed in 28 U.S.C. § 2111, which directs courts to review for legal errors “without regard” to those that do not affect the parties’ “substantial rights.” Consider how dramatically different this is from *Strickland*’s prejudice test, where the petitioner must establish “a reasonable probability that, but for counsel’s unprofessional errors, the result of the proceeding would have been different.”

A trial,” *Chapman*, 386 U.S. at 22, but the standard today applies to pretrial proceedings, post-trial proceedings, plea negotiations, etc.


223. *Id.* at 776. The application of this standard to collateral attack was the work of *Brecht v. Abrahamson*. See *Brecht v. Abrahamson*, 507 U.S. 619, 637–38 (1993) (“The *Kotteakos* standard is thus better tailored to the nature and purpose of collateral review.”); see also *Fry v. Pliler*, 551 U.S. 112, 116 (2007) (favoring the *Kotteakos* standard for collateral review). In *Brecht*, Justice Stevens, whose concurrence provided the fifth vote, wrote that although the *Kotteakos* standard was “less stringent” than *Chapman*, “[g]iven the critical importance of the faculty of judgment in administering either standard . . . that difference is less significant than it might seem.” *Brecht*, 507 U.S. at 643 (Stevens, J., concurring).


225. *Id.* See *Fed. R. Crim. P.* 52(a) (“Any error, defect, irregularity, or variance that does not affect substantial rights must be disregarded.”).

WHY STRICKLAND IS THE WRONG TEST

A. Harmless Error Versus Strickland Prejudice?

But aren’t harmless error and Strickland prejudice essentially the same inquiry? Aren’t both concerned with the question: Did it matter in the end? Only superficially. In Strickland the burden is on the defendant;227 in Chapman and Kotteakos it is on the government.228 Strickland imposes a “strong presumption of reliability” about the result;229 the harmless-error statute expresses a “congressional preference for determining ‘harmless error’ without the use of presumptions.”230 Strickland tells us to presume in favor of attorney competence;231 harmless error, again, is freighted with no such presumptions.232 Strickland looks for a “reasonable probability” that the result was affected by the error;233 harmless error asks whether error was harmless “beyond a reasonable doubt”234 or caused a “substantial and injurious effect.”235 Strickland’s thrust is to require a court to be quite sure, before reversing, that the verdict was attributable

227. See id. at 696 (stating that the defendant has the burden of proving prejudice in a claim of ineffective counsel).
228. See Chapman v. California, 386 U.S. 18, 24 (1967)
   Certainly error, constitutional error, in illegally admitting highly prejudicial evidence or comments, casts on someone other than the person prejudiced by it a burden to show that it was harmless. It is for that reason that the original common-law harmless-error rule put the burden on the beneficiary of the error either to prove that there was no injury or to suffer a reversal of his erroneously obtained judgment.
See also Kotteakos v. United States, 328 U.S. 750, 771–72 (1946) (finding that the government did not meet its burden because it did not adequately justify the errors as harmless).
229. Strickland, 466 U.S. at 696.
231. See Strickland, 466 U.S. at 690 (“[T]he court should recognize that counsel is strongly presumed to have rendered adequate assistance and made all significant decisions in the exercise of reasonable professional judgment.”).
232. See Kotteakos, 328 U.S. at 765 (rejecting the use of presumptions in determining prejudicial effects of error).
to counsel’s error;\footnote{236} Chapman’s thrust is to require a court to be quite sure, before affirming, that it was “surely unattributable to the error.”\footnote{237} Kotteakos, meanwhile, tells a reviewing judge, when error is present in the record, to set aside the verdict unless he or she “is sure that the error did not influence the jury, or had but very slight effect.”\footnote{238}

Both standards, to be sure, are rarely met. But the harmless-error test at least has the virtue of giving the defendant a shot, unburdened by irrelevant presumptions. The difference is borne out in the success rates. A 2007 study considered federal circuit habeas cases between 2003 and 2006 in which a state court found error but declared it harmless;\footnote{239} the circuits, on average, disagreed in 23.5% of cases.\footnote{240} By contrast, a study that same year found that less than 1% of Strickland challenges result in habeas relief.\footnote{241} This is the difference between asking a court to say that a mistake was made (what trial is free of them?) and

\footnote{236. See Strickland, 466 U.S. at 687 (requiring the defendant to show that “counsel’s errors were so serious as to deprive the defendant of a fair trial”).}
\footnote{238. Kotteakos v. United States, 328 U.S. 750, 764 (1946).}
\footnote{240. Id. at 809. Similarly, a study coauthored by Judge Posner and William Landes considers the 963 federal appellate criminal cases (apparently direct appeals) decided between 1996 and 1998 in which the majority decided whether an error was “harmless.” William M. Landes & Richard Posner, Harmless Error 1 (Univ. Chi. Law Sch. John M. Olin Law & Econ., Working Paper No. 101 (2d Series), 2000), available at http://www.law.uchicago.edu/files/files/101.WML_.Harmless.pdf. In 19% of the cases, defendants had part of their sentence or conviction reversed, remanded, or vacated. Id. at 21.}
\footnote{241. Nancy J. King & Joseph L. Hoffman, Envisioning Post-Conviction Review for the Twenty-First Century, 78 Miss. L.J. 433, 440 (2008). Another study considering federal cases in 1990 and 1992 found that of 584 ineffective-assistance claims, the petition was “granted” in less than 1%. Victor E. Flango, Nat’l Ctr. for State Cts., Habeas Corpus in State and Federal Courts 62 tbl. 17 (1994). A third study examined more than 2,500 California state and federal appellate decisions in which an ineffective-assistance claim was raised. Laurence A. Benner, The Presumption of Guilt: Systematic Factors that Contribute to Ineffective Assistance of Counsel in California, 45 Cal. W. L. Rev. 263, 323 (2009). It found only 104 decisions in which both deficiency and prejudice were found, a success rate of 4%. Id. at 323–24.}
WHY STRICKLAND IS THE WRONG TEST

asking a court to declare an officer of the bar incompetent (rarely true, one hopes). Professor Nancy J. King, Assistant Reporter to the Advisory Committee for the Federal Rules of Criminal Procedure and a coauthor of the latter study, writes that success under Strickland remains “essentially hypothetical” in noncapital cases.

Courts are ill disposed toward ineffective-assistance petitions because they are the “inevitable” claim. Worse yet, they are unlimited in scope: they can be alleged against everything from egregious acts like pleading a man guilty against his will to (let us imagine) a simple Homeric nod over one possible objection that could have been made at some point during a ten-week trial. Strickland claims also evade the general prohibition on raising new arguments or presenting new evidence. A decade after Strickland, the Tenth Circuit, sitting en banc, described ineffectiveness claims as “the perfect tactical ‘open sesame’ to force re-reviews of close cases,” which exact a great toll, though

242. See United States v. Gaya, 647 F.3d 634, 638–39 (7th Cir. 2011)

The defendant who has a lawyer, even an incompetent one, must to establish a violation of his constitutional right to effective assistance of counsel prove that he was prejudiced by the lawyer’s incompetence . . . and that’s a lot harder to do than opposing a prosecutor’s claim of harmless error, for the prosecutor must prove the harmlessness of a constitutional error—and prove it beyond a reasonable doubt.

(citations omitted).


244. The term is Chief Judge Easterbrook’s. See United States v. Ramsey, 785 F.2d 184,193 (7th Cir. 1986) (“We are left with the inevitable claims of ineffective assistance of counsel.”); see also Wallace v. Davis, 362 F.3d 914, 919 (7th Cir. 2004) (“Thus we arrive at what seems to be the inevitable argument in capital cases: that counsel at sentencing was ineffective.”); Palermo v. United States, No. 98-2890, 1999 WL 417867, at *1 (7th Cir. June 17, 1999) (stating that the motion included “the inevitable staple of § 2255 litigation—a claim that his prior lawyer was ineffective”). The best figures show that in U.S. district courts ineffective-assistance claims are raised in 81% of capital cases and 50.4% of noncapital cases. Nancy J. King, Fred L. Cheesman II & Brian J. Ostrom, Final Technical Report: Habeas Litigation in U.S. District Courts 64 tbl.15 (2007), available at http://www.ncjrs.gov/pdffiles1/nij/grants/219559.pdf. This is consistent with the figures from 1990 and 1992. See Flango, supra note 241, at 45, 59 tbl.16.

245. See United States v. Pellerito, 701 F. Supp. 279, 281 (D.P.R. 1988) (claiming ineffective assistance of counsel because counsel allegedly encouraged the defendant to plead guilty against his will).
to little end, on courts, prosecutors, and defense lawyers alike. 246

In one case, the two standards, harmless error and *Strickland*, were actually considered side-by-side to review the same acts. 247 In *United States v. Herschberger*, 248 decided before the Tenth Circuit adopted *Strickland* for right-to-testify claims, the defendant alleged that he told his attorney that he wanted to testify but “counsel said he would decide.” 249 The defendant made both a right-to-testify claim on grounds that this right was suppressed and a *Strickland* claim for counsel’s failure to advise him that it was the defendant’s decision. 250 The court assumed that the first claim was amenable to harmless-error analysis; 251 the other claim was under *Strickland*. 252 The panel remanded for a hearing on the first claim but not the second. 253

Last year the Eleventh Circuit, apostle of *Strickland* in right-to-testify cases, was forced to revisit application of the harmless-error standard (as opposed to *Strickland*) because there was no assistance of counsel: the defendant acted pro se.

---

246. United States v. Galloway, 56 F.3d 1239, 1242 n.2 (10th Cir. 1995) (en banc).
248. Id.
249. Id. at *2.
250. See id. at *1, *3 (describing the defendant’s claims).
251. See id. at *2 (determining that the district court should have held an evidentiary hearing to assess whether the defendant’s right to testify was violated).
252. See id. at *3 (using the *Strickland* test to evaluate the defendant’s claim that his counsel’s failure to advise him of his right to testify amounted to ineffective assistance of counsel).
253. See id. at *4 (“Therefore, the district court did not abuse its discretion in failing to grant defendant a hearing on his claim for ineffective assistance of counsel. In the evidentiary hearing on remand, the district court need only consider whether defendant’s right to testify was violated.”). Petitioners still occasionally make this claim. See, e.g., Franklin v. United States, 227 F. App’x 856, 860 (11th Cir. 2007) (“Franklin makes a number of arguments on appeal, including that the question is not whether his testimony would have altered the final outcome, rather it is whether he was denied the right to testify. Franklin, however, must establish deficient performance and prejudice to obtain relief in this § 2255 motion.”).
WHY STRICKLAND IS THE WRONG TEST

In United States v. Hung Thien Ly, the court recalled the Fifth Circuit’s application of harmless error in the great case of Wright v. Estelle, noted that it since applied the ineffectiveness framework in Teague, and concluded that the issue “will be resolved in another case.”

A 2012 case in the Fifth Circuit illustrates our proposition about the decisive significance of the standard of review. In United States v. Wines, a man received a 35-year sentence for drug dealing. The Government had one witness, another dealer, who testified under a plea; the defense had one real witness, Wines’s mother. Wines never testified. In a postconviction hearing, Wines claimed he told his lawyer that he wanted a “chance to fight for [his] life” by testifying; his attorney recalled that they had decided that Wines would not testify because of prior convictions.

A dissenting Judge Higginbotham was sure that Strickland prejudice was established. “This is no easy

255. Wright v. Estelle, 549 F.2d 971 (5th Cir. 1977), aff’d, 572 F.2d 1071 (5th Cir. 1978) (en banc). All the same, the fractures over the right and the standard of review for errors were clear. The en banc decision said it “adheres to the panel opinion as published,” Wright, 572 F.2d at 1072, yet five judges, concurring, thought it a “disservice” to assume the existence of the right, and then declare a denial of it to be harmless error. Id. (concurring opinion). Three judges dissented on grounds that infringement of the right was structural error. See id. at 1080 (dissenting opinion) (arguing that harmless error should not be applied in this case). In Hung Thien Ly, Judge Tjoflat (by some accounts the longest serving active federal judge) counted six of fourteen judges who seemed to agree with harmless-error review. Hung Thien Ly, 646 F.3d at 1318 n.8. He himself was among the three dissenters some thirty-three years earlier! Id.; Wright, 572 F.2d at 1074 (dissenting opinion) (listing Judge Tjoflat as a dissenter).
256. Hung Thien Ly, 646 F.3d at 1318 n.8.
258. Id. at 601.
259. See id. at 600–01 (describing the witnesses’ testimony).
260. See id. at 601 (“The defense attorney did not call Wines to testify.”).
261. Id. at 602 (alteration in original).
262. See id. at 603 (discussing defense counsel’s assertion “that if Wines had been called to testify, the government would have quickly asked him about his prior conviction of a drug-related charge”).
263. See id. at 606 (Higginbotham, J., dissenting) (“I am persuaded that counsel’s failure to call Wines was not an objectively reasonable strategic decision and that Wines was prejudiced.”).
he wrote. Judges Jolly and Southwick, meanwhile, emphasized Strickland’s “heavy burden” and wrote that “as far as we can determine, no defendant in any court in the United States has been able to prove Strickland prejudice on the basis of his counsel advising him not to testify in his own defense at trial.”

The point is that the case was close enough to require a tense, forty-page published opinion from a divided panel. Would review under a harmlessness standard have produced a different result? The judges would have considered the effect of the attorney’s deficiency in failing to call the defendant (they agreed on that much) without speaking of a “heavy” burden—so heavy it has never been satisfied. Nor would the majority have referred repeatedly to the “presumption” of reasonable, strategic action by the lawyer. Losing under Strickland has a momentum of its own.

It is perhaps a final irony that when the Strickland Court chose to rely on a “materiality” standard (as with Brady violations) instead of a harmless-error standard, it did so out of a belief that the materiality standard was actually easier for

---

264. Id. at 620.
265. Id. at 606 (majority opinion).
266. See, e.g., id. at 603 (discussing the magistrate judge’s finding “that Wines had not overcome the strong presumption that his counsel’s decision was the result of sound trial strategy”).
267. See Strickland v. Washington, 466 U.S. 668, 694 (1984) (“[T]he appropriate test for prejudice finds its roots in the test for materiality of exculpatory information not disclosed to the defense by the prosecution . . . and in the test for materiality of testimony made unavailable to the defense by Government deportation of a witness . . . .” (citing United States v. Agurs, 427 U.S. 97, 104 (1976); United States v. Valenzuela-Bernal, 458 U.S. 858, 872–74 (1982))). United States v. Valenzuela-Bernal did not apply harmless error or constitutional standards of review. See United States v. Valenzuela-Bernal, 458 U.S. 858, 871 (1982) (“We thus conclude that the respondent can establish no Sixth Amendment violation without making some plausible explanation of the assistance he would have received from the testimony of the deported witnesses.”). United States v. Agurs explicitly avoided reliance on the harmless-error standard:

[S]ince we have rejected the suggestion that the prosecutor has a constitutional duty routinely to deliver his entire file to defense counsel, we cannot consistently treat every nondisclosure as though it were error. It necessarily follows that the judge should not order a new trial every time he is unable to characterize a nondisclosure as harmless under the customary harmless-error standard.

WHY STRICKLAND IS THE WRONG TEST

defendants to meet.268 But despite this solicitude for the fate of
defendants, the importation of that test into the area of denial of
the right to testify has resulted in a right without a remedy.

One might argue that shifting from Strickland to harmless
error will lose the competence prong of review for effective
assistance of counsel. But this is not a problem because, as stated
above, it is almost never within a lawyer’s authority to prevent
the defendant from exercising his personal right to testify.269

V. An Independent “Right to Testify,” a Proposed Test,
and Other Considerations

A. A Suggested Rule for Testing Claims and Ordering Hearings

We should simply speak of an independent “right to testify,”
an undisputed guarantee “implicit” in the Due Process, Self-
Incrimination, and Compulsory Process Clauses. It bears a
kinship to Faretta and the autonomy cases, but mostly it is an
application of the foremost right of all: the opportunity to be
heard.270 Usually the accused speaks through his lawyer, but
when a lawyer actively prevents a determined defendant from
testifying, has that defendant really been heard? Claims about
abridgement of the right to testify deserve a freestanding inquiry,
decoupled from Strickland. Courts already use the harmless-error
standard when a statute, a judge, or a prosecutor is to blame.271
Our argument is to use this standard with defense errors, too.

268. See Strickland, 466 U.S. at 697 (comparing the burden on the
defendant imposed by the materiality standard to that of other tests).

269. We say “almost” because there are scenarios that could prove
exceptions to the rule. Suppose a man is on trial while under medication for
mental illness and one day is given the wrong pills. If in a schizophrenic state he
demands to take the stand, the lawyer might properly reject his request and
later claim that the defendant was not acting voluntarily or was incompetent to
make the decision.

270. See Windsor v. McVeigh, 93 U.S. 274, 277 (1876) (“A sentence of a court
pronounced against a party without hearing him, or giving him an opportunity
to be heard, is not a judicial determination of his rights, and is not entitled to
respect in any other tribunal.”).

271. See supra notes 99–106 and accompanying text (discussing cases in
which a statute, a court, or a prosecutor caused error). And Rock’s balancing test
applies in cases of conflict between a statute and the right. See Rock v.
Arkansas, 483 U.S. 44, 54–56 (1987) (balancing the interest served by a state’s
The rule might read:

A defendant has a right at trial to testify truthfully in his own defense, contrary to advice of counsel, no matter the evidentiary or strategic detriment, if he timely and clearly announces his desire to do so to his attorney. A petitioner is entitled to a hearing as to the abridgment of this right if he makes a strong showing, based on specific allegations, that his attorney deprived him of this right.272

In collateral attack it is dangerously easy for a disgruntled convict to allege that he got bum advice;273 or that he was pressured into abandoning his chance to testify;274 or that some medical condition kept him off the stand;275 or that he should have been asked to waive on the record, etc.276 Before a time-

evidentiary rule with the limitation imposed on the defendant’s constitutional right to testify).

272. We use the term “at trial” here, but one might also include other occasions, like crucial pre-trial proceedings.

273. See, e.g., United States v. Hood, No. 88-4046, 1989 WL 102017, at *1 (9th Cir. Aug. 25, 1989) (stating that the defendant claimed his attorney “provided ineffective assistance by giving [the defendant] inaccurate information”).

274. See, e.g., Nichols v. Butler, 953 F.2d 1550, 1552 (11th Cir. 1992) (describing how the defendant’s attorney pressured the defendant into not testifying by threatening to withdraw mid-trial).

275. See, e.g., United States v. Pondelick, No. 11-30057, 2012 WL 907488, at *1 (9th Cir. Mar. 19, 2012) (noting that the defendant claimed that an abscessed tooth and infection kept him off the stand, even though he had stated earlier after “lengthy colloquy” that he did not wish to testify).

276. See, e.g., United States v. Aldea, 450 F. App’x 151, 152 (3d Cir. 2011) (“Aldea admitted on cross-examination, ‘[b]asically, at the end, yeah, it was my decision [not to testify] . . . .’”); Lott v. Attorney General of Florida, 594 F.3d 1296, 1302–03 (11th Cir. 2010) (stating that the defendant replied “Yes, ma’am” when the court asked whether it was “a joint choice . . . that [he] would not testify in the trial); United States v. Bailey, 245 F. App’x 768, 770–71 (10th Cir. 2007) (rejecting a claim of ineffective assistance because petitioner’s affidavit was “ambiguous,” and counsel submitted a “detailed” account advising Mr. Bailey to keep silent and listing the “numerous reasons for not wanting to testify,” like dangers on cross-examination). Further, despite several opportunities, Mr. Bailey “never once suggested to the court that he wished to testify.” Id. at 771; see also Winfield v. Roper, 460 F.3d 1026, 1035 (8th Cir. 2006) (“Winfield’s claim that he did not waive his right to testify was fully explored at the evidentiary hearing in the post-conviction proceedings, and the state circuit court found that penalty counsel’s testimony was credible, unlike that of Winfield, Kessler, and Bates.”); Gonzales v. Elo, 233 F.3d 348, 357 (6th Cir. 2000) (“Indeed, as the magistrate noted, Petitioner did not raise this claim until nearly six years after his conviction, and after his appeal as of right was
WHY STRICKLAND IS THE WRONG TEST

A consuming probe into uncertain credibility and faded (but self-serving) memory is permitted—sometimes years after the events in question—a defendant must be required to offer a detailed, who-said-what account of how his right was denied. In rare circumstances the district court might require an affidavit from the lawyer. If a persuasive showing is made, an evidentiary hearing could find facts and assess the probable effect of the ungiven testimony. Although a court sometimes may violate the right, violations by counsel pose the special problem of unreviewability, since his or her acts are almost always off the record. Unlike other “personal” rights—to plead guilty, to defend pro se, to appeal, to waive jury trial—the right to testify is usually waived behind closed doors, without any assurance that the client understands his rights beyond what his lawyer tells him.

Then a district court would need to make a preliminary inquiry under a harmless-error standard into whether it would have mattered. This would turn on the strength of the Government’s evidence, the likely significance to jurors of a defendant’s appearance on the stand, the centrality (or not) of the defendant’s ungiven testimony, the probable harms of cross-rejected[,] . . . let alone voice any such concern at trial.”); El-Tabech v. Hopkins, 997 F.2d 386, 388 (8th Cir. 1993) (“Although El-Tabech asserts that he raised his hand several times at trial and was admonished by the judge to address the court only through his attorneys, our review of the trial transcript has uncovered no reference to these events.”).

277. See Underwood v. Clark, 939 F.2d 473, 476 (7th Cir. 1991) (stating that “particularity” is necessary to give defendant’s claim sufficient credibility to warrant further judicial investment); Siciliano v. Vose, 834 F.2d 29, 31 (1st Cir. 1987) (finding that no hearing was required over a claim that the defendant was forbidden to testify because no “specific” and “credible” allegations of compulsion by counsel were provided in the record).

278. For instance, in Chang v. United States, the court found that the defendant’s affidavit claiming his right to testify was not explained to him, although “generic,” could possibly have merited a hearing. Chang v. United States, 250 F.3d 79, 86 (2d Cir. 2001). But because the district court supplemented the record with a “detailed affidavit from trial counsel credibly describing the circumstances concerning appellant’s failure to testify,” it was justified in dismissing the complaint without a hearing. Id. at 85. But Jiles v. United States considered whether a defendant should be obliged to get an affidavit from his lawyer. Jiles v. United States, 72 F. App’x 493, 494 (7th Cir. 2003). The circuit held that a defendant’s “specific affidavit does not need corroboration from the very person accused of wrongdoing.” Id. (citing Taylor v. United States, 287 F.3d 658 (7th Cir. 2002)).
examination and impeachment, and the jury's presumptive view of the defendant's credibility. If there was an erroneous abridgement of the right, and it might have mattered, a record now exists for district and circuit courts to consider alongside prior proceedings.

B. The Waiver Question

The right-to-testify waiver jurisprudence is well-settled. All circuits agree that defendants need not waive on the record and that a court is not obliged to explain the right to the defendant.

279. In Ortega v. O'Leary, the panel felt that Van Arsdall, a Confrontation Clause case, sets the framework for determining whether an error is harmless beyond a reasonable doubt under Chapman: "[T]he factors [to consider] include the importance of the witness' testimony to the . . . case, whether the testimony was cumulative, the presence or absence of evidence corroborating or contradicting the testimony of the witness on material points . . . and of course, the overall strength of the . . . case."

280. See Arredondo v. Huibregtse, 542 F.3d 1155, 1165 (7th Cir. 2008) ("[T]he Supreme Court of the United States never has held that a trial court must engage in a personal colloquy with a defendant to determine whether he wishes to testify or that a waiver of the right to testify must occur formally on the record."); see also Berkovitz v. Minnesota, 505 F.3d 827, 828 (8th Cir. 2007) (declining to adopt a rule requiring all defendants who do not testify to waive the right on the record); United States v. Stover, 474 F.3d 904, 908 (6th Cir. 2007) (stating that defendant's waiver of his right to testify does not require a colloquy with the court); United States v. Glenn, 389 F.3d 283, 287 (1st Cir. 2004) (affirming that the trial court is not constitutionally required to advise the client of his right to testify); United States v. Manjarrez, 258 F.3d 618, 623 (7th Cir. 2001) ("[W]e have repeatedly held that the Constitution does not require a trial court to question a defendant sua sponte in order to ensure that his decision not to testify was undertaken knowingly and intelligently . . . .''); United States v. Brown, 217 F.3d 247, 258 (5th Cir. 2000) (agreeing with the majority of courts that say a "district court generally has no duty to explain to the defendant that he or she has a right to testify or to verify that the defendant who is not testifying has waived the right voluntarily"); United States v. Richardson, 195 F.3d 192, 197–98 (4th Cir. 1999) (stating that a trial court generally has no duty to inform a defendant of his right to testify); United States v. Leggett, 162 F.3d 237, 246 (3d Cir. 1998) ("[A] trial court not only has no duty to make an inquiry but, as a general rule, should not inquire as to the defendant's waiver of the right to testify."); United States v. Van De Walker, 141 F.3d 1451, 1452 (11th Cir. 1998) (affirming that the appellate court is not
WHY STRICKLAND IS THE WRONG TEST

unless perhaps if he is a pro se defendant.281 This is because while a judge must inquire of the defendant before taking a guilty plea or allowing him to waive jury trial or forgo assistance of counsel,282 there is too great a risk in this context of impeding his lawyer’s strategy, interfering with the client–counsel relationship, leading the defendant to believe that testifying is being suggested, or tempting him to waive the right not to testify.283 A defendant may not know whether he wishes to testify.

---

281. See United States v. Hung Thien Ly, 646 F.3d 1307, 1317 (11th Cir. 2011)

[The district court was required to correct Ly’s misunderstanding of his right to testify. By not informing Ly that he could testify in narrative form, the district court denied his right to choose whether to testify “knowingly and intelligently.” This case falls within one of the “exceptional, narrowly defined circumstances” that trigger a district court’s duty to discuss with a criminal defendant his decision of whether to testify.

282. See United States v. Ortiz, 82 F.3d 1066, 1070–71 (D.C. Cir. 1996) (citing Supreme Court cases that require this inquiry).

283. See United States v. Webber, 208 F.3d 545, 551–52 (6th Cir. 2000) (recognizing that waiver of the right to testify “qualitatively differs” from the right to enter a plea of guilty, waive a jury trial, or forego the assistance of
until he hears the State’s evidence; the court will not know until
the defense rests. 284 Waiving at trial is unlike waiving at a plea
hearing, for a plea is “itself a conviction.” 285 A colloquy is probably
necessary only when counsel or defendant tips off a court that a
dispute between them over testifying has arisen. 286 A colloquy
then helps to prove waiver later on. 287 If a colloquy occurs,
however, caution is necessary so that the court explains to the
defendant that he possesses the right without venturing into
comment on advisability. 288

284. And at this point, of course, a desire to testify would come too late. See
Pennycooke, 65 F.3d at 11 (stating that “a colloquy on the right to testify” when
the defense rests “not only would be awkward . . . but more importantly
inadvertently might cause the defendant to think that the court believes the
defense has been insufficient” (citation omitted)).


286. See Arredondo v. Huibregtse, 542 F.3d 1155, 1157 (7th Cir. 2008)
(involving such interruptions); Ortega v. O'Leary, 843 F.2d 258, 260 (7th Cir.
1988) (same). The Seventh Circuit's jurisdiction seems stocked with unruly
defendants.

287. See United States v. Ramsey, 785 F.2d 184, 194–95 (7th Cir. 1986)
(finding that despite the defendant’s claim that his trial counsel would not allow
him to take the stand, the defendant “bypassed an opportunity to support such a
claim in the district court”). The court noted that “[n]othing in the record
supports the assertion that counsel thwarted [the defendant’s] desire to testify,”
and “[t]o the contrary, the district judge asked [the defendant] during the trial
whether he wanted to testify.” Id.

288. See, e.g., United States v. Hung Thien Ly, 646 F.3d 1307, 1315 (11th
Cir. 2011) (“Experience demonstrates that district courts sometimes provide
inappropriate commentary when they inject themselves into a defendant’s
choice of whether to testify.”); United States v. Yono, 605 F.3d 425, 426 (6th
Cir. 2010) (advising against a colloquy into the defendant’s waiver of his
right to testify); United States v. Leggett, 162 F.3d 237, 248 (3d Cir. 1998)
(stating that a trial court should avoid encouraging or discouraging a
defendant to testify); United States v. Pennycooke, 65 F.3d 9, 13 (3d Cir.
1995) (noting that sometimes a trial court might have to “inquire discreetly”
to ensure that a defendant’s constitutional rights are not suppressed);
Joelson, 7 F.3d at 178 (discussing the danger of “judicial interference
WHY STRICKLAND IS THE WRONG TEST

The waiver inquiry tests whether the defendant made a knowing, intelligent relinquishment. In nearly every case in which a defendant does not testify or inform the court of a desire to do so, waiver should be presumed or inferred. Defense counsel would always have primary responsibility for advising the defendant on the right and the implications of exercising it. A good test must catch instances in which a defendant is stripped of his right without adding new layers to the already extensive protections for criminal defendants.

289. See United States v. Goodwin, 770 F.2d 631, 635–38 (7th Cir. 1985) (stating that “[t]he trial judge . . . went beyond his limited function of ensuring that Goodwin’s decision not to testify was voluntary when he expressed surprise . . . , explained some of the pros and cons of her taking the stand, and strongly implied that her only chance for acquittal was to testify”); United States v. Leggett, 162 F.3d 237, 246 (3d Cir. 1998) (“If a defendant does waive this right [to testify], the waiver must be knowing, voluntary and intelligent.”).

290. See United States v. Ly, 646 F.3d 1307, 1313 (11th Cir. 2011) (“Like other fundamental trial rights, the right to testify is truly protected only when the defendant makes his decision knowingly and intelligently.”); United States v. Leggett, 162 F.3d 237, 246 (3d Cir. 1998) (“If a defendant does waive this right [to testify], the waiver must be knowing, voluntary and intelligent.”).

291. See Hung Thien Ly, 646 F.3d at 1313 (“In cases where a defendant is represented by counsel, counsel is responsible for providing the advice needed to render the defendant’s decision on whether to testify knowing and intelligent.”); United States v. Ortiz, 82 F.3d 1066, 1070 (D.C. Cir. 1996) (“Thus, defense counsel, not the court, has the primary responsibility for advising the defendant of his right to testify and for explaining the tactical implications of doing so or not.”).
The right to testify remains circumscribed by other legitimate trial interests and rules of evidence or procedure. These operate to disallow testimony on subjects ruled inadmissible after in limine motions to prevent false testimony, to prohibit unreliable forms of evidence, to maintain fair, orderly proceedings to ensure that any testimony is given after oath or affirmation, to keep an unruly defendant from degrading the trial, or perhaps to inhibit a mentally defective defendant from sinking himself on the stand. A defendant may not claim that his right to testify is

---

292. See United States v. Bifield, 702 F.2d 342, 350 (2d Cir. 1983) (agreeing with the district court that although defendant wanted to testify about duress, such testimony was inadmissible because irrelevant under Fed. R. Evid. 402).

293. See United States v. Flores-Martinez, 677 F.3d 699, 709 (2012) (considering the rights of a defendant seeking to testify on matters specifically excluded by the ruling of the trial court).


295. See Rock v. Arkansas, 483 U.S. 44, 56 (1987) (“In applying its evidentiary rules a State must evaluate whether the interests served by a rule justify the limitation imposed on the defendant’s constitutional right to testify.”). Arkansas could not, without better evidence, ban entirely a defendant’s use of hypnotically refreshed memory. See id. at 61.

296. See United States v. Jones, 880 F.2d 55, 59–60 (8th Cir. 1989) (allowing a court to exercise discretion, in the interests of fairness and order, to refuse to reopen evidence for testimony when defendant asserted his right to testify only after the evidence-taking stage of the trial was closed, though before it was sent to the jury). A court might find that the defendant is trying to engage in delay or other improper purpose.

297. See United States v. Looper, 419 F.2d 1405, 1406 (4th Cir. 1969) (disallowing a defendant who was a conscientious objector from testifying because of a religious inability to swear on the Bible and raise his hand). The circuit remanded for a new trial, instructing that “all the district judge need do is to make inquiry as to what form of oath or affirmation would not offend defendant’s religious beliefs but would give rise to a duty to speak the truth.” Id. at 1407.

298. See United States v. Bentvana, 319 F.2d 916, 944 (2d Cir. 1963) (discussing the misbehaving defendant’s right to testify). The defendant had engaged in “outbursts and serious misconduct” and the judge was bound to “maintain the order and decorum necessary for a fair trial.” Id. But the circuit nonetheless found that the court should have allowed the testimony “[i]n view of the importance of the privilege” and instead have had a marshal subdue or gag him if he continued to insist on disobeying court orders. Id.

299. See People v. Robles, 466 P.2d. 710, 716 n.1 (Cal. 1970) (“In some situations a defendant’s persistence in testifying contrary to his attorney’s advice might raise questions as to the defendant’s present sanity.”); see also
tread upon by rules allowing cross-examination or impeachment with prior convictions and bad acts. The Supreme Court has also held that obstruction of justice enhancements are no burden on the right. An accused who testifies is in the position of any other witness: duty-bound to speak truthfully, entitled to the same privileges, and exposed to the same perils of impeachment, stress, embarrassment, and so on.

It is well-established that the right to testify may be waived only by the defendant. According to Chief Judge Frank Ward v. Sternes, 334 F.3d 696, 699–700 (7th Cir. 2003)

Ward’s counsel admitted that it wasn’t his client, but he who had made the decision to keep Ward off the stand. He further told the court that he didn’t believe he could have an informed discussion with Ward about the decision, since most of his prior exchanges with his client were one-sided, generating only an occasional ‘uh-uh’ response from Ward.

300. See United States v. Siddiqui, 699 F.3d 690, 705–06 (2d Cir. 2012) (discussing how an apparently unstable defendant took the stand, to her detriment, over counsel’s strenuous urging). On appeal, Ms. Siddiqui’s lawyer argued that “in some cases a defendant may be competent to stand trial yet incompetent to exercise her right to testify without the approval of defense counsel.” Id. at 705. The court declined to decide this question. Id. at 705–06. The case is also a fine example of the lengths to which a zealous defense lawyer and a concerned judge can go to help a defendant avoid self-inflicted wounds, short of compulsion. See id. at 698 n.4 (discussing the lengths to which they went). See also People v. Robles, 466 P.2d 710, 716 n.1 (Cal. 1970) (“In some situations a defendant’s persistence in testifying contrary to his attorney’s advice might raise questions as to the defendant’s present sanity.”); Ward v. Sternes, 334 F.3d 696, 699–700 (7th Cir. 2003)

Ward’s counsel admitted that it wasn’t his client, but he who had made the decision to keep Ward off the stand. He further told the court that he didn’t believe he could have an informed discussion with Ward about the decision, since most of his prior exchanges with his client were one-sided, generating only an occasional ‘uh-uh’ response from Ward.

301. See United States v. Dunnigan, 507 U.S. 87, 96 (1993) (rejecting the defendant’s argument that courts will enhance sentences whenever the accused takes the stand and is found guilty).

302. See United States v. Babul, 476 F.3d 498, 500 (7th Cir. 2007) (ruling that only the defendant may waive the right to testify in his own defense (citing Rock v. Arkansas, 483 U.S. 44 (1987))). Dozens of cases confirm this proposition. See, e.g., Galowski v. Murphy, 891 F.2d 629, 636 (7th Cir. 1989) (providing an early post-Strickland, post-Rock statement of this rule of law). Teague acknowledged that the “right is personal to the defendant and cannot be waived either by the trial court or by defense counsel . . . . [T]here can be no effective waiver of a fundamental constitutional right unless there is an ‘intentional relinquishment or abandonment of a known right of privilege.’” United States v.
Easterbrook, the distinction between this and the ordinary rights of trial management (whether to call a witness, object to hearsay, etc.), is that the former are choices in which “one does not need a legal education to appreciate the issues.”

The presumption against the waiver of constitutional rights is why a judge who accepts a plea must ensure that the defendant grasps what he is waiving. Federal Rule of Criminal Procedure 11 requires a court to inquire about the defendant’s awareness of his rights to plead not guilty, have a jury trial, and secure counsel. Another provision of Rule 11 directs the court to inform the defendant of his “right at trial to confront and cross-examine adverse witnesses, to be protected from compelled self-incrimination, to testify and present evidence, and to compel the attendance of witnesses.” Each of the rights just listed appears in Madison’s Bill of Rights—except the “right to testify.” Yet today a plea is invalid unless the court ensures waiver of the right “personally in open court.”

Finally, any rule requiring a defendant to gainsay his attorney in front of a judge to avoid an inference of waiver seems unfair; if anyone, it is the lawyer’s job to expose disagreement.

---

303. Babul, 476 F.3d at 500.
304. See Schneckloth v. Bustamonte, 412 U.S. 218, 241 (1973) (“A strict standard of waiver has been applied to those rights guaranteed to a criminal defendant to insure that he will be accorded the greatest possible opportunity to utilize every facet of the constitutional model of a fair criminal trial.”); Glasser v. United States, 315 U.S. 60, 70–71 (1942) (stating that the court must “indulge every reasonable presumption against the waiver of fundamental rights” but must also protect those rights by ensuring that the waiver is “intelligent” and “competent”); see also Brookhart v. Janis, 384 U.S. 1, 5 n.4 (1966) (“When constitutional rights turn on the resolution of a factual dispute we are duty bound to make an independent examination of the evidence in the record.”).
305. At the same time, there can really be no presumption against waiver of the right to testify without presuming an intent to abandon the right to remain silent. No presumptions should apply to this choice between silence and speech.
307. Id. 11(b)(1)(E) (emphasis added). The right to testify is recognized elsewhere in the rules. See, e.g., id. 12.3(c) (stating that a public-authority defense can be barred under certain circumstances but that this “does not limit the defendant’s right to testify”).
308. Id. 11(b)(2).
309. See United States v. Martinez, 883 F.2d 750, 760–61 (9th Cir. 1989) (stating that a defendant may not waive his right to testify under his lawyer’s
Instances in which a defendant does pipe up seem to involve uncommonly self-assured men, like the one who shouted of his lawyer—“He lies. Lies. Lies.”—until the court threatened to remove him.\textsuperscript{310} From the case law it also appears that a trial judge is likelier to regard such efforts as irksome impertinence rather than an alarm that a right is going unheeded.\textsuperscript{311}

\section*{C. Due Process Clause Inquiry?}

The Due Process Clause generally applies to violations committed by the state, not a lawyer.\textsuperscript{312} One might argue that because the State brings the case (prosecutor), conducts the trial advice and then try to invalidate the trial because he decided to abide by his lawyer's counsel), \textit{vacated}, 928 F.2d 1470 (9th Cir. 1991); \textit{see also} United States v. McMeans, 927 F.2d 162, 163 (4th Cir. 1991) (holding that waiver was appropriate partly because the defendant’s testimony at an earlier trial evinced knowledge of the right to testify); United States v. Bernloehr, 833 F.2d 749, 751–52 (8th Cir. 1987) (emphasizing that the defendant did not object when his counsel rested without calling him to the stand). \textit{But see} Chang v. United States, 250 F.3d 79, 84 (2d Cir. 2001) (recognizing the inconsistency of requiring a defendant to rely on his attorney to speak for him in the courtroom and holding that by failing to speak the defendant has waived his right to testify); United States v. Ortiz, 82 F.3d 1066, 1071 (D.C. Cir. 1996) (recognizing that it is impracticable to put a burden on the defendant who might not even be aware of the right he possesses and that the burden would conflict with the instruction that a defendant should speak through his counsel); Underwood v. Clark, 939 F.2d 473, 476 (7th Cir. 1991) (refusing to treat a defendant’s silence as a waiver of the right to testify because the defendant might feel “too intimidated to speak out of turn in this fashion”); United States v. Teague, 908 F.2d 752, 759–60 (11th Cir. 1990) (refusing to find “that by failing to speak out at the proper time a defendant has made a knowing, voluntary and intelligent waiver of a personal right of fundamental importance such as the right to testify”), \textit{vacated}, 932 F.2d 899 (11th Cir. 1991).

\textsuperscript{310.} \textit{See} Arredondo v. Huibregtse, 542 F.3d 1155, 1157–64 (7th Cir. 2008) (containing a tense exchange); \textit{see also} United States v. Leggett, 162 F.3d 237, 254 (3d Cir. 1998) (McKee, J., dissenting) (“How could he make such a request? Leggett could not very well have disrupted the proceedings by standing in open court and speaking directly to the judge without being asked anything.”).

\textsuperscript{311.} \textit{See}, \textit{e.g.}, \textit{Leggett}, 162 F.3d at 254 (noting that the judge “feared” that Leggett might “jump to his feet and assert his right to testify”).

\textsuperscript{312.} \textit{See} County of Sacramento v. Lewis, 523 U.S. 833, 845–46 (1998) (describing the Due Process Clause as limiting what the government may do); Lochner v. New York, 198 U.S. 45, 52 (1905) (stating that under the Due Process Clause “no state can deprive any person of life, liberty, or property without due process of law”).
(judge), and executes the sentence (warden), failure to ensure a trial in which a defendant’s right to testify is respected is state action. This is the theory behind why certain attorney conduct—like failing to disclose a material conflict of interest—still implicates state action.\footnote{313} But again, a stand-alone “right to testify” inquiry is needed.\footnote{314}

VI. Conclusion

Taking the long view, the constitutional right to testify is a novelty. Over three centuries, what was prohibited to a defendant became a statutory privilege, then a sort of assumed right, and finally an explicit constitutional command.\footnote{315} In federal courts it took a century for the statutory privilege (1878)\footnote{316} to harden into constitutional right (1987).\footnote{317} During the 1960s and ‘70s it was

\footnote{313. \textit{See} Cuyler v. Sullivan, 446 U.S. 335, 343 (1980) (“This Court's decisions establish that a state criminal trial, a proceeding initiated and conducted by the State itself, is an action of the State within the meaning of the Fourteenth Amendment.”).

314. Under the Sixth Amendment’s “presumed prejudice” line, courts decline to scour the record to ascertain the degree of prejudice. These are cases where, for instance, a lawyer carelessly forfeited an appeal. \textit{See} Roe v. Flores-Ortega, 528 U.S. 470, 483 (2000) (“The even more serious denial of the entire judicial proceeding itself . . . similarly demands a presumption of prejudice.”). Or where a lawyer failed to comply with \textit{Anders} and left his client uncounseled on appeal. \textit{See} Penson v. Ohio, 488 U.S. 75, 88–89 (1988) (“It is therefore inappropriate to apply either the prejudice requirement of \textit{Strickland} or the harmless-error analysis of \textit{Chapman.”). Presumed prejudice is a form of structural error: both involve mistakes that foreclosed an avenue, disabled the creation of a record, or raised barriers to determining what might have happened. We feel prejudice should not be presumed for the same reasons the error here is not structural. Every circuit has rejected the claim, occasionally made, for presumed prejudice. \textit{See, e.g.,} Palmer v. Hendricks, 592 F.3d 386, 396–97 (3d Cir. 2010) (determining there is no presumption of prejudice for right-to-testify cases).

315. Few would argue that if competency statutes in the states and U.S. Code were repealed, courts of this country would once again be permitted to deny defendants the power to choose to testify. It might also be remembered that both the right to testify and the common law rule against defendant testimony were alike intended to bring out the truth and protect the defendant (which, as noted, was once thought best done by prohibiting all defendants from testifying, so that no harmful inferences for failure to speak would be made by the jury).


WHY STRICKLAND IS THE WRONG TEST

still treated as a trial tactic. But when finally constitutionalized it became necessary for courts to decide how to review alleged infringements. Because it seemed like a matter of lawyer ineffectiveness, the U.S. courts of appeals, in the 1990s, reached for the familiar Strickland standard. The problem, we see, is that Strickland frustrates the right. Rarely will an overbearing attorney’s conduct provoke a postconviction inquiry; almost never will it lead to a new result. The court is usually satisfied to concede a violation of right while squinting in vain to find prejudice. And no wonder: the lawyer is almost always right to restrain his client. Yet the proper avenue is not Strickland but an independent, constitutional inquiry. The first court to do so gets to name it.

318. See United States v. Von Roeder, 435 F.2d 1004, 1009 (10th Cir. 1971) ("[T]he decision of counsel to place or not to place his client on the stand has been described as particularly difficult, and has generally been treated as a question of trial tactics."); see also United States v. Garguilo, 324 F.2d 795, 797 (2d Cir. 1963) (expressing the statement of old view, before Rock and Strickland, that denial of right to testify is strategic and there is no remedy for strategic error); Seth Dawson, Due Process v. Defense Counsel’s Unilateral Waiver of the Defendant’s Right to Testify, 3 Hastings L.Q. 517, 523 (1975) ("There is a growing recognition that the right to testify has transcended its statutory origins and is now emerging as a constitutionally protected right, inherent in the ever-broadening concept of due process."). In 1993, the Ninth Circuit said it “essentially is a strategic trial decision with constitutional implications.” United States v. Joelson, 7 F.3d 174, 178 (9th Cir. 1993).
Vertical Boilerplate

James Gibson*

Abstract

Despite what we learn in law school about the “meeting of the minds,” most contracts are merely boilerplate—take-it-or-leave-it propositions. Negotiation is nonexistent; we rely on our collective market power as consumers to regulate contracts’ content. But boilerplate imposes certain information costs because it often arrives late in the transaction and is hard to understand. If those costs get too high, then the market mechanism fails.

So how high are boilerplate’s information costs? A few studies have attempted to measure them, but they all use a “horizontal” approach—i.e., they sample a single stratum of boilerplate and assume that it represents the whole transaction. Yet real-world transactions often involve multiple layers of contracts, each with its own information costs. What is needed, then, is a “vertical” analysis, a study that examines fewer contracts of any one kind but tracks all the contracts the consumer encounters, soup to nuts.

This Article presents the first vertical study of boilerplate. It casts serious doubt on the market mechanism and shows that existing scholarship fails to appreciate the full scale of the information cost problem. It then offers two regulatory solutions. The first works within contract law’s unconscionability doctrine, tweaking what the parties need to prove and who bears the burden of proving it. The second, more radical solution involves forcing

* Professor of Law and Director of the Intellectual Property Institute, University of Richmond. My thanks to Suzanne Corriell, Chris Cotropia, David Epstein, Jessica Erickson, David Frisch, Jeanne Fromer, Meredith Harbach, Ann Hodges, Sonia Katyal, Corinna Lain, Shari Motro, Kristen Osenga, Kimberly Robinson, Noah Sachs, Roberta Sachs, Jessica Silbey, and Kevin Walsh; to the participants at WIPIP 2011, the Fordham Intellectual Property Colloquium, the 2011 Texas IPLJ symposium, the Richmond Law Faculty Colloquy, and the New England Intellectual Property Colloquium; to the Hunton and Williams Summer Research Endowment Fund; and of course to Jane Savoca, the best deal I ever made.
both sellers and consumers to confront and minimize boilerplate’s information costs—an approach I call “forced salience.” In the end, the boilerplate experience is as deep as it is wide. Our empirical work should reflect that fact, and our policy proposals should too.

Table of Contents

I. Introduction ............................................................... 163

II. The State of the Debate .............................................. 166
   A. The Common Ground: Contracts as Product Features ........................................ 166
   B. The Battleground: Market Function or Market Failure ........................................ 169
      1. Information Acquisition ......................................... 171
      2. Information Processing ........................................ 174

III. A Vertical Study of Boilerplate ................................. 180
   A. Verticality’s Advantages ......................................... 180
   B. Study Design .......................................................... 184
      1. Assumptions About Enforceability ........................................ 185
      2. Assumptions About the Scope of the Transaction ........................................ 189
   C. Study Results ............................................................ 190
      1. Information Acquisition Costs ......................................... 192
      2. Information Processing Costs ........................................ 195

IV. Responding to Boilerplate’s Defenders .................... 199
   A. Informed-Minority Theory ......................................... 199
   B. Reputational Theory .................................................. 205

V. Toward Enforceable Boilerplate ............................... 212
   A. Two Extremes .......................................................... 212
      1. Get Rid of All Boilerplate ......................................... 212
      2. Enforce All Boilerplate ........................................ 215
   B. Two In-Betweens ...................................................... 218
      1. Unconscionability .................................................. 218
      2. Forced Salience .................................................... 224

VI. Conclusion .............................................................. 228

VII. Appendix: Boilerplate Breakdown ............................. 230
I. Introduction

Forget what your contracts professor told you about offers, counteroffers, and meetings of the minds. Most contracts are merely boilerplate—i.e., take-it-or-leave-it propositions. In theory, that’s fine because the same is true of most market transactions. We negotiate over almost nothing. No one haggles with a supermarket cashier over the price of a loaf of bread, or how thinly it is sliced, or whether it’s covered by a warranty. Instead, we rely on our collective power as consumers to drive unwanted terms (contractual or otherwise) out of the marketplace. Competition, not negotiation, is the answer.

Scholars have observed for some time, however, that when it comes to boilerplate contracts, market competition may not work as well as theory would have us believe.1 Boilerplate, these observers maintain, is particularly resistant to market forces because of the high information costs it imposes on consumers: it often arrives late in the transaction, and once it arrives it often consists of overlong, impenetrable gobbledygook. This means that consumers routinely disregard boilerplate in their purchasing decisions, which in turn means that the market cannot be trusted to regulate it.2

Recent empirical evidence supports this view. But the studies so far have all been “horizontal”—they examine a single stratum of boilerplate (e.g., an array of software licensing agreements) and assume that it represents the whole transaction. Such studies are certainly useful in that they reveal some of the information costs that boilerplate creates for consumers. But they capture only a single moment in time, one aspect of the overall

---


2. See Michael I. Meyerson, The Efficient Consumer Form Contract: Law and Economics Meets the Real World, 24 GA. L. REV. 583, 601 (1990) (“[T]here generally will be too few informed consumers to produce a competitive market for contract terms.”).
consumer experience. A more realistic approach would recognize that a consumer's ability to evaluate boilerplate is a function of the information costs of the entire purchase, soup to nuts. What is missing, then, is a “vertical” analysis—a study that examines fewer contracts of any one kind, but places the consumer’s encounter with those contracts in a more representative, real-world context, in which the boilerplate represents part of the transaction rather than its entirety. In short, a horizontal study is a still life, whereas a vertical study is a film—a moving picture of the consumer’s entire transaction.

This Article presents the results of the first-ever vertical study of boilerplate. The subject of the study, like the subject of the emerging empirics and the foundational case law in the field, is the computer industry. To gauge overall information costs, I purchased ordinary desktop computers from four major vendors, tallied every word of boilerplate to which I became contractually bound, and recorded the point in the transactions at which each term arrived. The result? An average of twenty-five different contracts, comprising almost as many words as a *Harry Potter* novel. And nine out of every ten of those boilerplate terms arrived late in the transaction, long after the seller had been paid.

This study casts doubt on the law’s current approach to boilerplate terms, which is to enforce them whenever minimal disclosure and assent requirements have been satisfied.\(^3\) And when the study’s results are situated in the current scholarship, they show that even those who doubt the market’s efficacy fail to appreciate the full scale of the information cost problem. In the end, these observations lead me to offer two regulatory solutions. The first and more conventional consists of tweaks to the unconscionability doctrine—modest changes in what the parties need to prove under the doctrine and who bears the burden of proving it. The second, more radical solution involves forcing both sellers and consumers to confront and minimize the information

\(^3\) See Korobkin, *supra* note 1, at 1204 (“If the non-drafting party indicates his general assent to the form, courts will enforce the terms contained therein whether or not that party approves of the terms provided, understands those terms, has read them, or even has the vaguest idea what the terms might be about.”).
costs that boilerplate creates—an approach I call “forced salience.”

Two clarifications before we continue. First, as the discussion so far implies, I use the term “boilerplate” to mean any written contract drafted by one party and not subject to revision by the other (what others variously call form contracts or contracts of adhesion). For my purposes, then, it does not matter whether boilerplate terms are the same across a given industry or whether they vary from seller to seller. As long as a seller’s contract is a take-it-or-leave-it proposition, it falls within my definition.

Second, this Article examines information costs and their effect on the market from the point of view of the individual consumer—“bottom up,” so to speak. It therefore does not make “top down,” legislative-style judgments about what particular terms are bad for consumers as a class and should therefore be forbidden by fiat. Such judgments have their place, but this is an Article about boilerplate, not about arbitration clauses or class action waivers or hidden credit card fees (important though those topics are). Indeed, as we will see, even boilerplate that goes unread can contain provisions that benefit both parties, and it is not the province of the law to forbid such terms, even if they emerge from an impaired market.

The Article proceeds as follows. Part II summarizes the current debate between those who believe that the market for boilerplate functions reasonably well and those who don’t, and it discusses emerging empirical evidence on the issue. Part III adds a new dimension to those empirics by presenting the results of my vertical study and demonstrating the overwhelming information costs that a real-world transaction imposes on consumers. Part IV uses this real-world perspective to review and reject two common defenses of boilerplate’s enforceability. Finally, Part V details my approach to regulating boilerplate in a world of vertical transactions. Take it or leave it.

---

4. See Rakoff supra note 1, at 1177 (listing characteristics of a “contract of adhesion”).

5. See infra Part V.A (discussing boilerplate as an example of the classic market for lemons).
II. The State of the Debate

To understand why boilerplate’s information costs are important, one must first understand how consumers and boilerplate interact in the marketplace. I therefore begin by defining a core issue on which almost everyone agrees—namely, that the enforceability of boilerplate depends on how well the relevant market is functioning. I then describe how consumers’ information costs can threaten that market function. This will set the stage for Part III’s vertical study of boilerplate’s real-world information costs.

A. The Common Ground: Contracts as Product Features

Anyone who has taken (or taught) Contracts in law school is familiar with the typical reaction when the topic turns to contracts of adhesion. By that point in the semester, students have learned to think of a contract as a negotiated arrangement of parties’ personal preferences—a meeting of the minds, mediated by the give-and-take of offer and counteroffer. Then, suddenly, they encounter boilerplate and its adhesive terms, and they realize that in real life the vast majority of contracts are take-it-or-leave-it propositions. No negotiation takes place. No minds meet. One party gives; one party takes. Madness!

If the market is functioning properly, however, these objections are unavailing. Nonnegotiable terms are the norm in the modern marketplace. Price is a prime example: consumers don’t usually negotiate over the price of the products they buy. Rather, price tends to be a take-it-or-leave-it term, and we trust that competition will punish those sellers that set theirs too high.6 Likewise, if a consumer desires a product feature that one

6. “Seller” is my shorthand for the party that introduces the boilerplate into the transaction. Obviously there are circumstances in which the buyer originates the contract, but this Article focuses on consumers, who will usually be buyers and will usually be contract “takers” rather than contract “givers.” In a similar vein, this Article limits its scope to consumers and the market failure that attends their contractual transactions. But it may well be the case that the market is no more adept at responding to the contractual preferences of sophisticated business entities than those of consumers. Many of the adhesion
seller does not provide (say, a sunroof on a rental car), the answer is not negotiation; the answer is that the consumer takes his or her business elsewhere—namely, to a competing seller that does offer that feature.

Instead of relying on the individual power of negotiation, then, we rely on the collective power of competition to provide consumers with the array of product features they desire, at a price they are willing to pay. Why shouldn’t we approach boilerplate the same way? Let the market punish those sellers whose contracts are too onerous. Consumers will express their contracts that I encountered in my study would have applied equally to businesses or had a separate but equally complex contractual counterpart for such purchasers. And Mitu Gulati and Robert Scott have recently shown that even in the sovereign-debt market, where multibillion-dollar banks and hedge funds transact with nation-states, boilerplate language fails to respond to the parties’ clear preferences—a phenomenon that Gulati and Scott blame on a suspect familiar to this Article’s readers, namely “a business model that relies on herd behavior, fails to provide incentives for innovation and thus rises and falls on volume-based, cookie-cutter transactions.” G. Mitu Gulati & Robert E. Scott, Introduction: The Three and a Half Minute Transaction: Boilerplate and the Limits of Contract Design 8 (Columbia Univ. Law Sch., Working Paper No. 410, 2011), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1945988; see also Florencia Marotta-Wurgler, What’s in a Standard Form Contract? An Empirical Analysis of Software License Agreements, 4 J. EMPIRICAL LEGAL STUD. 677, 680 (2007) [hereinafter Marotta-Wurgler, Empirical Analysis] (finding that “[end-user license agreements] associated with products targeted toward the general public are not significantly more pro-seller than the [end-user license agreements] associated with business-oriented products”).

7. For some reason, cars (with and without sunroofs) tend to be the example of choice when discussing market responses to consumer preferences. See, e.g., Clayton P. Gillette, Rolling Contracts as an Agency Problem, 2004 Wis. L. REV. 679, 688 (using the example of a red sports car with a sunroof and standard transmission); Korobkin, supra note 1, at 1220 (using the example of different colored cars, each with different options available).

8. See, e.g., RICHARD A. POSNER, ECONOMIC ANALYSIS OF LAW 144 (8th ed. 2011) (“[W]hat is important is not whether there is haggling in every transaction but whether competition forces sellers to incorporate in their standard contracts terms that protect the purchasers.”); Robert A. Hillman & Jeffrey J. Rachlinski, Standard-Form Contracting in the Electronic Age, 77 N.Y.U. L. REV. 429, 442 (2002) (“[T]he aggregate decisions of many consumers can pressure businesses into providing an efficient set of contract terms in their standard forms.”); Korobkin, supra note 1, at 1209 (describing “the market discipline established by the ability of buyers to shop among sellers for the most desirable package of product attributes, including contract terms”); Rakoff, supra note 1, at 1251 (“[B]argaining is not essential . . . as long as shopping concerning the particular term takes place.”).
preferences by rejecting those terms in favor of more attractive terms offered by another seller. Unwanted boilerplate will simply disappear from the market, just as ridiculously high prices do.9

In other words, a boilerplate term is merely a product feature—no different from price or a sunroof. The law would not normally regulate the price that a car rental company charges its customers or force it to offer sunroofs in its vehicles. For the same reason, the argument goes, the law should enforce a boilerplate contract without regard to its content. Let the market work it out.

This notion of boilerplate as a product feature, regulated by market forces, has become a staple of both the scholarly literature and the case law. Scholars consistently equate contract terms to noncontractual attributes of a product.10 And perhaps the most well-known case on boilerplate, ProCD, Inc. v. Zeidenberg,11 wholeheartedly embraced this idea in discussing the adhesive “terms of use” that the seller had included with its digital database software:

Terms of use are no less a part of “the product” than are the size of the database and the speed with which the software

---

9. See Korobkin, supra note 1, at 1219 (explaining that the standard economic model assumes that consumers engage in a cost–benefit analysis of products before they buy and that such behavior drives out of the market all products with undesirable attributes or with production costs that are higher than buyers are willing to pay).


compiles listings. Competition among vendors, not judicial revision of a package’s contents, is how consumers are protected in a market economy. ProCD has rivals, which may elect to compete by offering superior software, monthly updates, improved terms of use, lower price, or a better compromise among these elements.\footnote{Id. at 1453 (citation omitted).}

This widespread acceptance of boilerplate as just another product feature approach has occasioned a shift in the scholarly literature. Scholars no longer worry about individual negotiation; instead, they worry about whether ProCD’s assertion is correct: does competition among sellers over the terms of contracts adequately protect consumer interests? We turn to that question next.

**B. The Battleground: Market Function or Market Failure**

We have now seen that contract law and scholarship treat boilerplate as simply another product feature, which consumers are free to accept or reject according to their own preferences. Under this view, a functioning market will regulate the content of adhesive contracts just as it regulates other features, such as price. Enforceability of boilerplate accordingly depends on how well the market responds to consumer preferences regarding that “feature” of the transaction—the boilerplate’s terms.

To evaluate the market’s responsiveness to the boilerplate “feature,” however, one must first understand how the market regulates the features of a product as a general matter. In an ideal world, a consumer identifies and evaluates all the features of a product, rating each against the others in an internal cost–benefit calculus that reflects his or her particular preferences. So when evaluating a car rental, I decide how much each of the many product features matters to me, regardless of whether that feature is contractual in nature (e.g., mileage allotment, insurance coverage) or noncontractual (e.g., the model of the car,
its color, whether it has a sunroof). I then perform the same evaluation on the competing car rentals that the market offers.

When I have evaluated all the rental offerings, I choose the one that most closely approaches my optimal balance of features, so as to maximize the utility I derive from the transaction—or I exit the market entirely, having failed to identify any option that justifies the price that the seller demands. Either choice sends a signal to the marketplace about what features I desire and how much I am willing to pay for them, and that signal is combined with the signals from other consumers to produce an efficient array of market options.

This idealized model, known as “compensatory” decisionmaking, makes many assumptions about consumer behavior. Most important, it assumes that consumers are capable of both acquiring the information they need and then processing that information in a sophisticated cost–benefit analysis, under which the merits of unrelated features like sunroofs and insurance policies are reduced to some common utility metric (price, presumably) by which they can be compared and traded off.

---

13. In a sense, of course, the latter features could also be considered contractual, in that the rental agency promises to provide a car with those features. But the “contractual” label here is meant to capture the more abstract, intangible kinds of promises, which are often found in boilerplate.

14. Shoshana Shiloh et al., Individual Differences in Compensatory Decision-Making Style and Need for Closure as Correlates of Subjective Decision Complexity and Difficulty, 30 Personality & Indiv. Diff. 699, 701 (2001) (“A compensatory strategy entails that the alternative chosen is superior to the other alternatives in the sum of the weighted utilities of all the attributes considered, and leads to maximization of utilities—the main criterion of normative decision making.”). For a more detailed (and quite excellent) explanation of this idealized process in a contracts context, see Korobkin, supra note 1, at 1219–22.

15. See R. Ted Cruz & Jeffrey J. Hinck, Not My Brother’s Keeper: The Inability of an Informed Minority to Correct for Imperfect Information, 47 Hastings L.J. 635, 638 (1996) (explaining that in an ideal world with perfect information, contracts “will contain only efficient terms... because all of the terms will, by definition, be fully understood and properly valued”).

16. See Shiloh et al., supra note 14, at 701 (explaining that “compensatory” decisionmaking means that “the decision maker clarifies objectives, surveys an array of alternatives, searches for relevant information, assimilates information in an unbiased manner, and evaluates alternatives carefully before making a choice”).
against one another.\textsuperscript{17} As discussed below, both assumptions are suspect.

\textit{1. Information Acquisition}

The first flaw in the idealized model of compensatory decisionmaking is that it assumes that consumers can efficiently acquire the necessary information about the options the market offers.

Suppose I want to rent a car with two particular features: I want it to be reliable and I want it to have a sunroof. It is relatively easy to distinguish cars that have sunroofs from those that do not. It might be harder to determine whether a car is reliable. And if the cost of acquiring information about reliability exceeds the value I attach to that feature, I will make my decision without regard to reliability.\textsuperscript{18} It is rational for me to do so, given the cost of information acquisition, but my choice nevertheless sends the wrong signal to the marketplace because the information acquisition cost prevented me from expressing my preference with regard to reliability.

If we view a boilerplate contract as just another product feature, then the information acquisition problem requires us to examine how difficult it is for consumers to learn the terms of the contract and make compensatory judgments accordingly. Contract law has several mechanisms that address this issue, such as the requirement that terms be reasonably certain before a contract is formed,\textsuperscript{19} the distinction between an acceptance and a counteroffer,\textsuperscript{20} and limitations on the modification of contracts after formation.\textsuperscript{21} In various ways, these mechanisms mediate the tension between the need to arrive at a meaningful agreement

\begin{itemize}
\item \textsuperscript{17} Grether et al., \textit{supra} note 10, at 287 (“Consumers could fail to choose the best because of high costs of acquiring information about market choices . . . or because of high costs of processing information about market choices . . . ”).
\item \textsuperscript{18} See \textit{id.} at 287–88 (explaining that “even rational consumers fail to consider all options in the face of high information costs”).
\item \textsuperscript{19} \textit{Restatement (Second) of Contracts} § 33 (1981).
\item \textsuperscript{20} \textit{Id.} § 59.
\item \textsuperscript{21} \textit{Id.} § 89.
\end{itemize}
and the fact that all contracts are incomplete—i.e., that some contractual terms are unknown or become known only after the parties are heavily invested in the transaction.22

Despite these existing doctrinal mechanisms, there is compelling evidence that current contract law does not adequately address consumers’ information acquisition costs, such that consumers often cannot engage in compensatory decisionmaking with regard to the boilerplate terms that accompany many common transactions. Indeed, the significance of those costs has been one of the most hotly contested topics in contract case law and scholarship over the last fifteen years.23

On one side of the debate are those who have no objection to delaying the availability of boilerplate terms until other aspects of the transaction are well underway. The best example of this approach is, again, Judge Easterbrook’s famous ProCD, Inc. v. Zeidenberg,24 which stands for the proposition that contractual terms withheld from a consumer until late in the transaction are nevertheless enforceable, as long as the consumer had a chance to return the product if the terms were not acceptable.25 After all, the argument goes, consumers know that modern transactions often come with boilerplate attached.26 Some courts have followed ProCD, and some scholars have defended it.27

In other quarters, however, ProCD’s belief in a functioning market has not been as well received. The main objection is that even if the late-arriving boilerplate gives consumers the option to

25. Id. at 1452.
26. See, e.g., Hill v. Gateway 2000, Inc., 105 F.3d 1147, 1149–50 (7th Cir. 1997) (Easterbrook, J.) (following ProCD and noting that late-arriving boilerplate contains terms that consumers value as part of the transaction).
reject its terms, by that point they have invested too much time and attention to return the product (i.e., reject the “offer”) and go back to square one. For example, in ProCD’s world, consumers who want to comparison-shop for personal computers must purchase a computer, bring it home, and search it for boilerplate (both within its box and once it is started up)—and then must repeat this process for each and every computer they are considering.28 This imposes arguably insuperable information acquisition costs, and, for this reason, many commentators and several courts have rejected ProCD and argued against the enforceability of late-arriving boilerplate.29

As we will see below, a common response to the problem of information acquisition costs is to require the disclosure of

---

28. Cf. Hill, 105 F.3d at 1150 (relying on ProCD to enforce contractual terms found in the box of a mail-order computer).

29. For courts, see, e.g., Klocek v. Gateway, Inc., 104 F. Supp. 2d 1332, 1339–41 (D. Kan. 2000) (discussing why the court was not persuaded to follow the reasoning used in Hill and ProCD); Novell, Inc. v. Network TradeCtr., Inc., 25 F. Supp. 2d 1218, 1230–31, 1230 n.17 (D. Utah. 1997) (following the “majority” position that late-arriving shrinkwrap licenses are invalid “contracts of adhesion, unconscionable, and/or unacceptable pursuant to the U.C.C.”), vacated on other grounds, 187 F.R.D. 657 (D. Utah 1999); Wachter Mgmt. Co. v. Dexter & Chaney, Inc., 282 Kan. 365, 377–78 (2006) (adhering to “traditional contract principles” and treating shrinkwrap agreements as a proposal to modify the terms of the contract pursuant to UCC 2-209). For commentators, see, e.g., Jean Braucher, Amended Article 2 and the Decision to Trust the Courts: The Case Against Enforcing Delayed Mass-Market Terms, Especially for Software, 2004 Wis. L. Rev. 753, 755 [hereinafter Braucher, Decision to Trust the Courts] (arguing that a buyer does not agree to a seller’s delayed mass-market terms); Niva Elkin-Koren, Copyright Policy and the Limits of Freedom of Contract, 12 Berkeley Tech. L.J. 93, 110–11 (1997) (arguing that agreements received so late in the transaction dissuade consumers from rejecting the terms because of the high transaction costs already endured by the consumer); Korobkin, supra note 1, at 1265 (arguing against enforcement of late-arriving boilerplate because buyers, “[a]fter the purchase, . . . [have] already invested in the particular products, and returning them would . . . require[] expending additional time and effort”). Note that the sunk cost effect might cause an actual consumer to be even less likely to return the computer than the theoretical rational consumer. Shmuel I. Becher, Behavioral Science and Consumer Standard Form Contracts, 68 La. L. Rev. 117, 129 (2007) (“Since the efforts to become familiar with the transaction’s details are sunk, a natural tendency, according to the framework proposed by behavioral law and economics, is to ignore potentially adverse terms that the [standard form contract] may contain, although this tendency is irrational according to traditional law and economics.”).
boilerplate earlier in the transaction. Such an approach can indeed substantially decrease those costs—a necessary and welcome step towards a functioning market. It is not, however, sufficient all on its own. Something else stands in the way, a second and even more problematic shortcoming of the idealized model of compensatory decisionmaking: the cost of processing information about a product. It is to that issue that we now turn.

2. Information Processing

The second flaw in the idealized model of compensatory decisionmaking has to do with consumers’ ability to process information once they acquire it. Even if information about each product and all its features is available, such that information acquisition costs are low, consumers frequently lack the capacity to evaluate that information and express their preferences accordingly.

The problem here is information overload. Again, the car rental example: Rental cars come in many shapes and sizes, with a variety of features (both contractual and not): make, model, color, transmission, mileage allotment, insurance coverage, and more. Compensatory decisionmaking would require me to decide how much each feature matters, rate it against the other features, produce some sort of aggregate score, and then do the same for the many other rental options the market offers.

This is a demanding information processing task. I must make a high number of feature-to-feature comparisons and then account for both the gradations of difference among them and the value I attach to each such gradation. Do I value a sunroof more than an automatic transmission? If so, by how much? And that’s just two features. Contemplate the added complexity that would come with a third—say, color. Would I choose an orange car with a sunroof and manual transmission over a blue car with no sunroof and an automatic transmission?

One can see how quickly the cost of processing such information rises as market options expand. Consider the variety of choices available for everyday purchases like food:

An ordinary supermarket contains 285 varieties of cookies, including 21 chocolate chip options alone; 20 different types of
Goldfish crackers; 13 “sports drinks,” 65 “box drinks,” and 85 flavors of juice; a dozen varieties of Pringles potato chips; 80 pain relievers; 40 lipstick shades; 16 varieties of instant mashed potatoes, 75 different instant gravies, and 120 different pasta sauces; 175 different salad dressings; and a whopping 275 types of cereal.30

And keep in mind that not only is there often a huge number of competing products, but the products differ from one another along multiple dimensions such that each has multiple features to be weighed and compared. Do I want low-fat, Chewy Chips Ahoy or Double-Stuf Oreos?31

Not surprisingly, then, empirical studies show that, for all but the most basic transactions, the cost of processing information makes compensatory decisionmaking a mere pipe dream.32 Consumers abandon a purely compensatory strategy when faced with as few as six options,33 process no more than five

31. The correct answer is the Oreos.
32. See Jacob Jacoby, Perspectives on Information Overload, 10 J. CONSUMER RES. 432, 435 (1984)

[The key finding to emerge [from studies of consumer decisionmaking] is that consumers stop far short of overloading themselves. They tend to examine only small proportions of the brand and attribute information that is available. . . . Few findings regarding consumer behavior have proven to be as consistent across so many different procedures, contexts, products, investigators, and so on.]

See also Naresh K. Malhotra, Information Load and Consumer Decision Making, 8 J. CONSUMER RES. 419, 427 (1982) (summarizing studies that show—with remarkable consistency—a decline in optimal decisionmaking after the number of data inputs exceeds ten). These findings seem to be offshoots of the theory of the “magic number seven,” the conclusion from cognitive science that people generally cannot retain and process more than approximately seven bits of information in their short-term memories. See George A. Miller, The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information, 63 PSYCHOL. REV. 81, 95 (1956) (synthesizing research on information processing in different contexts and concluding that “the span of absolute judgment and the span of immediate memory impose severe limitations on the amount of information that we are able to receive, process, and remember”).

features of any particular option, and make decisions essentially at random when faced with just four options that have four features each. And the low levels of complexity at issue in these studies do not even begin to represent the amount of information that consumers frequently encounter in the real world. Moreover, not only is it clear that the typical consumer can only process a limited (and surprisingly small) amount of product information, but there is also some empirical evidence that, once that limit is reached, providing more information and options is worse than useless: it may actually reduce the consumer’s ability to make the right choice.

34. Id. at 162; see also Malhotra, supra note 32, at 423 (“[T]he probability of correct choice decreases significantly as the number of attributes on which information is provided increases from five to 15, 20, or 25.”).

35. Grether et al., supra note 10, at 297. Grether and his co-authors curiously conclude that these findings support the proposition that “consumers do not experience serious problems as a result of the amount of information that markets and the state now generate,” id. at 294—a conclusion for which other commentators rightly take them to task. See Melvin Aron Eisenberg, Text Anxiety, 59 S. Cal. L. Rev. 305, 308–09 (1986); Roberta Romano, A Comment on Information Overload, Cognitive Illusions, and Their Implications for Public Policy, 59 S. Cal. L. Rev. 313, 317 (1986).

36. The largest of the studies involve options that number in the teens or twenties, with a similar number of features per option. E.g., Malhotra, supra note 32, at 420 (twenty-five options with up to twenty-five features). As Jacob Jacoby points out in commenting on a sixteen-option, sixteen-feature study, the real world presents consumers with considerably more information than that. See Jacoby, supra note 32, at 434 (noting that there are approximately 150 cereal brands on the market and their packaging can contain over 100 items of information).

37. See Kristin Diehl, When Two Rights Make a Wrong: Searching Too Much in Ordered Environments, 42 J. MKTG. RES. 313, 314 (2005) (listing reasons why more information can lead to worse decisions); Sheena S. Iyengar & Emir Kamenica, Choice Proliferation, Simplicity Seeking, and Asset Allocation, 94 J. PUb. ECON. 530, 533–34 (2010) (finding small but significant negative effect on 401(k) investment decisions as employees were offered more fund options); Sheena Sethi-Iyengar et al., How Much Choice Is Too Much? Contributions to 401(k) Retirement Plans, in Pension Design and Structure: New Lessons from Behavioral Finance 83, 88–91 (Olivia S. Mitchell & Stephen P. Utkus eds., 2004) (finding a 1.5% to 2% decrease in employee 401(k) participation for every ten funds added to plan); Paul Slovic, Toward Understanding and Improving Decisions, in 2 Human Performance and Productivity: Information Processing and Decision Making 157, 168 (William C. Howell & Edwin A. Fleishman eds., 1982) (reporting study in which horse-race handicappers’ predictions failed to improve and actually became
So how do consumers respond to these insuperable information costs? They simplify, usually by identifying just a few salient, easily processed features and focusing on those, or at least using them as a screening device to reduce the number of products under consideration to a more manageable level at which compensatory comparison is possible. They try to make a satisfactory choice by sacrificing inquiry into certain features in favor of pursuing inquiry into few, salient others—an approach known as “satisficing.”

Replacing a purely compensatory approach with a satisficing strategy makes sense for consumers. Indeed, they may have no other choice, given their inherent cognitive limitations and the presence of so much information in the marketplace. But satisficing has important implications for whether we can rely on the market to adequately account for consumer preferences. After all, by definition satisficing results in something other than an optimal expression of consumer preferences. In other words, if a particular product feature does not make the cut during the satisficing process—i.e., if it is not one of the features that are salient to consumers when they screen available options—then there is less reason to believe that the market is sending the right signal to sellers about the desirability of that feature.

more inconsistent as available information increased); see also Jeffrey Davis, Protecting Consumers from Overdisclosure and Gobbledygook: An Empirical Look at the Simplification of Consumer-Credit Contracts, 63 VA. L. REV. 841, 847–49 (1977) (citing studies).

38. Grether et al., supra note 10, at 287–88. The term satisficing was coined by Herbert Simon, one of the godfathers of behavioral economics, in Herbert A. Simon, Rational Choice and the Structure of the Environment, 63 PSYCHOL. REV. 129, 136 (1956).

39. Over the years, different commentators have used different definitions of satisficing. See, e.g., Eisenberg, supra note 1, at 214–15 (defining satisficing as searching until one finds an option that meets certain predetermined criteria at which point the choice is made and the search stops); Grether et al., supra note 10, at 287–88 (defining satisficing as used in main text above); see also Korobkin, supra note 1, at 1223–25 (describing various simplifying strategies that consumers use in response to information processing costs). For present purposes, these distinctions make no difference because they all stand for the proposition that, in the real world, consumers depart from compensatory decisionmaking.

40. See Botti & Iyengar, supra note 30, at 27 (“[T]he objective of reducing the cognitive costs involved in making the choice can produce suboptimal
What do information processing costs mean for boilerplate? According to contract law, very little. Courts and legislatures tend to assume that the problem, if any, is information acquisition, and thus that early disclosure of terms is all that consumers need. Other than the anemic doctrine of unconscionability (to which we will return later), current law is largely silent on the issue.

The law may be silent, but a chorus of commentators has argued that information processing costs can routinely prohibit the reading of boilerplate even when it is provided early in the transaction. And empirical studies (mostly of the software industry) have begun to confirm this suspicion that information acquisition is only half of the problem. One study of more than six hundred software contracts found a lack of correlation between competitive market conditions and the content of boilerplate

decisions and subsequent dissatisfying outcomes.


42. See infra Part V.A.

43. The incapacity doctrine is one way in which contract law already accepts the proposition that cognitive ability should play a role in enforceability, at least at the extreme. See Eisenberg, supra note 1, at 212–13 (“[T]he doctrine of capacity rests on the ‘assumption that incompetents, properly defined, require protection from their own actions,’ so that the premise of the bargain principle, that a contracting party will act with full cognition to rationally maximize his subjective expected utility, is not fulfilled.” (footnote omitted)). For a comprehensive review of contract doctrines that address information processing costs for boilerplate, see Hillman & Rachlinksi, supra note 8, at 454–60.

44. E.g., Eisenberg, supra note 1, at 247; Gillette, supra note 7, at 682; Hillman, supra note 1, at 850; Korobkin, supra note 1, at 1217; Rakoff, supra note 1, at 1226.
terms, suggesting that boilerplate does not respond to competitive pressures. And in a study of online software purchases, Florencia Marotta-Wurgler tracked one month of Internet click-stream data for 47,399 website visitors. Her finding? Even if the seller forces consumers to click “I agree” and provides a direct link to the boilerplate, only one in every two hundred consumers reads it—and that’s under a very liberal definition of “read” that includes any consumer that spends at least one second on the page where the adhesive terms are available. (Relaxing these limitations led to an even lower reading rate.)


46. A related study found that boilerplate that was available early in the transaction was no more pro-consumer than boilerplate that arrived later, suggesting that lowering the costs of information acquisition does not necessarily lead to different contract terms. Florencia Marotta-Wurgler, Are “Pay Now, Terms Later” Contracts Worse for Buyers? Evidence from Software License Agreements, 38 J. LEGAL STUD. 309, 333 (2009). Of course, this finding might be read to show that all boilerplate is responsive to consumer preferences, rather than that no boilerplate is. And the study’s absolute measure of “pro-consumer” versus “pro-seller” should be taken with a grain of salt because it uses the methodology from an earlier study in which the maximum pro-consumer score for a contract was six and the maximum pro-seller score was seventeen. Marotta-Wurgler, Empirical Analysis, supra note 6. Nevertheless, for comparative purposes—determining whether one contract is more pro-consumer than another—the methodology is sufficiently reliable.


48. See id. at 108 (finding that only 0.52% of such consumers spend more than one second on the contract web page); see also id. at 110 (concluding that “the primary cost [to consumers] lies not in locating and accessing EULAs, but rather in reading and assessing contract terms”).

49. See id. at 108 (finding a rate of only 0.13% when the boilerplate is not directly called to the consumer’s attention and 0.00% when it takes more than one click of the mouse to locate it).
This, in sum, is the boilerplate problem. If a contract term is merely a product feature, then more terms mean more features. More features mean more complexity. More complexity increases the use of satisficing strategies that eliminate boilerplate from consumers’ decisionmaking calculus. And if consumers routinely eliminate boilerplate from their decisionmaking, the justification for its enforcement—the protection of the marketplace through the collective power of competition—goes away.

III. A Vertical Study of Boilerplate

We have now seen that the problem of information costs casts doubt on whether the market registers consumer preferences regarding boilerplate. We have also seen some emerging empirics that suggest that this concern is more than theoretical. In this Part, I offer a different lens through which to view the boilerplate problem: a “vertical” study that contextualizes both information acquisition and information processing within the realities of consumer decisionmaking.

A. Verticality’s Advantages

The existing empirical studies that examine information costs in the world of boilerplate share one important limitation: they examine one stratum of individual contracts, all of a kind, and assume that this single contract defines the entire transaction.50 Even the click-stream study, which tracked actual consumer interactions with boilerplate, limited itself to interactions with a single kind of contract in a simple one-off transaction (the purchase of a stand-alone item of software).51

50. See studies cited supra notes 6, 45–49. This is not meant as a criticism; those studies were designed to measure something other than the overall consumer experience. For example, one study was primarily interested in identifying the balance of pro-seller versus pro-consumer terms. Marotta-Wurgler, Empirical Analysis, supra note 6, at 679.

51. See Marotta-Wurgler, Does Contract Disclosure Matter?, supra note 47 (examining software licensing agreements in the context of purchasing a single item of software). The issue is not that the click-stream study focuses only on software licensing; the issue is that by focusing on any one kind of contract and
That sort of “horizontal” analysis is revealing, but its focus on a single kind of contract, and nothing else, means that it sacrifices depth for breadth. In contrast, a “vertical” analysis would look at fewer contracts of any one kind, but would situate the consumer’s encounter with those contracts in the real world—an encounter in which the boilerplate is only part of a transaction, not its entirety. Consider software licensing. Sometimes consumers encounter software boilerplate in the sort of one-off horizontal context that earlier studies have focused on. But other times they encounter software boilerplate as part of an overall shopping experience, an experience that also presents noncontractual features to be evaluated, and indeed often involves multiple vendors and multiple contracts. Measuring the consumer experience vertically, from top to bottom, therefore provides a uniquely instructive view of information costs and the tradeoffs that consumers make in deciding which features to evaluate and when.

In other words, a horizontal study tells us something about the practices of a given industry, but it tells us less about the consumer experience in dealing with that industry and the true information costs that the industry imposes on consumers. A vertical study can overcome that deficiency. After all, if boilerplate is just one feature of a product, then a consumer’s ability to acquire and process information about boilerplate is not a function of the complexity of any one contract. It is a function of the complexity of the entire transaction, soup to nuts. Therefore, if we want to know when in the transaction the consumer

52. Thus the saying in the software world, “The license is the product.” See, e.g., Robert W. Gomulkiewicz, The License Is the Product: Comments on the Promise of Article 2B for Software and Information Licensing, 13 BERKELEY TECH. L.J. 891, 896 (1998) (“For most software products, the license is the product; the computer program provides functionality to the user, but the license delivers the use rights.”).

53. See Leff, supra note 10, at 146–47 (noting that “when one stands far enough back from the whole deal, from the whole process of goods buying, what one sees is a unitary, purchased bundle” and that boilerplate is just one part of “the whole ‘set’”).

assuming that it represents the totality of the transaction, one can draw only limited conclusions about consumer ability to acquire and process information in a real-world context.
encounters boilerplate (a factor of vital importance to information acquisition costs) and how much total transactional complexity the consumer encounters (a factor of vital importance to information processing costs), a vertical study is the way to go.54

What would such a study look like? As explained in more detail below, my approach was to purchase personal computers and measure both the volume of boilerplate (the “how much” factor) that came with them, and the point in the transactions at which I encountered each bit of boilerplate (the “when” factor). I chose this approach for four reasons.

First, a computer purchase is exactly the kind of transaction that shows the advantages of a vertical analysis. From the consumer’s point of view, buying a computer is a single transaction, with a one-time, lump-sum price term. Yet the product to be purchased presents the consumer with many different features that emerge at different points in time. Examining such a transaction therefore allows one to track both the “when” and the “how much” aspects of information cost.

Second, most of the scholarship and case law in this area focuses on the computer industry. Indeed, some of the foundational cases on consumers and boilerplate specifically involve computer purchases.55 Therefore, regardless of one’s views on the study’s broader application, it is germane to an important debate in the field—and to an industry that generates around $50 billion annually from such purchases.56

Third, a computer purchase involves a high degree of contractual complexity. It thus demonstrates the critical role that information costs play in boilerplate’s enforceability. One can

54. Of course, a vertical case study of this kind constitutes just one example of a consumer’s encounter with boilerplate. In a sense, then, this approach is the mirror opposite of the horizontal approach in that it sacrifices breadth for depth.


56. The recent economic downturn has affected personal computer sales, dropping domestic earnings from $51.3 billion in 2007 to a low of $45.8 billion in 2009. DATAMONITOR, P Cs IN THE UNITED STATES 10 (2011). But the numbers have begun to bounce back, with $47.2 billion earned in 2010—and volume rose steadily even during the recession, from 62.4 million units in 2006 to 75.7 million in 2010. Id. at 10–11.
imagine transactions with comparable vertical complexity (for example, a vacation package that includes flights, hotels, tours, and of course a rental car), as well as transactions that are much simpler (for example, buying a loaf of bread). But as we will see in Part V,57 part of my thesis is that the enforceability of boilerplate should vary with the overall complexity of the transaction, and starting with an example that involves high complexity helps drive that point home.

Finally, contracts and complexity are inextricably intertwined in the market for information goods. Sellers have significant incentives to offer merchandise that, like a computer system, comprises many different products that could have been sold separately—and to then use boilerplate to restrict consumers’ resale of the individual products within. This strategy, known as bundling, allows sellers to sell a set of goods to consumers at a unitary price even though particular consumers attach disparate values to the set’s individual components.58 For example, the Microsoft Office Home and Student suite, available for $150, contains four separate programs (Word, Excel, PowerPoint, and OneNote).59 Some consumers may value Word, Excel, and PowerPoint at $40 each and OneNote at $30, whereas others value Word, Excel, and PowerPoint at $30 each and OneNote at $60. Selling the suite at $150 satisfies both groups and generates more surplus than could be gained from selling the same consumers each program as a

57. In particular, see the discussion of procedural unconscionability in Part V.A.

58. Nobel laureate George Stigler first introduced this concept in George J. Stigler, United States v. Loew’s Inc.: A Note on Block-Booking, 1963 SUP. CT. REV. 152, 152 (1963) (discussing the strategy as employed through the concept of “block-booking” of movies). The “bundling” label came later. See, e.g., Yannis Bakos & Erik Brynjolfsson, Bundling Information Goods: Pricing, Profits, and Efficiency, 45 MGMT. SCI. 1613, 1613 (1999) (studying the strategy of “bundling” a large number of information goods).

standalone product at a unitary price. But for bundling to succeed, the seller must be able to prevent arbitrage; the strategy will not work if, for example, a consumer who values OneNote at $30 can unbundle it from the suite and resell it as a standalone program to a consumer who values it at $60. Enter boilerplate, which restricts consumers from doing that exact thing.

For these reasons, then, a vertical study of the information costs inherent in the purchase of an entire computer system can teach us a lot about boilerplate enforceability in the real world. To such a study we now turn.

B. Study Design

The subject matter of the study was the purchase of desktop computers from the four top sellers of Windows-based computer systems (Acer, Dell, HP, and Toshiba), which together make up two-thirds of the domestic computer market. Each order comprised just a single basic desktop unit and the software and accessories that were included in the base price; any additional hardware, software, or other option that required additional fees was declined. I paid in full at the time the system was ordered, had it shipped, opened the box, set it up, and started up various

60. Microsoft sells the components individually as well as in a suite—a practice known as “mixed bundling.” See William James Adams & Janet L. Yellen, Commodity Bundling and the Burden of Monopoly, 90 Q.J. ECON. 475, 475 (1976). The standalone versions, however, are sold at a significant markup. See Buy Microsoft Office 2010, supra note 59 (listing OneNote at $80 and the other components at $140).

61. See Bakos & Brynjolfsson, supra note 58, at 1614 (noting that one of the conditions for analyzing bundling as a device for price discrimination is “no reselling”).

62. Id.

63. See Gartner, Inc., Gartner Says Worldwide PC Shipments in Fourth Quarter of 2011 Declined 1.4 Percent; Year-End Shipments Increased 0.5 Percent (Jan. 11, 2012), http://www.gartner.com/it/page.jsp?id=1893523 (last visited Feb. 4, 2013) (showing HP market share at 23.1%, Dell at 22.4%, Toshiba at 10.7%, and Acer at 9.8%, for a total of 66.0%) (on file with the Washington and Lee Law Review). Due to funding limitations, the other major vendor (Apple, at 11.6%) was not included.
programs that came with it. All along the way, I kept track of the boilerplate I encountered—every contract term that was presented to me in a take-it-or-leave-it fashion.64

The most difficult part of the study design was deciding which contracts should count. In the Dell purchase, for example, I collected data on 186 different potential contracts. But determining which of those 186 should be included in my tally required two judgment calls.

1. Assumptions About Enforceability

The first and most important judgment call was to decide which of the various terms I encountered would be viewed as enforceable under current law. My goal was to be very conservative—to include only those contracts that a court would enforce without any real controversy. I achieved this goal through the use of two narrowly conceived enforcement criteria: the acceptance criterion and the availability criterion.

The acceptance criterion meant that I counted only those contracts to which I had clearly and affirmatively manifested assent.65 Usually this assent took the form of the classic “I agree” or “I accept” mouse click—the classic “click-wrap” contract.66 So, for example, at the end of Dell’s online order process I encountered the screen depicted in Figure 1, which asked me to affirmatively manifest agreement to Dell’s Terms and Conditions of Sale. Any court would find that selecting the “I AGREE” option and then submitting the order, as I did, was a manifestation of

64. This meant that contractual terms for which I had an array of options (for example, shipping) were not included in the tally.

65. Some courts have upheld adhesion contracts without explicit manifestations of assent. See, e.g., Hill v. Gateway 2000, Inc., 105 F.3d 1147, 1150 (7th Cir. 1997). But I applied an explicit acceptance criterion because other cases suggest that terms that arrive after purchase are binding only if acceptance is unequivocal. See, e.g., Klocek v. Gateway, Inc., 104 F. Supp. 2d 1332, 1341 (D. Kan. 2000).

66. The origin of the term click-wrap is obscure, but it at least dates back to 1996. See Koh Su Haw, E-Commerce: Technology Can Bypass the Legal Pitfalls, BUS. TIMES (Singapore), Oct. 14, 1996, at 16 (defining “click-wrap”).
assent to be bound.\textsuperscript{67} Those Terms and Conditions of Sale therefore counted in my tally.

\textbf{Figure 1}

In contrast, the bottom of the screen in Figure 1 contained some classic, barely perceptible “fine print” (e.g., “Offers subject to change.”) in gray font on a black background. These terms were not included in my analysis. Although some courts would undoubtedly enforce them, the lack of an affirmative manifestation of assent meant that enforceability was too uncertain to merit inclusion under my conservative assumptions.\textsuperscript{68} Similarly, some courts would enforce the various policies that one could access by clicking on the hyperlinks above.

\textsuperscript{67} See Nathan J. Davis, \textit{Presumed Assent: The Judicial Acceptance of Click-Wrap}, 22 \textit{Berkeley Tech. L.J.} 577, 583 (2007) (“[C]ourts have almost uniformly found assent when the user clicks while having notice of the terms.”).

\textsuperscript{68} E.g., Specht v. Netscape Comm’ns Corp., 306 F.3d 17, 19–30 (2d Cir. 2002) (finding no online contract formation when assent was ambiguous and terms were relegated to the bottom of the web page).
the fine print (e.g., “© 2010 Dell” and “Limited Warranty”)—but not all courts would do so, so I did not count them either. The same goes for the hyperlinks indicated by the arrow in Figure 2; those links were prominently displayed next to a contract to which I did clearly manifest assent (the Terms and Conditions of Sale), but it was not clear that acceptance of the latter constituted acceptance of the additional contractual terms that one would reach by following those links, so I did not include them in my tally.

My second criterion, availability, meant that the terms to which I manifested assent had to be easily accessible if they were to count in the overall tally. In the online context, it is common for contract terms to be located on a separate web page from the assent mechanism, rather than being forced upon the user as part of that mechanism. Figure 1’s reference to Dell’s Terms and Conditions of Sale provides an example; those terms were available with a single click of a mouse on the prominent hyperlink next to “I AGREE.” To remain conservatively consistent with the relevant case law, I counted such terms only if one could follow the given hyperlink (or nonhyperlinked web address) and find them no more than two web pages away.

69. E.g., Burcham v. Expedia, Inc., No. 4:07CV1963, 2009 WL 586513, at *4 (E.D. Mo. Mar. 6, 2009) (finding defendant bound by browse-wrap agreement even absent affirmative assent because “[a] link to the full text of the user agreement is found at the bottom of the very web page that [defendant used]” and “[t]he user agreement specifically states that users consent to be bound . . . by . . . using the website”); Pollstar v. Gigmania, Ltd., 170 F. Supp. 2d 974, 982 (E.D. Cal. 2000) (refusing to dismiss breach of contract claim involving “browse wrap license” and noting that although “the user is not immediately confronted with the notice of the license agreement, this does not dispose of [plaintiff’s] breach of contract claim”).

70. E.g., Specht, 306 F.3d at 19–30.

71. See Cheryl B. Preston & Eli W. McCann, Unwrapping Shrinkwraps, Clickwraps, and Browsewraps: How the Law Went Wrong from Horse Traders to the Law of the Horse, 26 BYU J. PUB. L. 1, 20 (2011) (surveying eight popular service providers’ methods for acquiring consent to terms of service and finding that all of them provided links to the terms rather than displaying them directly).

addition, it had to be obvious which boilerplate applied; if a link led to a menu of undifferentiated contracts, and the consumer could not easily identify the correct one, then I would not count any of them.

Figure 2

These two criteria also guided the inclusion of terms that were incorporated by reference in qualifying boilerplate (a common occurrence in the online world). First, acceptance: the actual text of the originating contract had to include a reference to the incorporated terms, such that acceptance of the former implied acceptance of the latter. Second, availability: the incorporated terms had to be easily accessible to the reader of the originating contract. For example, the text of Dell’s Terms and Conditions of Sale contained a hyperlink to a Return Policy (in such a manner that it was clear that acceptance of the one constituted acceptance of the other), and the hyperlink led directly to the Return Policy terms. In contrast, a later “Service Contracts” hyperlink within the Terms and Conditions of Sale failed to satisfy the availability criterion; it took two mouse clicks
just to get to a menu of potentially applicable contracts, and it was not entirely clear which of them applied. Therefore, the Return Policy was included in the overall volume measurement but the Service Contracts were not.

2. Assumptions About the Scope of the Transaction

The second judgment call inherent in my analysis involved which software boilerplate to include. As we all know, when you start up a computer for the first time, you find all sorts of programs that you may not have known were part of your purchase—games, trial versions of software, system accessories, and so forth. Opening such programs often leads to click-wrap boilerplate. Should such contracts be counted in the study?

As before, my approach here was conservative; I only included contracts associated with programs and features that the seller had represented as part of the deal when I ordered the computer. A reasonable purchaser might expect more programs than that as a practical matter because extras like media players and DVD burners come standard on most computers. But a conservative approach to contract law would tell us that the purchaser only has a right to expect those programs that were explicitly promised.

Again, I did not pay extra for any of this software; all of it came with the computer. And of course the same enforceability criteria applied here; none of the software contracts counted unless I clearly and affirmatively manifested assent and could easily access their terms.

In the aggregate, what these conservative assumptions mean is that the study almost certainly underestimates information costs that boilerplate truly imposed in the transaction, perhaps by a wide margin.\footnote{To take just one example, my conservative criteria ended up excluding a lot of arbitration procedures, despite the fact that a court would almost certainly find that the presence of arbitration clauses binds the consumer to use those procedures. In most cases, however, those procedures were excluded because they failed to satisfy the availability criterion: the arbitration clauses usually linked to the home page of an arbitration organization (e.g., the American Arbitration Association or JAMS), rather than to the particular}
count includes only those contracts that very clearly alerted the consumer to the need to search for terms because it was very clear that proceeding with the transaction meant agreeing to be bound. And more important, the availability criterion means that only those contracts that could be easily located would qualify—even though in reality courts might enforce contracts that are much harder to find. In several important respects, then, the study makes information costs look less costly than they really are.

C. Study Results

Even under the conservative assumptions outlined above, the four purchases produced a weighted average of twenty-five binding contracts totaling 74,897 words. In other words, the average computer purchase binds the consumer to twenty-five contracts, comprising 74,897 words of boilerplate. To put that word count in perspective, it’s just a tad less than the number of words in the first *Harry Potter* book. Of course, *Harry Potter* is a

arbitration procedures to which the consumer would be bound—and it took several clicks of the mouse and some guesswork to find those procedures on the organization’s website. Nevertheless, a court would almost certainly say that the consumer had a contractual obligation to arbitrate and to use those procedures. (For instance, the arbitration clause that was upheld in *Cavalier Manufacturing, Inc. v. Clarke*, 862 So. 2d 634 (Ala. 2003), merely stated that disputes would be arbitrated under American Arbitration Association (AAA) procedures, with no hyperlink to the AAA website, let alone to the procedures themselves. Brief of Appellant Cavalier Manufacturing, Inc. at 5, Cavalier Mfg. v. Clarke, 862 So. 2d 634 (Ala. 2003)).

74. The exact figures were 24.75 contracts and 74,897.19 words. A weighted average was used so that data from sellers with a greater market share would receive proportionately more emphasis in the calculation, under the theory that consumers are more likely to buy products from them than from their competitors. The weights used correspond to the market share figures from *supra* note 63. (The non-weighted average would not have been much different: twenty-three contracts comprising 71,828 words.) See the Appendices for a full breakdown of the figures discussed in this section.

page-turner, whereas boilerplate contracts are anything but. So perhaps a better analogy is tax forms: you could read every word of the instruction booklet for IRS Form 1040a, cover to cover—all eighty-eight pages—and still be more than a thousand words short of the boilerplate total from this single computer purchase.76 Or, for the truly masochistic among you, try reading this Article, and then do it again, and then once more. And don’t skip the footnotes this time.


Table 1: Summary of Results

<table>
<thead>
<tr>
<th>Seller</th>
<th>No. of Contracts</th>
<th>... overall</th>
<th>... at purchase</th>
<th>... at computer startup</th>
<th>... at program startup</th>
<th>... per $</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acer</td>
<td>12</td>
<td>33,128</td>
<td>9,135</td>
<td>23,993</td>
<td>0</td>
<td>47.1</td>
</tr>
<tr>
<td>Dell</td>
<td>29</td>
<td>78,203</td>
<td>9,765</td>
<td>24,165</td>
<td>44,273</td>
<td>84.7</td>
</tr>
<tr>
<td>HP</td>
<td>25</td>
<td>79,340</td>
<td>0</td>
<td>24,328</td>
<td>55,012</td>
<td>103.4</td>
</tr>
<tr>
<td>Toshiba</td>
<td>27</td>
<td>96,641</td>
<td>18,678</td>
<td>34,744</td>
<td>43,219</td>
<td>131.5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>93</strong></td>
<td><strong>287,312</strong></td>
<td><strong>37,578</strong></td>
<td><strong>107,230</strong></td>
<td><strong>142,504</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Total Unique</strong></td>
<td><strong>56</strong></td>
<td><strong>161,767</strong></td>
<td><strong>39,065</strong></td>
<td><strong>38,225</strong></td>
<td><strong>84,477</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Raw Average</strong></td>
<td><strong>23.25</strong></td>
<td><strong>71,828</strong></td>
<td><strong>9,394</strong></td>
<td><strong>26,807</strong></td>
<td><strong>35,626</strong></td>
<td><strong>91.7</strong></td>
</tr>
<tr>
<td><strong>Weighted Average</strong></td>
<td><strong>24.75</strong></td>
<td><strong>74,897</strong></td>
<td><strong>7,698</strong></td>
<td><strong>25,912</strong></td>
<td><strong>41,286</strong></td>
<td><strong>93.2</strong></td>
</tr>
</tbody>
</table>

1. Information Acquisition Costs

The number of words is informative, and we will return to the question of total volume when we consider information processing costs (the “how much” question). First, however, what does the study tell us about the cost of information acquisition? Even with the conservative assumptions discussed above, one acquisition issue comes through loud and clear—an issue one can appreciate only through a vertical study: the “when” question. At what stage of the transaction was each contract encountered?

Of the 74,897 total words, only 7,699 (10.3%) were presented to me by the time I had to decide whether to order (and pay for) the computer.77 I had to wait until the computer arrived before the rest were made available. Of the remaining 67,198 words, 25,912 (34.6%) were presented when the computer arrived and

77. Note that if I had bought the computer in a store instead of online, many of the website’s boilerplate terms would have arrived as paperwork in the box, or would have been presented to me on startup. Whether those presentations would have satisfied my enforceability criteria is an issue I did not explore.
was first started up, and the other 41,287 (55.1%) when individual programs were opened.78

In other words, the verticality of the study shows how very pertinent the issue of late-arriving terms is. At the time that consumers tender payment, they will have had no opportunity to express their preferences regarding nine out of every ten words to which they will become contractually bound. Of course, they will eventually have the opportunity to explicitly say no to these late-arriving terms, and thus to communicate their preferences to the market.79 But that opportunity arises only after a considerable investment in acquiring the information that would bear on formation of those preferences—the time spent navigating the website, deciding which computer to buy, placing the order, waiting for the shipment, starting up the computer, and opening the various programs to find the applicable contractual language. That investment represents a considerable information acquisition cost. In essence, by the time I have an opportunity to express my preferences to the market, I am no longer at the market. I am already home.

Moreover, the true acquisition cost imposed by late-arriving terms would actually be a multiple of the cost I experienced in my study because, for a market to function most effectively, consumer decisionmaking should be not only compensatory but

78. I might have been able to find some of the late-arriving boilerplate by searching the websites of the software providers before deciding to purchase. But such an approach would have carried acquisition costs of its own, and it is far from certain that I could have located the correct terms. See Jean Braucher, Delayed Disclosure in Consumer E-Commerce as an Unfair and Deceptive Practice, 46 WAYNE L. REV. 1805, 1860–61 (2000) [hereinafter Braucher, Delayed Disclosure] (finding that 87.5% of software websites did not make boilerplate available to consumers pre-purchase); James F. Rodriguez, Software End User Licensing Agreements: A Survey of Industry Practices in the Summer of 2003, at 2 (unpublished and undated manuscript) (finding that only twelve of forty-three major software companies provided license terms on their websites, none provided an easily identifiable pre-purchase link to the terms, and only two offered a website search capability enabling users to find the terms) (on file with the Washington and Lee Law Review).

79. This opportunity to explicitly manifest or refuse express assent is what makes the boilerplate term enforceable, even under the most consumer-friendly cases, such as Klocek v. Gateway, Inc., 104 F. Supp. 2d 1332, 1339–41 (D. Kan. 2000) (noting that late-arriving terms are enforceable if consumers “expressly agree[] to them”).
comparative. Consumers do not decide whether they will purchase a given product in the abstract; they decide whether to purchase it in light of the options available from other sellers. This means that for a consumer to truly express his or her market preferences regarding late-arriving boilerplate, he or she would have to order multiple computers, start them all up, open the programs on each, and then examine the boilerplate within. To make such a comparison between the four computers in this study would have meant reading fifty-six unique contracts totaling 161,767 words.  

Finally, having made that comparison and decided which terms to reject, the consumer would have to register that rejection with the marketplace—e.g., by returning the rejected feature and receiving a corresponding refund. Even if we generously assume that the seller’s return policy allows for this possibility, it is far from clear how it would work in practice. How could one return, say, the antivirus software but retain the operating system (and what would the refund be)? If the return really turns out to be an all-or-nothing proposition, then it sends a weak signal to the marketplace because it would not be clear which features prompted the rejection. And even if these obstacles could be overcome, the return process adds more expense (including both the hassle and the possibility of restocking fees), which means more information acquisition costs.

80 When all the boilerplate in the four transactions is added together, there were ninety-three contracts totaling 287,312 words. Appendix A. But there was some overlap among these contracts (for example, each seller offered the same Microsoft Windows license), so the number of unique contracts was lower. This demonstrates the intuitive notion that comparison among products is easier if each seller offers the exact same features—but that is hardly an argument for market function; it simply substitutes an antitrust problem for an information cost problem. Note also that more of the unique contracts arrived early in the transaction, such that a consumer who bought all four computers for comparison purposes would have encountered nearly one out of every four unique terms (24.1%) before submitting payment, rather than one out of every ten (10.3%).
2. Information Processing Costs

The analysis so far has shown that the costs of acquiring information about boilerplate are a significant impediment to compensatory decisionmaking. As we have already seen, however, the cost of processing information can constitute a separate and equally problematic obstacle to market function. So what does the case study have to teach us about the information processing costs associated with boilerplate?

The issue here is one of sheer volume of terms—the “how much” question. A few empiricists have taken note, mostly in passing, of the volume of boilerplate terms that sellers present to consumers. For example, one of Marotta-Wurgler’s horizontal studies found an average length of 1,500 words for the typical end-user license agreement in the software industry.81 And a horizontal study of the boilerplate that accompanied the top fifty programs at Download.com for 2006 found an average of 2,752 words and a nearly-college-level readability.82

Here again, we see the advantages of a study that looks vertically at all the boilerplate in a single transaction. As noted above, even under my conservative assumptions, the average computer purchase results in 74,897 words of binding contracts. This finding casts the information overload problem in sharp relief. Even if all that boilerplate were presented up front—i.e., even if the acquisition problem were solved—consumers could not engage in compensatory decisionmaking unless they first paid the processing cost of reading and understanding the terms well enough to form preferences and compare them to other sellers’ contractual offerings.

Whether consumers would pay that cost depends on how high it is. Recall that consumers who confront complex products (like computers) minimize those costs by satisficing—ignoring certain features in favor of others. The higher the cost of processing boilerplate, then, the more likely consumers are to

81. Marotta-Wurgler, Empirical Analysis, supra note 6, at 694.
ignore it and turn their attention instead to other, more salient product features.\footnote{Supra note 38 and accompanying text.}

The question, then, is how to measure the cost of processing 74,897 words of boilerplate. One potential answer is to measure the time one would need to read all those terms. A few studies have examined how long it takes the average person to read legal text. The most relevant, conducted by Michael Masson and Mary Anne Waldron, presented subjects with short excerpts from four contracts (a mortgage, a property sale, a bank loan, and a lease renewal) and measured how long it took to read and understand them.\footnote{See generally Michael E.J. Masson & Mary Anne Waldron, Comprehension of Legal Contracts by Non-Experts: Effectiveness of Plain Language Redrafting, 8 APPLIED COG. PSYCHOL. 67 (1994).} That study’s application here is somewhat complicated by the fact that it also explored the effect of using “plain English” in contracts, so it tested four different versions of each contract (with varying degrees of complexity in wording) and thus did not produce a single average reading rate.\footnote{Id. at 74–75. There is no easy way to determine which of the experiment’s different wording categories the contracts at issue here would fall under. Note that one study of complexity in consumer contracts suggests that shorter contracts are easier to comprehend, although that result is confounded somewhat by the fact that the shorter contract was also written in simpler language. See Davis, supra note 37, at 869.} Furthermore, because it involved short excerpts of a few hundred words each rather than long contracts,\footnote{Masson & Waldron, supra note 84, at 71.} its results might not be a precise fit for the contracts at issue here, which averaged 3,016 words; perhaps reading speed increases as one goes along, or declines as one gets tired. Nevertheless, by averaging the reading rates for the various contracts in the study, we can reach a fair estimate of how fast the typical consumer can read contractual language: 177.5 words per minute.\footnote{I arrived at this figure by taking the study’s four entire-document reading rates and averaging them. Id. at 74 tbl.4. One might object to using an average, rather than one of the specific rates that Masson and Waldron found, but it would make little difference; the four rates only ranged from a low of 167.0 to a high of 193.6. Id.}

At that rate, the 74,897 words worth of boilerplate would take just over seven hours to read (assuming, optimistically, that
the reader could keep up the pace for that long without a moment’s rest). Therefore, even if all the contracts had been presented to the reader at the outset of the transaction, such that information acquisition costs were low, the information processing costs of actually reading the contracts would be significant—further evidence that consumers would focus instead on other product features and thus fail to register their preferences regarding boilerplate terms. And, as with acquisition costs, consumers who wish to make an apples-to-apples comparison would also have to read the contracts that competing vendors offered; to do so with the four competing products at issue here would have meant more than fifteen hours of reading.

Yet even this conclusion may underestimate the amount of information overload because, when it comes to boilerplate, compensatory decisionmaking requires not only reading but understanding. Masson and Waldron used two mechanisms (question-answering and paraphrasing) to measure how well their subjects understood the contracts at the above reading speed—and found them wanting. Comprehension improved when archaic terms were replaced, yet even then “performance of subjects on both the question-answering and the paraphrase tasks remained relatively poor (in the best cases average performance ranged from about one-third to two-thirds correct, depending on which aspect of comprehension was measured) and misconceptions were apparent across all versions of the documents.” In other words, even at the relatively slow reading rate of 177.5 words per minute, a substantial subset of consumers

---

88. 74,897 words ÷ 177.5 words/minute = 421.95 minutes, or 7.03 hours.
89. As explained supra note 80, there was some overlap among the boilerplate that each seller offered, so consumers who had already read one seller’s boilerplate would be able to save some time when they got to the next seller.
90. Note also that part of the problem with boilerplate is that consumers can’t know what it covers without at least skimming it. In a sense, then, consumers don’t know whether they should read boilerplate until after they have done so! At a minimum, it takes some effort even to determine whether learning the contents of boilerplate can be discounted in favor of learning about other product features.
91. Masson & Waldron, supra note 84, at 72.
92. Id. at 79.
would have a hard time understanding boilerplate and consequently would have a hard time forming and expressing compensatory and comparative preferences regarding contractual terms.93

Another way to get a sense of information processing costs is to express them in dollars per word. After all, computers are expensive—the four in the study averaged $782.24—and one might tolerate higher information costs when spending a lot of money, so as to ensure that a big expenditure produces a worthwhile payoff. Even under this metric, however, the information costs are significant. The 74,897 words of boilerplate worked out to a weighted average of ninety-three words per dollar spent. For comparison’s sake, imagine having to read ninety-three words of boilerplate each time you buy a can of soda, or 279 words when buying a three-dollar gallon of milk, or 5,580 words when filling a twenty-gallon tank with gas. And none of the contracts here seemed to be related to my choice to buy moderately powerful computers, so more affordable models would likely have come with exactly the same contracts.

In the end, however one chooses to measure information costs, the evidence strongly suggests that the market would not be responsive to consumer preferences regarding the binding boilerplate in this study. Consumers confronted with such purchases would almost certainly ignore the boilerplate entirely; the costs of finding and reading it are quite high, and they would likely consider its terms to be less important than other, more salient features of the transaction. (As a general matter, boilerplate terms relate to rare events—the defective product, the lost shipment, the litigious purchaser—for which the average consumer has little legal exposure.94)

Boilerplate would accordingly be one of those product features that would not make the cut when consumers, overloaded by information, abandon compensatory decisionmaking in favor of satisficing. And when consumers

93. For a discussion of how a representative subset of informed consumers might avert market failure, see infra Part IV.A.
94. Cruz & Hinck, supra note 15, at 663; Eisenberg, supra note 1, at 243; Hillman & Rachlinski, supra note 8, at 443; Rakoff, supra note 1, at 1226.
ignore contract terms, the market ignores contract terms—removing the justification for enforcement of those terms in the first place.

**IV. Responding to Boilerplate’s Defenders**

In Part II, we saw that in transactions of even moderate complexity, consumers satisfice; they abandon compensatory decisionmaking and focus their limited attention on a few product features. Horizontal studies of consumer decisionmaking have begun to provide evidence that such satisficing threatens the function of the boilerplate market. In Part III, I presented new, vertical evidence of this dysfunction, showing that both acquisition costs and processing costs are even higher than the horizontal studies have shown. All in all, then, the evidence suggests that boilerplate is high on the list of features that consumers ignore.

If the enforceability of boilerplate terms depends on their responsiveness to consumer preferences, it seems impossible to defend enforcement when most consumers ignore those terms. Yet scholars have offered two distinct arguments in favor of enforcing boilerplate even in the face of the mounting evidence that it goes unread. The first is an *ex ante* argument that the market works, due to a sufficiently large subset of consumers that actually do read the terms. The second is an *ex post* argument that even when the market fails, reputational concerns keep sellers from unduly aggressive enforcement of one-sided terms.

If either of these arguments is correct, then this Article has not proved anything, and no regulatory intrusion into the market is necessary notwithstanding boilerplate’s information cost problem. In the following discussion, however, I evaluate both arguments, on their own merits and in light of the verticality findings above, and find them wanting.

**A. Informed-Minority Theory**

No market functions perfectly. Information costs and other impediments to compensatory decisionmaking are the rule, not
the exception, even for transactions that involve no contractual terms whatsoever.95 Therefore, although there is compelling evidence that when it comes to boilerplate the market does not function particularly well, the question is whether it functions well enough.

The answer to this question depends partly on the available alternatives to pure market regulation (“well enough” in comparison to what?), an issue to which we will return in Part V. But the answer also depends on how dysfunctional the market is. In a perfectly functioning market, every consumer would costlessly find, read, and understand all adhesive contract terms and would then use compensatory decisionmaking to weigh their impact in the purchase decision. As Alan Schwartz and Louis Wilde have argued, however, a less-than-perfect market, in which only some consumers pay attention to those terms, can produce approximately the same outcome as a fully functional market, so long as sellers compete for those marginal, attentive consumers.96 In other words, if information costs cause most but not all consumers to ignore boilerplate, the consumers who do pay attention—the “informed minority”—might represent the interests of the ignorant, such that the resulting terms would adequately reflect overall consumer preferences.97

Numerous commentators have relied on the Schwartz and Wilde theory to argue that the market adequately responds to consumer preferences even when many consumers are ignorant

95. See Howard Beales et al., The Efficient Regulation of Consumer Information, 24 J.L. & ECON. 491, 512 (1981) (“[V]irtually no consumer product market or associated information market meets the textbook ideal of perfect information and perfect competition.”); Schwartz & Wilde, supra note 10, at 630 (“It is generally recognized . . . that information is never perfect; the decisionmaker’s task, therefore, is to characterize, in terms of the need for intervention, real world states that are intermediate between perfect information and perfect ignorance.”).

96. Schwartz & Wilde, supra note 10, at 638 (“Rather than asking whether an idealized individual is sufficiently informed to maximize his own utility, the appropriate normative inquiry is whether competition among firms for particular groups of searchers is, in any given market, sufficient to generate optimal prices and terms for all consumers.”); see also id. at 659–62 (offering some observations about the implications of this theory for boilerplate).

97. Id. at 660.
about its offerings. In the boilerplate context, however, this argument depends on a number of questionable premises. First, it assumes that the consumers who constitute the informed minority have the same preferences (with regard to contractual terms) as the ignorant majority. Yet we already know that the members of the informed minority are different in at least one important way: they read the boilerplate that the rest of us ignore. Why would we think that they would be different in only that one respect? By definition their payoff from reading a contract is higher than most people’s, which suggests a different valuation for the terms it contains. In short, for the informed-minority theory to work, we must assume heterogeneity in the

98. E.g., Baird, supra note 10, at 936 (“The sophisticated buyer provides protection for those that are entirely ignorant.”); Clayton Gillette, Pre-Approved Contracts for Internet Commerce, 42 Hous. L. Rev. 975, 976–77 (2005) (noting that the inability of a seller to distinguish attentive consumers from inattentive consumers will prevent the seller from acting opportunistically); George L. Priest, A Theory of the Consumer Product Warranty, 90 Yale L.J. 1297, 1347 (1981) (“If a small group of consumers reads warranties and selects among products according to warranty context, manufacturers may be forced to draft warranties responsive to the group’s preferences, even though the large majority of consumers generally neglect warranty terms.”).

99. If they don’t, they cannot represent the majority; they would send sellers the wrong market signal. See Schwartz & Wilde, supra note 10, at 638 (“When the preferences of searchers are positively correlated with the preferences of nonsearchers, competition among firms for searchers should tend to protect all consumers.”).

100. See Cruz & Hinck, supra note 15, at 671 (“All that we can know [about members of the informed minority] is that they are particularly sensitive to some aspect of the contract—not that the average consumer (or even just one other consumer outside the marginal minority) necessarily shares their preferences.”); id. at 676 (“[T]here is no guarantee—indeed it seems unlikely—that the marginal consumer will be typical of other consumers.”).

101. Another possibility is that the informed minority comprises repeat purchasers, who choose to incur the information costs because they can amortize them across a number of purchases. Such purchasers might have idiosyncratic preferences regarding the content of terms; for example, they would not care as much about a class action waiver because their high volume would make it cost-efficient to sue individually. See Gillette, supra note 7, at 694. But even if their preferences were the same as the ignorant majority’s, repeat buyers are more likely to be accommodated when something goes wrong (even if the contract does not require such accommodation)—which means that they will be less worried about contractual guarantees and therefore less able to act as a proxy for nonreaders. Id. at 692. And of course, the amortization argument assumes that the boilerplate does not change over time; try telling that to an iTunes user.
minority’s proclivity to read boilerplate but homogeneity in its preferences for boilerplate terms—a dubious pair of assumptions and a slim reed on which to build a theory of enforcement.

Second, even if the informed minority has the same boilerplate preferences as the ignorant majority, the former cannot adequately represent the latter if sellers can differentiate between the two groups.\textsuperscript{102} Sellers that can differentiate will offer better terms to the informed minority and continue to offer one-sided terms to the uninformed majority, thereby delinking the fates of the two groups and returning the latter market to a largely nonresponsive state.\textsuperscript{103} How easy such differentiation would be is an open question, dependent on the particular industry at issue.\textsuperscript{104} In the online context, for example, one could easily imagine a website offering different terms to different consumers based on their browsing habits and the attention they pay to boilerplate.\textsuperscript{105}

\begin{footnotesize}
\begin{itemize}
\item 102. See Schwartz & Wilde, supra note 10, at 638 (assuming that “it is usually too expensive for firms to distinguish among extensive, moderate, and nonsearchers” and that “it would often be too expensive to draft different contracts for each of these groups even if they could conveniently be identified”); id. at 663–64 (explicitly addressing this assumption).
\item 104. On a related note, Ted Cruz and Jeffrey Hinck see an inherent tension between, on the one hand, the number of informed consumers and, on the other, the ability of sellers to differentiate. The informed-minority theory grows more plausible as the number of attentive readers necessary to adequately represent the ignorant majority decreases—but Cruz and Hinck argue that “[t]he ability to identify and segregate consumers . . . becomes much easier as the number of informed buyers becomes smaller” and therefore “[u]nder such conditions, the seller needs only to target a small proportion of his buyers to defeat the effect of an informed minority upon the terms offered to uninformed consumers.” Cruz & Hinck, supra note 15, at 656.
\item 105. See Robert A. Hillman & Ibrahim Barakat, Warranties and Disclaimers in the Electronic Age, 11 Yale J.L. & Tech. 1, 17 (2009) (exploring the possibility that industries can “employ new e-technologies to segregate readers and offer them more advantageous terms”); Hillman & Rachlinski, supra note 8, at 471–72 (“E-businesses can use data on consumer behavior collected from their prior
\end{itemize}
\end{footnotesize}
Finally, the market is likely to underproduce informed consumers because the informed-minority theory creates a collective action problem: only a few consumers incur the information costs, but all consumers reap the resulting benefits. This means that even consumers who are inclined to read the boilerplate have an incentive to ignore it, sit back, and free-ride on the efforts of others.\textsuperscript{106} In contrast, the seller always pays careful attention to boilerplate terms because the seller is a party to every transaction to which the terms apply and thus internalizes all their benefits.\textsuperscript{107} (Indeed, the seller is the party that drafted the terms.)

These arguments give us reason to doubt the informed-minority theory, but the proof of the pudding is in the eating: are boilerplate-reading consumers really common enough to adequately represent the interests of their nonreading counterparts? Schwartz and Wilde themselves are skeptical, estimating that more than one in three consumers would have to read the boilerplate for the informed-minority concept to work.\textsuperscript{108} Modeling and industry-specific empirics suggest that their skepticism is well-founded.\textsuperscript{109} For example, one study of software transactions and offer different terms to those consumers who are most likely to read the boilerplate (or who have already read it during a prior site visit).\textsuperscript{110} See Beales et al., supra note 95, at 503; Cruz & Hinck, supra note 15, at 668; Hillman & Rachlinski, supra note 8, at 447.

\textsuperscript{106} See Beales et al., supra note 95, at 503; Cruz & Hinck, supra note 15, at 668; Hillman & Rachlinski, supra note 8, at 447.

\textsuperscript{107} This is merely another way of stating the information asymmetry problem familiar to those who study adhesion contracts. See, e.g., Posner, supra note 8, at 145; Omri Ben-Shahar, Fixing Unfair Contracts, 63 STAN. L. REV. 869, 901 (2011); Eisenberg, supra note 1, at 243.

\textsuperscript{108} Schwartz & Wilde, supra note 10, at 661.

\textsuperscript{109} For modeling, see Cruz & Hinck, supra note 15, at 648–55 (concluding that under reasonable assumptions “the proportion of consumers that must be term-informed is somewhere between 10% and 30%, both of which are relatively large given the impediments to formation of an informed minority”). For empirics, see Yannis Bakos et al., Does Anyone Read the Fine Print? Testing a Law and Economics Approach to Standard Form Contracts 3 (NYU Law & Econ. Working Paper No. 195, 2009) (finding an “informed minority . . . orders of magnitude smaller than the required informed minority size in the theoretical examples of Schwartz and Wilde”), available at http://lsr.nellco.org/nyu_lewp/195.
retailers who made their licensing agreements available online found that only two out of every one thousand shoppers accessed the agreements for more than one second—and even that small subset averaged less than a minute spent on the relevant web page.110

Yet even these caveats do not fully capture the limitations of the informed-minority theory because they consider boilerplate in the abstract rather than as part of a complex, integrated transaction. In other words, for a real-world perspective, we must examine the impact of verticality on the informed-minority theory. And as it turns out, verticality challenges the informed-minority theory in two particular ways.

First, this Article’s vertical study reveals information costs that are much higher than previous studies have shown; the informed minority acquires almost all of the boilerplate late in the transaction and faces more than seven hours of reading.111 As these information costs increase, the number of contract readers will shrink, all else being equal. After all, even those with an idiosyncratic tendency to read boilerplate must deal with their own finite resources and cognitive limitations. And as the number of readers shrinks, so does the likelihood that the informed minority will be of sufficient size to serve as an adequate proxy.

Second, the added complexity of a fully vertical transaction may increase the differences between the preferences of the informed minority and the preferences of the ignorant majority, such that the former cannot represent the latter in the marketplace. Homogeneity of demand with regard to a single product is one thing; homogeneity of demand with regard to an entire series of products is another.112 As discussed above, in a vertical-boilerplate situation it can be prohibitively hard for the consumer to reject a single contract or term; one usually either

110. Bakos et al., supra note 109, at 3, 26.
111. See supra note 88 and accompanying text.
112. It is possible, however, that increased complexity brings with it increased, rather than decreased, homogeneity—i.e., that heterogeneity in individual demand is smoothed out by bundling more products into the transaction. See supra notes 58–62 and accompanying text (discussing bundling). If this is true, then added complexity would not make the informed minority less able to represent the ignorant majority.
rejects or accepts the entire transaction.\textsuperscript{113} Therefore, as more sellers and features are introduced—i.e., as vertical complexity increases—the ability to send a signal to the market about a single product feature (such as a boilerplate term) becomes increasingly muted.

In the end, then, the informed-minority theory gives us little reason to be confident in the market’s responsiveness to consumers’ boilerplate preferences. In theory, a subset of attentive consumers could serve as a proxy for the inattentive remainder. But in practice, the empirical evidence points the other way, and the assumptions on which the informed-minority argument rests prove far-fetched—even more so once it is applied to a real-world vertical transaction that involves multiple adhesion contracts and other indicators of high product complexity.

\textbf{B. Reputational Theory}

Commentators have also offered a second, separate defense of boilerplate’s enforceability in the face of seemingly insuperable information costs: the reputational theory. The gist of the reputational theory is that it does not matter if boilerplate terms fail to respond to consumer preferences at the time of contract formation because, when the opportunity arises to enforce a term at the tail end of the transaction, sellers waive enforcement and accommodate the consumer.\textsuperscript{114}

And why would the seller do so? For reputational reasons. It wants to retain the consumer’s future business and avoid becoming known for poor customer service.\textsuperscript{115} Thus boilerplate is

\begin{footnotesize}
\footnotesize
\begin{itemize}
\item \textsuperscript{113} Supra Part III.C.1.
\item \textsuperscript{115} See sources cited supra note 114.
\end{itemize}
\end{footnotesize}
not a concern (the argument goes), because even if the market
fails to rid the transaction of unwanted terms \textit{ex ante},
reputational concerns will get rid of them \textit{ex post}.\textsuperscript{116}

If reputational constraints are sufficient to govern the
outcome of disputes, however, why do we need boilerplate at
all?\textsuperscript{117} The answer that the theory’s proponents give is that a
seller will not \textit{always} disregard contractual clauses in favor of a
customer service strategy.\textsuperscript{118} Instead, the seller will enforce
boilerplate selectively, reserving its use for those few
opportunistic consumers who seek to exploit the seller’s “the
customer is always right” instinct—e.g., the conniving buyer who
damages a product and then tries to return it as defective.\textsuperscript{119} The
boilerplate terms are thus relevant because they form an \textit{ex ante}

\begin{itemize}
  \item[116.] See sources cited \textit{supra} note 114.
  \item[117.] See Hillman & Barakat, \textit{supra} note 105, at 12 (“If accepted, the
\hspace{1em} \text{[reputational]} argument means that we do not need contract law at all.”).
\hspace{1em} Indeed, one well-known study of actual business-to-business practices found
\hspace{1em} that reputation dominates contracts as the key determinant in dispute
\hspace{1em} resolution.
\hspace{1em} Even where the parties have a detailed and carefully planned
\hspace{1em} agreement which indicates what is to happen if, say, the seller fails to
deliver on time, often they will never refer to the agreement but will
negotiate a solution when the problem arises apparently as if there
had never been any original contract.
\hspace{1em} Stewart Macaulay, \textit{Noncontractual Relations in Business: A Preliminary Study},
\hspace{1em} 28 \textit{A M. SOC. REV.} 55, 61 (1963) [hereinafter Macaulay, \textit{Noncontractual
\hspace{1em} Relations}].
\hspace{1em} 118. See sources cited \textit{supra} note 114.
\hspace{1em} 119. Clay Gillette uses a similar example in \textit{Rolling Contracts as an Agency
\hspace{1em} Problem}, \textit{supra} note 7, at 704; see also Bebchuk & Posner, \textit{supra} note 114, at
\hspace{1em} 834 (using an example of a hotel guest who checks out late). Some of the theory’s
\hspace{1em} proponents further explain that, although boilerplate gives sellers an advantage
\hspace{1em} over consumers, that advantage merely counterbalances an advantage that
\hspace{1em} consumers have: namely, sellers have a reputation in the marketplace that tells
\hspace{1em} consumers which sellers to avoid, but consumers have no reputation that tells
\hspace{1em} sellers which consumers to avoid. Thus there is a need for an \textit{ex post}
\hspace{1em} contractual mechanism to weed out the “bad” consumers. Bebchuk & Posner, \textit{supra} note
\hspace{1em} 114, at 829–30; Gillette, \textit{supra} note 7, at 704. \textit{But see eBay, All About Feedback
\hspace{1em} Policies}, http://pages.ebay.com/help/policies/feedback-ov.html (last visited Feb. 4,
\hspace{1em} 2013) (describing a feedback system that not only allows eBay buyers to rate
\hspace{1em} sellers but also allows sellers to rate buyers) (on file with the Washington and
\hspace{1em} Lee Law Review).
baseline from which the seller has the *ex post* discretion to depart.\textsuperscript{120}

Unlike the informed-minority theory, the reputational theory neatly solves the problem of information costs, in both their horizontal and vertical dimensions, because under its approach boilerplate terms are only enforced against the rare opportunist who deserves no better. My vertical empirics thus do not directly speak to the reputational theory. (Indeed, if the theory is true, there is no boilerplate problem to be solved.) Nevertheless, the theory fails to survive when evaluated in light of my broader theme: examining the boilerplate problem from the perspective of what actually takes place in the real world. When one views the reputational defense from that perspective, its foundations fall away; it is revealed to be merely an interesting theory, rather than a practical justification for enforcing real-world boilerplate against real-world consumers.

The reputational theory proves unrealistic for three reasons. First, the theory relies on questionable assumptions about sellers’ ability to differentiate among consumers. For the reputational dynamic to rescue boilerplate, we must expect a seller to be unable to distinguish opportunistic buyers *ex ante* but somehow be able to distinguish them *ex post* (because a seller that can make the distinction *ex ante* will simply impose different boilerplate terms on the different types of buyers).\textsuperscript{121} Both propositions are problematic. How certain are we that sellers can accurately identify the consumers who really received defective products and distinguish them from the opportunists? And how certain are we that sellers could draw such distinctions at the tail

\textsuperscript{120}. Posner, *supra* note 8, at 145 (“[T]he seller’s right to stand on the contract as written will protect it against opportunistic buyers.”); Johnston, *supra* note 114, at 878 (“The key to understanding why a firm can benefit by allowing its employees to forgive some customers’ contract breaches lies in the recognition that not all existing customers are worth keeping.”); Rakoff, *supra* note 1, at 1221 (“[I]f legal liabilities are set lower than the obligations that the firm recognizes in its actual practice, the gap can provide room to maneuver in the face of inevitable adversity.”).

\textsuperscript{121}. See Johnston, *supra* note 114, at 879 (describing sellers’ *ex post* enforcement discretion as “a substitute for ex-ante screening”).
end of the transaction but not at the front end? It seems just as likely, if not more so, that the presence of one-sided boilerplate would scare off the “good” customers from even asking for accommodation. Indeed, one reasonable definition of an opportunistic customer would be someone who hounds sellers for concessions to which he or she is not legally entitled—which would imply no accommodation for anyone who bothers to ask.


123. See Gillette, supra note 7, at 706 (“Sellers may use contract terms in an in terrorem effort to deter requests for redress, or as an initial response to buyer complaints.”); Dennis P. Stolle & Andrew J. Slain, Standard Form Contracts and Contract Schemas: A Preliminary Investigation of the Effects of Exculpatory Clauses on Consumers’ Propensity to Sue, 15 Behav. Sci. & L. 83, 91 (1997) (finding that exculpatory clauses in contracts have a deterrent effect on propensity to seek compensation when they are read); Charles A. Sullivan, The Puzzling Persistence of Unenforceable Contract Terms, 70 Ohio St. L.J. 1127, 1136 (2010) (explaining that unenforceable provisions in contracts deter uninformed parties from exercising contractual rights). It may seem odd for consumers to be scared off by onerous boilerplate terms, given the compelling evidence that they don’t read boilerplate. But here we are not talking about reading boilerplate before making the purchase decision; we are talking about reading it much later, once something has gone wrong. Consumers who do not read boilerplate when entering into a transaction may well read it later when the transaction takes a turn for the worse and they need to assess the possibility of legal recourse.

124. In addition, some empirical research suggests that lower-class consumers are more likely to view contracts as binding than upper-class consumers. See Zev J. Eigen, The Devil in the Details: The Interrelationship Among Citizenship, Rule of Law and Form-Adhesive Contracts, 41 Conn. L. Rev. 381, 421–22 (2008) (presenting preliminary findings). If this is true, then the willingness to seek extracontractual accommodation may depend not just on whether a consumer is “good” or “bad” but on his or her socioeconomic status—an unsettling possibility.
Second, the reputational theory resurrects the employee as an agent with discretion to negotiate. Such employee discretion contrasts with one of the longstanding justifications for enforcing boilerplate in the first place, namely the need to avoid the agency costs of individualized negotiation—e.g., to keep employees from succumbing too easily to Marshall Field’s “give the lady what she wants” philosophy. Of course, the ex post negotiation that the reputational theory calls for would occur less frequently, and thus would arguably impose lower agency costs, than the ex ante negotiation that would occur if all contracts were individually bargained. Nevertheless, for the reputational theory to work we must believe that the benefits of this approach would exceed the inevitable costs of giving employees the discretion to waive terms on a case-by-case basis. The theory’s proponents provide no evidence that this would actually be the case in the real world.

Finally, even if we assume that employees can differentiate between the good and bad customers, and we zero out the agency costs (so that the employee invariably acts in the best interests of the seller), we still must believe that it is in the seller’s best interests to accommodate the “good” consumer. Sellers may derive reputational benefit from forgoing enforcement of a boilerplate term, but that benefit comes at a price. For example, a seller incurs a cost each time it replaces a product.

---

125. See, e.g., Johnston, supra note 114, at 878–80 (highlighting the importance of managerial discretion to waive one-sided terms when dealing with “good, high value customers”).

126. As legend has it, Field—the founder of the eponymous Chicago department store—uttered this phrase as a rebuke to an employee who was arguing with a female customer. It soon became the store’s motto. Mark D. Bauer, “Give The Lady What She Wants”—As Long As It Is Macy’s, 80 Temp. L. Rev. 949, 950 (2007). Yet it is that very customer service attitude that boilerplate is meant to protect against. Stewart Macaulay, Private Legislation and the Duty to Read—Business by IBM Machine, the Law of Contracts and Credit Cards, 19 Vand. L. Rev. 1051, 1059 (1966) (warning against employees’ “the customer is always right” attitude); Macaulay, Noncontractual Relations, supra note 117, at 65 (citing the need to “keep . . . salemen from making concessions to the customer”); Rakoff, supra note 1, at 1222–24 (warning against “wayward sales personnel”). Perhaps that’s why Marshall Field & Co. is no more. See Bauer, supra note 126, at 950 (revealing that the store is now Macy’s).

127. See, e.g., Johnston, supra note 114, at 878–80 (discussing the benefits of “granting discretionary forgiveness” in loan-repayment agreements in order to build “profitable, long-term relationships” with “high-value” borrowers).
it is not contractually obligated to do so, then it will not accommodate a replacement request, even from a “good” customer, unless the reputational benefit exceeds that cost.

That cost–benefit calculation involves a daunting number of unknowns. On the benefit side of the ledger, it assumes that consumers generate and share meaningful quantities of accurate reputational information.\(^\text{128}\) But like any informational asset, reputation is a nonrival public good; consumers therefore do not fully internalize the gains from publicizing their experience with disreputable sellers, which means that such accounts are likely to be underproduced.\(^\text{129}\) And even if we ignore the public-goods problem, sellers may find it more cost-effective to buttress their reputations through advertising, marketing, and outright manipulation than through accommodating complaining customers one by one.\(^\text{130}\) Indeed, short-term and one-off sellers

---

\(^\text{128}\) Although no one has conducted an empirical study of boilerplate that focuses on the reputational factor, Victoria Plaut and Robert Bartlett found that associating a click-wrap contract with Google (presumably a company with a good reputation) had no statistically significant effect on whether consumers read the contract’s terms. Victoria C. Plaut & Robert P. Bartlett III, Blind Consent? A Social Psychological Investigation of Non-Readership of Click-Through Agreements, 36 L. & HUM. BEHAV. 293, 302, 305 (2012).

\(^\text{129}\) See Gillette, supra note 7, at 704; Hillman & Rachlinski, supra note 8, at 447; cf. Rakoff, Sociology of Boilerplate, supra note 122, at 1236 (criticizing reputation-theory proponents for inter alia “offer[ing] no model of how the market in reputation works, or of why the values it generates are responsive to anything other than firms’ fears of how much reputational damage particular claimants are, for a myriad of possible reasons, in a position to cause”). Gillette uses the public-goods issue to show that sellers are unlikely to share information on disreputable consumers, but it applies equally to the sharing of information by consumers about sellers. See Gillette, supra note 7, at 704.

\(^\text{130}\) See, e.g., ReputationReset, http://www.reputationreset.com (last visited Feb. 4, 2013) (“We help clean up the bad [search engine] results, restoring your reputation and making sure poor reviews, misinformation, or bad-mouthing competitors no longer hold you back.”) (on file with the Washington and Lee Law Review). The converse situation— websites or search engines that aggregate information on sellers’ boilerplate—is, like the reputational theory, an idea that is conceptually appealing but that has found no purchase in the real world. Bakos et al., supra note 109, at 33–34 (finding it “highly unlikely that shoppers are, to an important extent, becoming informed about EULA terms by consulting other online sources”); Ben-Shahar, Opportunity To Read, supra note 103, at 22–25 (evaluating the possibility of rating services for contracts but finding little hope); Florencia Marotta-Wurgler, Will Increased Disclosure Help? Evaluating the Recommendations of the ALI’s “Principles of the Law of Software
probably have little interest in building a reputation through any means at all.  

On the cost side of the ledger, some contractual prerogatives may be too valuable to give up, reputational consequences notwithstanding. For example, one of the most litigated issues in the world of boilerplate is the consumer class-action waiver. Does anyone seriously believe that a seller would ever voluntarily forgo enforcement of such a clause? In addition, those provisions that are not too valuable to forgo may come up too infrequently to have any reputational impact one way or the other, given that boilerplate terms tend to govern rare contingencies.  

For all these reasons, the reputational theory is too unreliable to rescue boilerplate from its shortcomings. True, it identifies a dynamic that might cause a seller to refrain from enforcing an adhesive term against a consumer. But only if the planets align in a particular—and particularly unrealistic—way. The theory falls short of providing support for the general proposition that allowing unread boilerplate to be enforced at the whim of the seller is in the public interest.  

We are therefore left with a question. If information costs render the market dysfunctional, and neither the informed-minority theory nor the reputational dynamic rescues it from that dysfunction, what other solutions to the boilerplate problem exist? For the answer, we turn to Part V.  

---

*Contracts*, 78 U. Chi. L. Rev. 165, 184 & n.65 (2011) (discussing the “EULAlyzer” boilerplate analysis program but finding it rarely used).  

131. See Hillman & Rachlinski, supra note 8, at 444.  

132. E.g., AT&T Mobility LLC v. Concepcion, 131 S. Ct. 1740, 1750–51 (2011) (upholding consumer class-action waiver against unconscionability challenge). Moreover, to be truly effective, the waiver of a class-action prohibition would have to be done on a grand scale, rather than made on a case-by-case basis, because by definition a class action cannot be filed by one consumer at a time.  

133. See Cruz & Hinck, supra note 15, at 663 (“[T]he probability of any single customer being affected by any given contract term is usually quite small.”); Hillman & Rachlinski, supra note 8, at 443 (noting that “standard terms cover events that are unlikely to occur”).
V. Toward Enforceable Boilerplate

We now know that once a transaction reaches a certain level of complexity, consumers simply will not read boilerplate. The functioning market that forms the basis for boilerplate’s enforceability is thus largely an illusion, particularly when one considers the full verticality of the transactions that consumers encounter in the real world. No informed minority rescues nonreading consumers. And reputational concerns will not reliably constrain sellers from enforcing the boilerplate.

So where do we go from here? In the following discussion, I first review the two solutions at the extreme: enforcing no boilerplate and enforcing all of it. I then suggest a pair of solutions that fall in between the extremes. The first of the two is the more conventional in that it operates within existing unconscionability doctrine, but it reforms the doctrine using insights from the verticality approach. The second is my more radical, “forced salience” mechanism, which makes sellers and consumers alike confront boilerplate’s information costs—and, by confronting them, minimize them.

A. Two Extremes

1. Get Rid of All Boilerplate

The most extreme solution to the boilerplate problem would be to refuse to allow sellers to even include boilerplate in any transaction of minimal complexity, or at least to declare it all unenforceable. Given the evidence of market failure, this approach is not as crazy as it first sounds. Yet it goes too far because even a one-sided boilerplate term is sometimes efficient. When that’s the case, undoing a term would be worse than preserving it.

To understand why some one-sided boilerplate is worth preserving, one must appreciate that boilerplate presents an example of the classic market for lemons.134 Although consumers

---

134. For the original conception of the lemons theory, see George A. Akerlof, The Market for “Lemons”: Quality Uncertainty and the Market Mechanism, 84
do not usually consider boilerplate terms when making a purchase, they do pay attention to other features of the product; indeed, the whole point of satisficing is that certain features are ignored so that others can be considered. So some features (such as price, presumably) will be salient even though others (such as boilerplate) will not. Thus the lemons problem: sellers will decrease the quality of the nonsalient features and use the resulting savings to make the salient features more attractive. Boilerplate, as a nonsalient feature, will accordingly be full of terms that reduce seller costs and shift risks to consumers, and sellers will use the money they save to lower the price of the product. Commentators have observed this dynamic in the adhesive contract terms that accompany products as diverse as cell-phone plans, bank accounts, and credit cards, and some empirics for the software industry lend support to its presence as well.

If the lemons effect consistently shifts boilerplate’s costs and risks to consumers, without any informed consent on their part, why would boilerplate ever be worth preserving? The answer is that sometimes consumers can bear those costs and risks more

---

Q.J. Econ. 488, 488–500 (1970). For those who first applied it to boilerplate, see Beales et al., supra note 95, at 510–11; Eisenberg, supra note 1, at 244 & n.158. For the most thorough explanation of boilerplate’s lemons dynamic, see Korobkin, supra note 1, at 1206.

135. See Korobkin, supra note 1, at 1206 (“[M]arket competition actually will force sellers to provide low-quality non-salient attributes in order to save costs that will be passed along to buyers in the form of lower prices.”).

136. Id.


138. Eisenberg, supra note 1, at 244.

139. Oren Bar-Gill, Seduction by Plastic, 98 Nw. U. L. Rev. 1373, 1401–02 (2004). For Bar-Gill, the nonsalience of credit-card terms is rooted in behavioral biases, rather than in information costs, but the effect is the same: consumers routinely discount certain features of the transaction, which gives sellers an incentive to reduce their quality so as to improve the more salient features. Id. at 1400–01.

140. Marotta-Wurgler, Empirical Analysis, supra note 6, at 680 (studying 647 software contracts from 598 different companies and 114 distinct markets and finding that they were “almost without exception tilted toward the seller, relative to the relevant default rules—some sharply so”).
efficiently than sellers. As Russell Korobkin explains, a “low quality” boilerplate term (i.e., a term that favors sellers over consumers) is not necessarily an inefficient one. There are some pro-seller terms to which a fully informed consumer would want to agree because the resulting savings in price would be greater than the cost of the term to the consumer. For example, a software company might realize substantial savings if it could contractually limit installation of its programs to one computer per customer. Such a term would arguably be “low quality,” in that it constrains consumers, but many consumers might nevertheless be happy to accept it because they never planned to install the program on multiple machines and so can pay a lower price without giving up much in return.

In short, if a term is efficient, it should be enforced whether it emerges by sheer luck from a dysfunctional market for lemons or by design from a functioning market with robust competition and universal salience of product features. The alternative is to allow consumers to have their cake and eat it too—i.e., to enjoy the lower price but then escape enforcement of the term. Faced with that possibility, sellers would change the boilerplate to allocate such risks to themselves, inefficiently, and charge higher prices to make up for it. Such an outcome would do no favors for seller or consumer.

141. Korobkin, supra note 1, at 1283. Korobkin also points out that the lemons effect can hurt sellers as well as consumers because it prevents sellers from competing on nonsalient features. Id. at 1206. Indeed, the main contribution of the lemons theory is to show how high-quality products are driven out of the marketplace whenever quality is not salient. See Akerlof, supra note 134, at 489–90 (explaining the lemons theory through the market for used cars).

142. See Rakoff, supra note 1, at 1222 (“Costs saved by shifting risks to the customer via form terms may well be returned to the customer by means of lower prices.”).


144. See Korobkin, supra note 1, at 1213 (explaining that, if courts would refuse to enforce market-efficient terms, “the majority of resulting contracts would be inefficient and the majority of buyers made worse off”).
2. Enforce All Boilerplate

The solution at the other extreme is to enforce all boilerplate, information costs or not, because we can’t do without it. This is a common response of boilerplate’s defenders to the information cost problem; they admit its shortcomings but nonetheless insist on its inevitability, under the assumption that transactions cannot proceed unless they are governed by reams of unread contracts. Eric Posner—one of ProCD’s few champions in the legal academy—bases his support for the case on such an argument: “Contracts are long and detailed by necessity. To sell goods, manufacturers need to be able to put just the crucial terms on the box (such as the price) along with useful information, and to omit information of little use to consumers, including obvious information.”¹⁴⁵ And Judge Easterbrook, the author of ProCD, envisions only two possible worlds: one in which consumers pay now but get the terms later, and another in which cashiers read pages of contracts to customers before ringing up sales, in a “droning voice [that] would anesthetize rather than enlighten many potential buyers.”¹⁴⁶

This insistence on boilerplate’s importance is overblown. Consider my Toshiba computer purchase. If all 96,641 words of boilerplate disappeared from that transaction, the heavens would not fall. Instead, default rules would fill in the blanks, and the transaction would proceed with surprisingly little disruption. In place of Toshiba customer service contracts, I would have an implied warranty of merchantability and express warranties based on statements Toshiba made about its goods.¹⁴⁷ In place of

¹⁴⁷. See U.C.C. § 2-314 (1977) (implied warranty of merchantability); id. § 2-313 (express warranties). Note that the warranty of merchantability allows for some variation in the quality of goods and that the generally accepted practice in the relevant industry is a major factor in determining merchantability. See Step-Saver Data Sys., Inc. v. Wyse Tech., 752 F. Supp. 181, 191 (E.D. Pa. 1990) (“Acceptance in the trade . . . has long been a reliable barometer for determining whether a particular product is merchantable.”), aff’d in relevant part, rev’d in irrelevant part, 939 F.2d 91 (3d Cir. 1991). Therefore, given the bugs that one encounters as a matter of course when dealing with computers, one would not expect the warranty to have much force. For example, the Step-Saver court held
Microsoft’s licenses, I would rely on the Copyright Act’s limited grant of user privileges to software purchasers. Instead of arbitrating various disputes with sellers, I would use the court system. And so forth.

This is not to say that a transaction can always proceed if the associated contracts disappear. To the contrary, in some instances contractual terms are essential, and getting rid of them would accordingly mean getting rid of the entire transaction. For example, no default rule will be able to tell me how soon I must return my rental car; a specific contractual term would be needed. Such essential terms, however, are generally not hidden in boilerplate. Rather, they tend to be presented up front, in a salient context, often with an array of options from which the consumer is forced to consciously choose. No one rents a car without having made a deliberate choice about the rental period. Or consider my computer purchase, in which the website walked me through the options for a number of important features, both contractual (shipping terms, service plan, extended warranty) and noncontractual (monitor, memory, size of hard drive). None of these web pages presented me with any boilerplate, and none of them was included in the boilerplate tally. Getting rid of the boilerplate would thus not have destroyed the transaction.

The prospect of unenforceable boilerplate does, however, draw attention to the importance of formulating appropriate default rules to fill the gap that unenforceable boilerplate would leave. The traditional view of default rules holds that they merely insert into the contract whatever term the parties would most likely have negotiated themselves—i.e., the efficient term.

that computer hardware’s incompatibility with popular software did not violate the merchantability standard. Id. The software programs in question included WordPerfect (which qualified as popular back in the early 1990s despite its near-mortibund existence today).


149. Cf. Preston & McCann, supra note 71, at 9 (commenting on ProCD by observing that “the seven-by-nine-by-three-inch box in which software is sold would provide plenty of space if the terms were limited to the reasonable number of terms necessary to protect intellectual property written in plain English”).

150. See, e.g., Posner, supra note 8, at 120 (calling for gap-filling default rules that “mimic the terms that the parties would have incorporated”); Frank
Modern scholarship, however, has revealed that substituting an inefficient term can sometimes be better because it forces a party to reveal welfare-enhancing information. Attractive as that prospect might be in some cases, it is exactly the wrong approach to take for boilerplate terms because the whole point of the boilerplate problem is that consumers already have more information than they can deal with. This suggests that when the possibility of unenforceable boilerplate is significant, the applicable default rule should be formulated based on the traditional gap-filling theory. Otherwise, we end up with the worst of both worlds: an unenforceable contract term displaced by an inefficient default rule.

In short, getting rid of all boilerplate is too extreme, but so is preserving all boilerplate. Therefore, the correct legal solution must appreciate that some boilerplate is good for consumers, even if produced by a dysfunctional market, and that doing away with boilerplate will not necessarily explode the transaction, as long as the law remains attentive to filling the resulting gap with appropriate default rules. With those caveats in mind, let us turn to a pair of more promising solutions to the boilerplate problem.

---

151. See Ian Ayres & Robert Gertner, Filling Gaps in Incomplete Contracts: An Economic Theory of Default Rules, 99 Yale L.J. 87, 97–100 (1989) (discussing penalty default rules and arguing that inefficient default contract terms should be enforced against the party that possesses more information, thus encouraging both information sharing and explicit “contract[ing] around” the inefficient rules). Ayres and Gertner also discuss using defaults to incentivize the sharing of information with third parties, such as courts, rather than with the other contracting party. Id. at 95–98. Those defaults are less relevant here.

152. See supra Part II.B.2 (discussing consumers’ cognitive limitations and the problem of information overload).
B. Two In-Betweens

Having rejected the solutions at the two extremes, let us now consider two approaches in between. The first is to use verticality to adjust contract law’s well-known approach to adhesion contracts—namely, the unconscionability doctrine—so as to better contextualize its inquiry and properly allocate its burdens of proof. The second is to directly require sellers to lower the information costs that their boilerplate imposes, both by making it more salient and by reducing its overall volume.

1. Unconscionability

When concerns about boilerplate arise, contract law turns to the unconscionability doctrine. The unconscionability analysis focuses on whether there were defects in the bargaining process (procedural unconscionability) and on whether the resulting contract contains grossly one-sided terms (substantive unconscionability). Some courts will invalidate a contract only

153. To be sure, there are other applicable doctrines, such as the “reasonable expectations” doctrine, see, e.g., Sparks v. St. Paul Ins. Co., 495 A.2d 406, 414 (N.J. 1985) (ruling that consent to an insurance contract “can be inferred only to the extent that the policy language conforms to public expectations and commercially reasonable standards”), and the Restatement’s rule for standardized agreements, see RESTATEMENT (SECOND) OF CONTRACTS § 211(3) (1981) (“Where the other party has reason to believe that the party manifesting such assent would not do so if he knew that the [standardized] writing contained a particular term, the term is not part of the agreement.”). But as Russell Korobkin points out, both doctrines suffer from infirmities similar to those of unconscionability, and in any event, they “appear to have been almost completely forgotten by courts, at least outside of the realm of insurance contracts.” Korobkin, supra note 1, at 1270–71. In any event, to the extent that these doctrines remain appealing, they can easily be folded into my substantive unconscionability proposal.

154. The distinction between procedural and substantive unconscionability originated in Arthur Allen Leff, Unconscionability and the Code—The Emperor’s New Clause, 115 U. Pa. L. Rev. 485, 487 (1967) (referring to “bargaining naughtiness as ‘procedural unconscionability’” and to “evils in the resulting contract as ‘substantive unconscionability’”) and has become a staple of the case law. See, e.g., Morrison v. Circuit City Stores, Inc., 317 F.3d 646, 666 (6th Cir. 2003) (discussing Ohio law); Williams v. Walker-Thomas Furniture Co., 350 F.2d 445, 449 (D.C. Cir. 1965) (not using the terms “procedural” and “substantive,” but explaining that “[u]nconscionability has generally been
when they find both kinds of unconscionability, whereas others are satisfied with just one of the two subcategories. The following proposal requires proof of both but makes a small adjustment to procedural unconscionability and a bigger adjustment to substantive unconscionability. In combination, these changes would go a long way toward addressing the infirmities in the way contract law approaches boilerplate.

Start with procedural unconscionability. As Russell Korobkin points out, the procedural inquiry asks whether the allegedly unconscionable term was presented in such a way as to attract the reader’s attention—was it on the first page? was it in boldface? but ignores the fact that cognitively overburdened consumers might fail to process even a prominent term. In contrast, Korobkin observes, a better procedural unconscionability analysis would not be satisfied with examining whether a boilerplate term was prominent, but would instead directly address whether it was salient to consumers as a class.

What does verticality add to this conversation? Context. In order to determine whether a term is truly salient, courts must look beyond the four corners of the contract itself and consider

recognized to include an absence of meaningful choice on the part of one of the parties together with contract terms which are unreasonably favorable to the other party”).

155. See, e.g., Gillman v. Chase Manhattan Bank, N.A., 534 N.E.2d 824, 828 (N.Y. 1988) (“A determination of unconscionability generally requires that a contract was both procedurally and substantially unconscionable when made . . . .”).

156. See, e.g., Maxwell v. Fid. Fin. Servs., Inc., 907 P.2d 51, 59 (Ariz. 1995) (concluding that “a claim of unconscionability can be established with a showing of substantive unconscionability alone”).


158. Korobkin, supra note 1, at 1272–73. Procedural unconscionability has other failings as well. See id. at 1258–68. Most relevant here, however, is its overly narrow emphasis on prominence of the term at issue.

159. See id. at 1279–83 (arguing that salience is the key feature of procedural unconscionability and calling for courts to “initially inquire into whether a challenged term is salient or non-salient” prior to “considering the possibility of invalidating a form contract term”).
the entire transaction from start to finish, with all its information costs. A term may be unresponsive to market forces, notwithstanding its prominence, if it arrives after the consumer has invested considerable time in the purchase (the acquisition cost issue) or has concentrated his or her limited attention on other product features (the processing cost issue). And a term that is not prominent may nonetheless be salient if it is an important part of the transaction, regardless of how inconspicuous it seems in the abstract.

Requiring courts to relax their focus on the particular text of the boilerplate and account for the big picture is not a radical change, and in fact is consistent with the rhetoric of procedural unconscionability jurisprudence. The consumer’s “age, education, intelligence, [and] business acumen and experience” are all relevant, at least in theory, to the procedural issue, as is the “commercial setting, purpose, and effect” of the contract. The law thus already calls for a contextualization of sorts. Yet in practice, factors like age and setting are difficult to weigh with any certainty, which is why courts tend to give them lip service and then focus on the text of the contract itself. Measuring boilerplate’s information costs, however, is much easier; evidence of word count, late-arriving terms, and related factors should be readily available to consumer and seller alike, and courts can evaluate that evidence in light of the emerging empirical findings on consumers’ cognitive limitations when dealing with boilerplate.

---


162. Korobkin would place the burden of proving nonsalience on the party that wishes to invalidate the term because all salient terms and some nonsalient terms are efficient. Korobkin, supra note 1, at 1280. Given the mounting evidence that boilerplate goes unread by nearly everyone, this should not be a particularly hard burden to bear. The key issue, then, is not who should bear the burden of proving procedural unconscionability; the question is
This simple expedient of adding vertical context to the analysis would result in more frequent findings of procedural unconscionability. But it would not upend boilerplate entirely. When information costs were low, as in the purchase of a straightforward product accompanied by a short contract, then the law would expect consumers to read the terms and register their individual preferences. Enforceability would merely scale with overall transactional complexity—and appropriately so.

Next, consider substantive unconscionability. Context is an important but overlooked factor here as well. A term may appear one-sided in isolation, but if the seemingly oppressed party got something more valuable in return (e.g., a significant price reduction), then invalidating the term would upset an efficient bargain. The inverse is also true: a term that appears reasonable on its face may in fact be oppressive if one party unilaterally imposed it on the other with no concomitant benefit. In short, by examining terms in isolation, rather than in the overall bargaining context, substantive unconscionability fails to weigh the costs and benefits of the transaction as a whole.

As Korobkin has pointed out, a better substantive unconscionability inquiry would focus not on whether the term at issue imposes a cost on the consumer, or on how high that cost is, but on whether imposing that cost on the consumer is efficient—whether it increases the parties’ joint wealth. As we saw in the lemons discussion above, unread boilerplate shifts costs to consumers as a matter of course, but such shifting is welfare-enhancing only when the consumer is the party better able to bear the risk. Therefore, to determine whether the term is “fair” to the consumer (as the substantive unconscionability analysis
purports to do), one must determine which party can most cheaply bear the risk being shifted. 166

Courts rarely make this determination, but there are exceptions. For example, the court in A&M Produce Co. v. FMC Corp. 167 evaluated a disclaimer of all warranties for agricultural equipment. Noting that “risk of loss is most appropriately borne by the party best able to prevent its occurrence,” the court ruled that the disclaimer was substantively unconscionable, determining that the seller could bear the risk of its own equipment’s failings more efficiently than the inexperienced buyer. 168

Granted, a cost–benefit analysis of this sort can be difficult. Even when a court knows that a boilerplate term is nonsalient, and has therefore already found procedural unconscionability, it will often be unclear whether a boilerplate term’s imposition of costs or risks on the consumer is efficient. 169 This means that the outcome will often depend on which party bears the burden of proof.

Korobkin suggests that the consumer should bear this burden (i.e., should have to prove that a boilerplate term is inefficient) because he is concerned that efficient terms might otherwise be thrown out—the danger of false positives. 170 Here, however, Korobkin and I disagree. Assigning burdens of proof generally is a function of two factors: (1) the probability of the event to be proved and (2) which party has better access to the

166. Id. at 1283–84.
168. A&M Produce Co., 186 Cal. Rptr. at 125 (“Rarely would the buyer be in a better position than the manufacturer-seller to evaluate the performance characteristics of a machine.”).
169. See Korobkin, supra note 1, at 1274–76 (revealing the complexities of such an inquiry through examples of arbitration clauses); cf. PUBLIC CITIZEN, THE ARBITRATION TRAP: HOW CREDIT CARD COMPANIES ENSNARE CONSUMERS 2 (2007) (arguing that credit-card arbitration agreements disfavor consumers).
170. See Korobkin, supra note 1, at 1285 (calling for an “implicit presumption against invalidating terms”).
relevant evidence.\textsuperscript{171} Regarding the first factor, we simply do not
know how frequently boilerplate contains inefficient terms. Korobkin is certainly right that there is a danger of false
positives if sellers bear the admittedly difficult burden of proving
efficiency.\textsuperscript{172} But there is a danger of false negatives if the
consumer bears the burden of proving inefficiency, and it is
impossible to know which danger is greater.\textsuperscript{173}

Given this uncertainty, the better basis for assignment of the
burden is to look to the second factor: which party has better
access to evidence about the matter to be proved?\textsuperscript{174} In the
boilerplate context, that party is the seller. After all, the seller is
the party that drafted the term in question, and of the two parties
it is the only one that fully internalizes the boilerplate's
information costs. So the seller presumably had a reason for
allocating costs and risks as it did. If the reason is that the
chosen allocation increases the parties' joint wealth, then the
seller can so demonstrate. If, however, the reason is that the
lemons dynamic forced it to inefficiently offload costs on the
consumer through nonsalient means, then forcing the seller to
prove efficiency not only leads to the correct result in the case at
hand, but it also helps solve the lemons problem going forward
and thus increases overall market efficiency.

In sum, two changes to the unconscionability doctrine can
help remedy the problems with boilerplate that this Article has
identified. First, as always, the consumer must prove procedural
unconscionability by showing that the disputed term was not

\textsuperscript{171} 2 \textsc{McCormick on Evidence} § 337 (John W. Strong gen'l ed., 5th ed.
1999); Bruce L. Hay and Kathryn E. Spier, \textit{Burdens of Proof in Civil Litigation:

\textsuperscript{172} See Korobkin, \textit{supra} note 1, at 1285 (arguing that false positives occur
because courts have difficulty weighing “industry-wide benefits and costs in the
context of an individual dispute”).

\textsuperscript{173} See \textsc{Posner, supra} note 8, at 827 (“[I]n the context of civil liability,
there is no reason to prefer one danger over the other.”).

\textsuperscript{174} \textsc{James Fleming Jr. et al., Civil Procedure} 344 (4th ed. 1992) (“The
burden of proof traditionally is placed on the party having the readier access to
knowledge about the fact in question.”); \textsc{McCormick, supra} note 171, § 337 (“A
doctrine often repeated by the courts is that where the facts with regard to an
issue lie peculiarly in the knowledge of a party, that party has the burden of
proving the issue.”).
salient—but this inquiry should take into account emerging empirical evidence and the full context of the overall transaction. Second, if the consumer satisfies this burden, then the burden on the issue of substantive unconscionability shifts to the seller, who must show that the term was efficient. Failure to do so means the term is unenforceable.

2. Forced Salience

Changing the unconscionability doctrine is an attractive approach to the boilerplate problem, not least because it works within an existing and familiar contract law construct. Its application, however, is likely to be infrequent because an unconscionability analysis arises only in those rare instances in which the parties have begun suing each other. In addition, its effect will be indirect because it encourages, rather than explicitly requires, boilerplate to be responsive to the market.

Is there a more direct and universal approach to boilerplate’s information costs problem—one that would fix the marketplace rather than supplant it? Consider again the two aspects of the information costs problem: the acquisition cost and the processing cost. Reducing acquisition cost is a matter of making the boilerplate more immediately accessible to consumers, such as through a disclosure requirement, so that they are better able to make a rational decision.175 Yet even when the acquisition cost is reduced to zero—i.e., even when the adhesive terms are readily

175. A number of commentators have accordingly suggested solving the boilerplate problem through early, mandatory disclosure. See, e.g., Braucher, Decision to Trust the Courts, supra note 29, at 755–56 (criticizing “the practice of holding back terms for mass-market products . . . even when it would be easy to provide them in advance” and arguing that “[t]he customer has a right to know what the product is before deciding whether to order it); Braucher, Delayed Disclosure, supra note 78, at 1860–62 (calling for “require[d] disclosure of key terms in software licenses,” and presenting data that only 12.5% of surveyed software companies disclosed their license agreement prior to requiring payment online); Robert A. Hillman & Maureen O’Rourke, Defending Disclosure in Software Licensing, 78 U. CHI. L. REV. 95, 104 (2011) (describing mandatory disclosure as a “safe harbor” for sellers).
available—consumers still face the processing cost, which in a complex transaction can pose just as big a problem.\footnote{176}

Requiring early disclosure is therefore not enough; it is a necessary but not sufficient part of the solution whenever the transaction as a whole is complex enough to impose considerable information processing costs. Any solution must also reduce processing costs. And until scientists develop a way to improve human cognition, processing costs are going to remain a function of the total amount of information presented.\footnote{177} There is, however, a regulatory approach that would lower both acquisition and processing costs, and it is revealed by examining the verticality of the computer purchase described above.

The computer purchases involved plenty of product features, both contractual and noncontractual, that were not presented as take-it-or-leave-it propositions. For example, the Dell website guided me through an array of options on such matters as the amount of memory, size of the hard drive, shipping terms, in-home service plan, and extended warranty—with one web page dedicated to each such feature. Indeed, the website all but forced me to choose between those options, with corresponding costs or savings depending on the choices I made.\footnote{178}

\footnote{176. See Ford Motor Credit Co. v. Milhollin, 444 U.S. 555, 568 (1980) ("\textit{Meaningful} disclosure does not mean \textit{more} disclosure. Rather, it describes a balance between \textquote{competing considerations of complete disclosure . . . and the need to avoid . . . [informational overload.]}\textquotenotemark{15} (alterations in original) (quoting S. Rep. No. 96-73, at 3 (1979), \textit{reprinted in} 1980 U.S.C.C.A.N. 280, 281)); Ben-Shahar \& Schneider, supra note 41, at 688 (noting that \textquote{incomplete disclosure leaves people ignorant, but complete disclosure creates crushing overload problems}). Indeed, relying on disclosure as the sole solution to the information costs problem actually makes things worse because it blinds courts to the processing problem and prompts them to enforce any disclosed term, no matter how nonsalient it may be. See Hillman, supra note 1, at 839 (noting that increased disclosure \textquote{may backfire . . . because it may not increase reading or shopping for terms or motivate businesses to draft reasonable ones, but instead, may make heretofore suspect terms more likely enforceable}.

177. See Korobkin, supra note 1, at 1246 (\textquote{[C]ontractual inefficiency results primarily from suboptimal information processing rather than from incomplete information}).

178. I was not literally forced to go through each and every feature because there was a \textquote{Finish[] Personalizing} option on each screen that would have skipped right to the order confirmation process—although even then I would have had to make choices with regard to terms such as shipping and payment.
Suppose that the law required Dell to force me to review all the contractual terms this way, rather than just the very few that it deemed worthy of inclusion up front? By making me confront each term before submitting any payment, Dell would be solving the information acquisition problem. And by not allowing me to skip the boilerplate, Dell would at least be increasing the odds that my attention would focus on the transaction’s contractual features, thus addressing the information processing problem.179

One can imagine more or less intrusive versions of this requirement. For example, the law could merely require consumers to click “I agree” with regard to each individual contract term. Or the law could require sellers to offer options with real consequences (in terms of costs or savings) and force consumers to select among them. Those approaches, however, would likely produce little real increase in salience; a consumer overloaded with information would just mindlessly click through, selecting options without really making choices.180 Information acquisition would occur, but information processing would not.

More effective would be to require the website to really force the consumer to read—e.g., by scrolling the text line by line, at a slow speed, or otherwise preventing the consumer from advancing to the next term until a set period of time had elapsed.181 This approach, which I call “forced salience,” helps solve the boilerplate problem in two ways. The most obvious is that by presenting each term up front and making the consumer confront

---

179. There is some empirical evidence that offering a consumer options can increase the rate of acceptance of a contract. E.g., Plaut & Bartlett, supra note 128, at 305.

180. See Hillman & Barakat, supra note 105, at 26 (speculating that “extra clicking would be cumbersome for little gain because consumers would simply click without digesting the disclaimer”); Hillman & O’Rourke, supra note 175, at 108 (considering a requirement that consumers click “I agree” but deciding that it “ultimately may not promote any additional reading anyway”); Korobkin, supra note 1, at 1246 (dismissing the “specific assent requirement”).

181. One might combine this approach with a “plain English” requirement, so that the now unavoidable boilerplate could actually be comprehended. Masson and Waldron had some success in increasing consumer comprehension through simplified diction and sentence structure. See Masson & Waldron, supra note 84, at 71–72, 75–76 (finding a significant increase in study participants’ ability to answer questions and correctly paraphrase prepositional phrases after reading the “plain-language” version of a contract).
it before moving on, it directly seeks to reduce the costs of information acquisition and information processing. If this results in compensatory consumer decisionmaking, great; the boilerplate problem is solved and the market properly registers the consumer’s preferences.

But would forced salience really result in significantly more reading of boilerplate? Perhaps not. Even if reading did not increase, however, this approach would address the boilerplate problem in a second, more subtle way. Consider that, as a consumer, your first reaction to forced salience is probably to scream in horror at the prospect of sitting through all of that boilerplate. It’s the digital equivalent of Judge Easterbrook’s “droning voice” that would “anesthetize rather than enlighten.” Yet this seeming liability is in fact an asset because, by adding new costs—i.e., time and frustration—to each contract term in the transaction, forced salience effectively imposes a tax on bloated boilerplate. Sellers know that consumers have limited time, limited cognitive abilities, and low thresholds for

182. Empirics suggest that reducing the length of click-wrap contracts increases comprehension and may increase readership. See Plaut & Bartlett, supra note 128, at 302–04 (finding 11% overall comprehension increase in all test contracts, but finding significant increase in readership time in only two of the six contracts). Reducing length also increases the likelihood that the consumer will reject the contract, suggesting that some expression of preferences due to increased comprehension might be occurring. See id. at 304–05 (suggesting that shorter contracts most likely allowed study participants to understand and analyze the terms). Note, however, that the contracts being tested by Plaut and Bartlett were associated with a free product. See id. at 301 (registering for and using online music website). Such contracts tend to have more readers than priced products. See Bakos et al., supra note 109, at 34 (pointing out that “consumers may fear that there is a ‘catch’ in products offered for free”).


184. Indeed, if one was interested only in this aspect of the forced salience approach, one could achieve it directly by taxing each word of boilerplate. (Thanks to Fred Yen for pointing this out.) One might also view forced salience as a classic formality, a “check against inconsiderate action,” which focuses parties’ attention on the legal consequences of their actions. See Baird, supra note 10, at 944 (“It is much cheaper to sign a document than to melt wax and use a signet ring. But this is the point.”); Lon L. Fuller, Consideration and Form, 41 Colum. L. Rev. 799, 800 (1941) (describing formalities’ “cautionary function”).
frustration. This means that sellers will respond to a forced salience requirement by reducing the overall volume of contract terms to the bare minimum, thus reducing information overload and increasing the chances that the remaining terms can truly be processed.\textsuperscript{185}

In other words, forced salience makes consumers confront the reality of boilerplate’s information costs, and the result is that sellers can no longer lade the transaction with as much boilerplate as they please. Instead, sellers will have to make a call about how badly they want each term to be part of the transaction. The inevitable paring of boilerplate to its essentials, and the resulting reduction in sheer volume of terms, would serve to make it easier for consumers to process those terms that survive. Even when forced salience fails, it succeeds.

Forced salience is an admittedly radical solution to the boilerplate problem. This Article has demonstrated, however, that the problem itself is radical and is radically underappreciated. Taking half-hearted measures will not restore integrity to so dysfunctional a market. Forcing the market to confront its dysfunction will.

\textit{VI. Conclusion}

Consumers do not encounter boilerplate in the abstract. They encounter it in the course of real-world transactions. And in the real world, the information costs of boilerplate loom large.

This Article has accordingly presented the results of a first-ever study of the information costs that boilerplate imposes on

\textsuperscript{185} One might also seek to reduce complexity to more manageable levels by regulating the noncontractual features of a transaction and their associated information costs. In a sense, the law already does so through product liability, false advertising law, and similar consumer-oriented regulation. But contractual features are a low-hanging fruit; it is comparatively simple to reduce complexity by declaring a contract term unenforceable—or at least that is the sort of regulation in which the law routinely engages. It is a much different proposition to reduce complexity through regulation of noncontractual product features (for example, requiring computer vendors to offer hard drives in only one size or a set number of USB ports). See Grether et al., supra note 10, at 289 ("[T]o limit the number of products that firms could sell or the number of attributes that products could have would require a costly, complex regulatory process.").
consumers as they make their way through an actual transaction—a fully contextualized, vertical experience. Its findings are that, in a typical computer purchase, those costs are so high as to be insuperable. They loom so large that real-world consumers have little choice but to disregard boilerplate entirely.

The solution to the information cost problem is neither to forbid all use of boilerplate nor to permit its indiscriminate use. Instead, the key lies in appreciating that, just as a contextualized analysis reveals the full scope of the problem, so can a contextualized solution solve it. The proper regulatory approaches thus involve changing existing contract doctrine to take into account the informational challenges that boilerplate actually presents, or incentivizing market actors to confront those costs and, by doing so, minimize them.
### VII: Appendix A—Summary of Results

Table 1: Summary of Results\(^{186}\)

<table>
<thead>
<tr>
<th>Seller</th>
<th>No. of Contracts</th>
<th>... at overall</th>
<th>... at purchase</th>
<th>... at computer startup</th>
<th>... at program startup</th>
<th>... per $</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acer</td>
<td>12</td>
<td>33,128</td>
<td>9,135</td>
<td>23,993</td>
<td>0</td>
<td>47.1</td>
</tr>
<tr>
<td>Dell</td>
<td>29</td>
<td>78,203</td>
<td>9,765</td>
<td>24,165</td>
<td>44,273</td>
<td>84.7</td>
</tr>
<tr>
<td>HP</td>
<td>25</td>
<td>79,340</td>
<td>0</td>
<td>24,328</td>
<td>55,012</td>
<td>103.4</td>
</tr>
<tr>
<td>Toshiba</td>
<td>27</td>
<td>96,641</td>
<td>18,678</td>
<td>34,744</td>
<td>84,477</td>
<td>131.5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>93</strong></td>
<td><strong>287,312</strong></td>
<td><strong>37,578</strong></td>
<td><strong>107,230</strong></td>
<td><strong>142,504</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Total Unique</strong></td>
<td><strong>56</strong></td>
<td><strong>161,767</strong></td>
<td><strong>39,065</strong> [24.1%]</td>
<td><strong>38,225</strong> [23.6%]</td>
<td><strong>84,477</strong> [52.2%]</td>
<td></td>
</tr>
<tr>
<td><strong>Raw Average</strong></td>
<td><strong>23.25</strong></td>
<td><strong>71,828</strong></td>
<td><strong>9,394</strong></td>
<td><strong>26,807</strong></td>
<td><strong>35,626</strong></td>
<td><strong>91.7</strong></td>
</tr>
<tr>
<td><strong>Weighted Average</strong></td>
<td><strong>24.75</strong></td>
<td><strong>74,897</strong></td>
<td><strong>7,698</strong> [10.3%]</td>
<td><strong>25,911</strong> [34.6%]</td>
<td><strong>41,286</strong> [55.1%]</td>
<td><strong>93.2</strong></td>
</tr>
</tbody>
</table>

186. Weighted averages are based on the market share set forth supra note 63.
# VIII: Appendix B—Boilerplate Breakdown

<table>
<thead>
<tr>
<th>Contract No.</th>
<th>Title</th>
<th>Word Count</th>
<th>When Encountered?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Standard Terms of Sale</td>
<td>3,005</td>
<td>At purchase</td>
</tr>
<tr>
<td>2</td>
<td>Limited Warranty Agreement</td>
<td>3,353</td>
<td>At purchase</td>
</tr>
<tr>
<td>3</td>
<td>Legal</td>
<td>665</td>
<td>Incorporated in 1</td>
</tr>
<tr>
<td>4</td>
<td>Products &amp; Services</td>
<td>136</td>
<td>Incorporated in 3</td>
</tr>
<tr>
<td>5</td>
<td>Privacy</td>
<td>1,972</td>
<td>Incorporated in 1</td>
</tr>
<tr>
<td>6</td>
<td>Contacts</td>
<td>4</td>
<td>Incorporated in 5</td>
</tr>
<tr>
<td>7</td>
<td>End User License Agreement</td>
<td>1,487</td>
<td>Computer startup</td>
</tr>
</tbody>
</table>

**Acer Subtotal:** 10,622 contracts

<table>
<thead>
<tr>
<th>Contract No.</th>
<th>Title</th>
<th>Word Count</th>
<th>When Encountered?</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>License Terms, Windows 7 Home</td>
<td>5,115</td>
<td>Computer startup</td>
</tr>
<tr>
<td></td>
<td>Premium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Windows 7 Privacy Statement</td>
<td>784</td>
<td>Incorporated in 8</td>
</tr>
<tr>
<td>10</td>
<td>Windows 7 Privacy Supplement</td>
<td>14,872</td>
<td>Incorporated in 8</td>
</tr>
<tr>
<td>11</td>
<td>.NET Framework Benchmark</td>
<td>429</td>
<td>Incorporated in 10</td>
</tr>
<tr>
<td></td>
<td>Testing Terms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Privacy Statement for the Microsoft Error Reporting Service</td>
<td>1,306</td>
<td>Incorporated in 10</td>
</tr>
</tbody>
</table>

**Microsoft Subtotal:** 22,506 contracts

**Acer Purchase Total:** 33,128 contracts

<table>
<thead>
<tr>
<th>Contract No.</th>
<th>Title</th>
<th>Word Count</th>
<th>When Encountered?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Terms and Conditions of Sale</td>
<td>2,339</td>
<td>At purchase</td>
</tr>
<tr>
<td>2</td>
<td>Return Policy</td>
<td>700</td>
<td>Incorporated in 1</td>
</tr>
<tr>
<td>3</td>
<td>Site Terms</td>
<td>1,077</td>
<td>Incorporated in 1</td>
</tr>
<tr>
<td>4</td>
<td>Warranties</td>
<td>5,649</td>
<td>Incorporated in 1</td>
</tr>
<tr>
<td>5</td>
<td>Notice for Dell End User Software License Agreement</td>
<td>78</td>
<td>Computer startup</td>
</tr>
<tr>
<td>6</td>
<td>Software License Agreement</td>
<td>1,581</td>
<td>Computer startup; incorporated in 7</td>
</tr>
<tr>
<td>7</td>
<td>DataSafe Online Terms and Conditions</td>
<td>2,799</td>
<td>DataSafe startup</td>
</tr>
<tr>
<td>8</td>
<td>Privacy Policy</td>
<td>1,665</td>
<td>Incorporated in 7</td>
</tr>
<tr>
<td>9</td>
<td>Change of Address/Request</td>
<td>256</td>
<td>Incorporated in 8</td>
</tr>
<tr>
<td>10</td>
<td>Security</td>
<td>1,051</td>
<td>Incorporated in 8</td>
</tr>
</tbody>
</table>

**Dell Subtotal:** 17,195 contracts

[Dell Purchase continued on next page]
<table>
<thead>
<tr>
<th>Contract No.</th>
<th>Title</th>
<th>Word Count</th>
<th>When Encountered?</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>License Terms, Windows 7 Home Premium</td>
<td>5,115</td>
<td>Computer startup</td>
</tr>
<tr>
<td>12</td>
<td>Windows 7 Privacy Statement</td>
<td>784</td>
<td>Incorporated in 11</td>
</tr>
<tr>
<td>13</td>
<td>Windows 7 Privacy Supplement</td>
<td>14,872</td>
<td>Incorporated in 11</td>
</tr>
<tr>
<td>14</td>
<td>.NET Framework Benchmark Testing Terms</td>
<td>429</td>
<td>Incorporated in 11</td>
</tr>
<tr>
<td>15</td>
<td>Privacy Statement for the Microsoft Error Reporting Service</td>
<td>1,306</td>
<td>Incorporated in 13</td>
</tr>
<tr>
<td>16</td>
<td>Software License Terms—Office 2010 Desktop Application Software</td>
<td>2,819</td>
<td>Office Starter startup</td>
</tr>
<tr>
<td>17</td>
<td>Genuine Microsoft Software Program Privacy Statement</td>
<td>911</td>
<td>Incorporated in 16</td>
</tr>
<tr>
<td>18</td>
<td>Privacy Statement for Office 2010 Content</td>
<td>4,994</td>
<td>Incorporated in 16</td>
</tr>
<tr>
<td>19</td>
<td>Use of Microsoft Copyrighted Content</td>
<td>1,628</td>
<td>Incorporated in 16 and 22</td>
</tr>
<tr>
<td>20</td>
<td>Privacy Supplement for Office Starter</td>
<td>455</td>
<td>Incorporated in 18</td>
</tr>
<tr>
<td>21</td>
<td>Trademark Guidelines</td>
<td>47</td>
<td>Incorporated in 19 and Internet Explorer startup;</td>
</tr>
<tr>
<td>22</td>
<td>Service Agreement</td>
<td>10,444</td>
<td>Incorporated in 19</td>
</tr>
<tr>
<td>23</td>
<td>Online Privacy Statement</td>
<td>4,614</td>
<td>Incorporated in 22</td>
</tr>
<tr>
<td>24</td>
<td>Anti-Spam Policy</td>
<td>464</td>
<td>Incorporated in 22</td>
</tr>
<tr>
<td>25</td>
<td>Authorized 3rd Party Software and Services</td>
<td>130</td>
<td>Incorporated in 22</td>
</tr>
<tr>
<td></td>
<td><strong>Microsoft Subtotal:</strong></td>
<td><strong>49,012</strong></td>
<td>15 contracts</td>
</tr>
<tr>
<td>26</td>
<td>Warranty Disclaimer and Software License Agreement</td>
<td>3,438</td>
<td>Reader startup</td>
</tr>
<tr>
<td>27</td>
<td>Online Privacy Policy</td>
<td>4,551</td>
<td>Incorporated in 26</td>
</tr>
<tr>
<td>28</td>
<td>Adobe.com Opt-Out page</td>
<td>201</td>
<td>Incorporated in 27</td>
</tr>
<tr>
<td></td>
<td><strong>Adobe Subtotal:</strong></td>
<td><strong>8,190</strong></td>
<td>3 contracts</td>
</tr>
<tr>
<td>29</td>
<td>End User License Agreement</td>
<td>3,806</td>
<td>Anti-Virus startup</td>
</tr>
<tr>
<td></td>
<td><strong>McAfee Subtotal:</strong></td>
<td><strong>3,806</strong></td>
<td>1 contract</td>
</tr>
<tr>
<td></td>
<td><strong>DELL PURCHASE Total:</strong></td>
<td><strong>78,203</strong></td>
<td>29 contracts</td>
</tr>
<tr>
<td>Contract No.</td>
<td>Title</td>
<td>Word Count</td>
<td>When Encountered?</td>
</tr>
<tr>
<td>-------------</td>
<td>-------</td>
<td>------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>1</td>
<td>End User License Agreement</td>
<td>1,822</td>
<td>Computer startup</td>
</tr>
<tr>
<td></td>
<td><strong>HP Subtotal:</strong></td>
<td><strong>1,822</strong></td>
<td><strong>1 contract</strong></td>
</tr>
<tr>
<td>2</td>
<td>License Terms, Windows 7 Home Premium</td>
<td>5,115</td>
<td>Computer startup</td>
</tr>
<tr>
<td>3</td>
<td>Windows 7 Privacy Statement</td>
<td>784</td>
<td>Incorporated in 2</td>
</tr>
<tr>
<td>4</td>
<td>Windows 7 Privacy Supplement</td>
<td>14,872</td>
<td>Incorporated in 2</td>
</tr>
<tr>
<td>5</td>
<td>.NET Framework Benchmark Testing Terms</td>
<td>429</td>
<td>Incorporated in 2</td>
</tr>
<tr>
<td>6</td>
<td>Privacy Statement for the Microsoft Error Reporting Service</td>
<td>1,306</td>
<td>Incorporated in 4</td>
</tr>
<tr>
<td>7</td>
<td>Software License Terms—Office 2010 Desktop Application Software</td>
<td>2,819</td>
<td>Office Starter startup</td>
</tr>
<tr>
<td>8</td>
<td>Genuine Microsoft Software Program Privacy Statement</td>
<td>911</td>
<td>Incorporated in 7</td>
</tr>
<tr>
<td>9</td>
<td>Privacy Statement for Office 2010</td>
<td>4,994</td>
<td>Incorporated in 7</td>
</tr>
<tr>
<td>10</td>
<td>Use of Microsoft Copyrighted Content</td>
<td>1,628</td>
<td>Incorporated in 7 and 13</td>
</tr>
<tr>
<td>11</td>
<td>Privacy Supplement for Office Starter</td>
<td>455</td>
<td>Incorporated in 9</td>
</tr>
<tr>
<td>12</td>
<td>Trademark Guidelines</td>
<td>47</td>
<td>Incorporated in 10</td>
</tr>
<tr>
<td>13</td>
<td>Service Agreement</td>
<td>10,444</td>
<td>Incorporated in 10</td>
</tr>
<tr>
<td>14</td>
<td>Online Privacy Statement</td>
<td>4,614</td>
<td>Incorporated in 13</td>
</tr>
<tr>
<td>15</td>
<td>Anti-Spam Policy</td>
<td>464</td>
<td>Incorporated in 13</td>
</tr>
<tr>
<td>16</td>
<td>Authorized 3rd Party Software and Services</td>
<td>130</td>
<td>Incorporated in 13</td>
</tr>
<tr>
<td></td>
<td><strong>Microsoft Subtotal:</strong></td>
<td><strong>49,012</strong></td>
<td><strong>15 contracts</strong></td>
</tr>
<tr>
<td>17</td>
<td>Software License Agreement</td>
<td>9,217</td>
<td>Premiere Elements startup</td>
</tr>
<tr>
<td>18</td>
<td>Privacy Policy</td>
<td>4,551</td>
<td>Incorporated in 17</td>
</tr>
<tr>
<td>19</td>
<td>Adobe.com Opt-Out page</td>
<td>201</td>
<td>Incorporated in 18</td>
</tr>
<tr>
<td>20</td>
<td>General Terms of Use</td>
<td>8,130</td>
<td>Incorporated in 17</td>
</tr>
<tr>
<td>21</td>
<td>Trademark Information</td>
<td>1,300</td>
<td>Incorporated in 20</td>
</tr>
<tr>
<td>22</td>
<td>Trademark Guidelines</td>
<td>3,001</td>
<td>Incorporated in 22</td>
</tr>
<tr>
<td>23</td>
<td>Trademark Database for General Distribution</td>
<td>1,484</td>
<td>Incorporated in 22</td>
</tr>
<tr>
<td>24</td>
<td>Fonts</td>
<td>101</td>
<td>Incorporated in 17</td>
</tr>
<tr>
<td>25</td>
<td>Flash Runtime Distribution License</td>
<td>521</td>
<td>Incorporated in 17</td>
</tr>
<tr>
<td></td>
<td><strong>Adobe Subtotal:</strong></td>
<td><strong>28,506</strong></td>
<td><strong>9 contracts</strong></td>
</tr>
<tr>
<td></td>
<td><strong>HP PURCHASE TOTAL</strong></td>
<td><strong>79,340</strong></td>
<td><strong>25 contracts</strong></td>
</tr>
</tbody>
</table>
### Toshiba Purchase

<table>
<thead>
<tr>
<th>Contract No.</th>
<th>Title</th>
<th>Word Count</th>
<th>When Encountered?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Terms and Conditions of Sale</td>
<td>2,175</td>
<td>At purchase</td>
</tr>
<tr>
<td>2</td>
<td>NAF Code of Procedure</td>
<td>16,503</td>
<td>Incorporated in 1</td>
</tr>
<tr>
<td>3</td>
<td>Notices</td>
<td>405</td>
<td>Shrinkwrap on computer</td>
</tr>
<tr>
<td>4</td>
<td>Limited Warranty</td>
<td>2,522</td>
<td>Incorporated in 3</td>
</tr>
<tr>
<td>5</td>
<td>JAMS Streamlined Arbitration Rules &amp; Procedures</td>
<td>7,282</td>
<td>Incorporated in 4</td>
</tr>
<tr>
<td>6</td>
<td>End User License Agreement</td>
<td>2,029</td>
<td>Computer startup</td>
</tr>
<tr>
<td>7</td>
<td>Eco Utility</td>
<td>847</td>
<td>Eco Utility Startup</td>
</tr>
</tbody>
</table>

**Toshiba Subtotal:** 31,763 7 contracts

<table>
<thead>
<tr>
<th>Contract No.</th>
<th>Title</th>
<th>Word Count</th>
<th>When Encountered?</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>License Terms, Windows 7 Home Premium</td>
<td>5,115</td>
<td>Computer startup</td>
</tr>
<tr>
<td>9</td>
<td>Windows 7 Privacy Statement</td>
<td>784</td>
<td>Incorporated in 8</td>
</tr>
<tr>
<td>10</td>
<td>Windows 7 Privacy Supplement</td>
<td>14,872</td>
<td>Incorporated in 8</td>
</tr>
<tr>
<td>11</td>
<td>.NET Framework Benchmark Testing Terms</td>
<td>429</td>
<td>Incorporated in 8</td>
</tr>
<tr>
<td>12</td>
<td>Privacy Statement for the Microsoft Error Reporting Service</td>
<td>1,306</td>
<td>Incorporated in 10</td>
</tr>
<tr>
<td>13</td>
<td>Software License Terms—Office 2010 Desktop Application Software</td>
<td>2,819</td>
<td>Office Starter startup</td>
</tr>
<tr>
<td>14</td>
<td>Genuine Microsoft Software Program Privacy Statement</td>
<td>911</td>
<td>Incorporated in 13</td>
</tr>
<tr>
<td>15</td>
<td>Privacy Statement for Office 2010</td>
<td>4,994</td>
<td>Incorporated in 13</td>
</tr>
<tr>
<td>16</td>
<td>Use of Microsoft Copyrighted Content</td>
<td>1,628</td>
<td>Incorporated in 13 and 19</td>
</tr>
<tr>
<td>17</td>
<td>Privacy Supplement for Office Starter</td>
<td>455</td>
<td>Incorporated in 15</td>
</tr>
<tr>
<td>18</td>
<td>Trademark Guidelines</td>
<td>47</td>
<td>Incorporated in 16</td>
</tr>
<tr>
<td>19</td>
<td>Service Agreement</td>
<td>10,444</td>
<td>Live Essentials startup</td>
</tr>
<tr>
<td>20</td>
<td>Online Privacy Statement</td>
<td>4,814</td>
<td>Incorporated in 19</td>
</tr>
<tr>
<td>21</td>
<td>Anti-Spam Policy</td>
<td>464</td>
<td>Incorporation in 19</td>
</tr>
<tr>
<td>22</td>
<td>Authorized 3rd Party Software and Services</td>
<td>130</td>
<td>Incorporated in 19</td>
</tr>
</tbody>
</table>

**Microsoft Subtotal:** 49,012 15 contracts

<table>
<thead>
<tr>
<th>Contract No.</th>
<th>Title</th>
<th>Word Count</th>
<th>When Encountered?</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>Chrome Terms of Service</td>
<td>6,552</td>
<td>Chrome startup</td>
</tr>
<tr>
<td>24</td>
<td>Google Privacy Policy</td>
<td>2,240</td>
<td>Incorporated in 26</td>
</tr>
<tr>
<td>25</td>
<td>Key Terms</td>
<td>598</td>
<td>Incorporated in 27</td>
</tr>
<tr>
<td>26</td>
<td>Google Chrome Privacy Notice</td>
<td>3,225</td>
<td>Incorporated in 26</td>
</tr>
</tbody>
</table>

**Google Subtotal:** 12,615 4 contracts

[Toshiba Purchase continued on next page]
<table>
<thead>
<tr>
<th>Contract No.</th>
<th>Title</th>
<th>Word Count</th>
<th>When Encountered?</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>Norton License Agreement</td>
<td>3,251</td>
<td>Internet Security startup</td>
</tr>
</tbody>
</table>

**Norton Subtotal:** 3,251 1 contract

**TOSHIBA PURCHASE TOTAL:** 96,641 27 contracts
When Certainty Dissolves into Probability: A Legal Vision of Toxic Causation for the Post-Genomic Era

Steve C. Gold*

Abstract

Proof of causation in toxic torts has presented persistent problems for the legal system, because the probabilities that science can know fit poorly with the demands for particularistic proof imposed by the law’s deterministic model of causation. Some scholars have hoped that genomic and molecular information will at last provide scientific certainty—definitive, individualized proof of toxic causation.

This Article argues that the opposite is true. Scientific research will increasingly elucidate the ways in which environmental exposures and human genes interact to produce disease, but this deeper knowledge will extend rather than resolve the problem of causal indeterminacy in toxic torts. Genomic and molecular understanding, instead of sounding the death knell for proposals to reform toxic tort causation law, will strengthen the argument for those reforms.

This Article proposes a probabilistic causal contribution model to replace the model of deterministic causation in toxic torts, building on earlier scholarly proposals and the creativity of

* Associate Professor of Law, Rutgers School of Law-Newark. This research benefited from the insights of Rutgers-Newark colloquium participants and was funded in part by the school’s Dean’s Research Fund. Thanks for suggestions are particularly due to Stuart Deutsch, Michael Green, Howard Latin, John Leubsdorf, Brandon Paradise, Twila Perry, James Pope, Sabrina Safrin, Sandy Steel, and Katherine Van Tassel. Thanks also to research assistants Tara Elliott, Kevin Geary, Kady Keen, Jason LaMarca, Karina Levitian, Brian Matthews, Olivia Pomann, and Caitlin Stephens; librarian Susan Lyons; and volunteer motivators and editors Jenny Aley and Sylviane Gold.
a handful of courts. The Article explores how the model would work and argues that it is superior to present doctrine when assessed against the goals of the tort system.

Table of Contents

I. Introduction ..................................................................... 239

II. The Persistent Problem: Toxic Injury and the
    Traditional Deterministic Model of Causation .......... 241
    A. The Traditional, Deterministic Model of
        Causation................................................................. 241
    B. The Lack of Fit in Toxic Injury Cases............. 244

III. Solutions Illusory and Real: The False—and
    True—Promise of Genomics ............................................ 252
    A. A Bit of Genomics ...................................................... 253
        1. Genes and Disease ............................................... 254
        2. Toxic Exposure and Disease—and Genes ........... 257
            3. The Tools of Toxicogenomics ............................... 259
    B. The False Promise ..................................................... 265
    C. The True Promise and Its Implications .......... 276

IV. An Alternative Vision: Probabilistic Causal
    Contribution .................................................................... 281
    A. A Metaphor ............................................................ 282
    B. Contributing Factor Causation ................................. 284
    C. Risk as Cause, Not as Harm ............................. 297
    D. The Question of Substantiality .......................... 304
    E. Discounted Recoveries ............................................. 307
    F. Boundaries ................................................................. 319

V. Probabilistic Causal Contribution and Tort Goals .......... 322
    A. Deterrence Goals: Is Probabilistic Precision
        Probable? .............................................................. 322
    B. Corrective Justice Goals: Is Causation
        Different? ................................................................. 330
    C. Truth-seeking Goals: Wrong All the Time? .......... 336
    D. Administration of Justice Concerns: Too Much
        Information? ............................................................. 337

VI. Conclusion........................................................................ 338
I. Introduction

Forty years after a jury first found that several asbestos product manufacturers each had caused someone's asbestosis and pleural mesothelioma, sick people continue to come to court seeking compensation for illnesses that they allege were caused by exposure to various toxins. The defendants sued for making, selling, using, or disposing of the allegedly toxic products continue to protest that something else—dumb luck, bad genes, someone else’s similar product, some entirely different toxin, or mother nature—is what really made the plaintiffs sick, and, more important, that the plaintiffs cannot prove otherwise.

Proof of toxic tort claims conforms poorly to the traditional deterministic legal model of but-for causation, because toxic injuries almost never involve an observable chain of physical events allowing easy inference of a causal relation between a particular defendant’s conduct and a particular plaintiff’s harm. Courts turn to science to replace causal intuition, but a disjunction remains between the probabilities that science can know and the determined result that the law wants proven. The resulting problem has produced hundreds of court opinions and numerous calls for doctrinal reforms in recognition of the difficulties that toxic causation presents. Yet, despite decades of thoughtful jurisprudence and scholarship, the core legal conception of toxic causation has hardly changed at all.

Science has changed, however, quickly and at an accelerating pace. The ability to peer into the genome heralds new insights into human susceptibility to toxic substances. Molecular technologies provide glimpses of previously inaccessible toxic mechanisms. Computing power allows this research to be conducted on a scale and at a rate never before possible.

1. Borel v. Fibreboard Paper Prods. Corp., 493 F.2d 1076, 1094 (5th Cir. 1973) (holding that “the jury could find that each defendant was the cause in fact of some injury to Borel” although “it is impossible, as a practical matter, to determine with absolute certainty which particular exposure to asbestos dust resulted in injury”); see PAUL BRODEUR, OUTRAGEOUS MISCONDUCT 39–70 (1985) (describing the Borel trial); Joseph Sanders, Risky Business: Causation in Asbestos Cancer Cases (and Beyond?), in PERSPECTIVES ON CAUSATION 11, 15–17 (Richard Goldberg ed., 2011) (describing the indivisible injury causation rationale of the Borel decision).
The law looks to this new science, hoping that its increased resolving power will at last build a bridge from probability to certainty. The dream is that genomic knowledge will lead to deterministic proof of the existence or absence of causal mechanism in individual cases. That “holy grail” of particularistic evidence would at last eliminate much of the indeterminacy that has plagued proof of specific causation in toxic tort claims; it also could fatally undermine arguments for exceptional judicial treatment of toxic tort claims, restoring to tort law a universally applicable deterministic model of but-for causation.

This Article argues that the predominant reality is more likely to be exactly the opposite: at the highest magnifications, certainty will dissolve into probability. Scientific research will increasingly elucidate the ways in which environmental exposures and human genes interact to produce disease, but for the legal system, this deeper knowledge will extend rather than resolve the causal indeterminacy problem in toxic torts—and will therefore increase the justification for doctrinal adjustments to address that problem.

Part II describes the traditional deterministic model of causation and how it fails in many toxic tort cases. Part III explains why it is likely that scientific success will exacerbate rather than solve that failure. Part IV proposes an alternative vision based on a probabilistic causal contribution model of causation. Part V assesses the proposal against goals of tort law.

2. See Joseph Sanders, Apportionment and Proof in Toxic Injury Cases, 10 Kan. J.L. & Pub. Pol'y 200, 202 (2000) (“Specific causation evidence seems to be the holy grail of toxic torts.”). “Specific causation” refers to causation of the individual plaintiff’s case of disease, in contrast to the exposure-disease link that is often referred to as “general causation.” See infra notes 20–25 and accompanying text.

II. The Persistent Problem: Toxic Injury and the Traditional, Deterministic Model of Causation

A. The Traditional, Deterministic Model of Causation

Some notion of causation inheres in the most elementary formulation of a tort claim. The claim that “defendant wronged plaintiff” implies a relation between the parties that is distinguishable from plaintiff’s relation with persons other than the defendant. To further say that the wrong was more than abstract—that “defendant harmed plaintiff”—is to connect defendant’s conduct and plaintiff’s injury via a causal relation that is distinguishable from the vagaries of a universe in which people just get injured sometimes. The concept is so intuitive that torts casebooks typically apply it well before they get around to expressing it. Yet articulating the meaning of causation is sufficiently difficult that the effort has long occupied philosophers, scientists, legal scholars, and judges.

Part of that effort involved disentangling the factual from the normative in everyday assessments of causal responsibility. We have come a long way from the legal fiction that any given event has a single objective cause, but the frank acknowledgment that the doctrine of proximate cause invokes policy choices did not extinguish the law’s reliance on the idea that events have

---

4. See, e.g., MARC A. FRANKLIN ET AL., TORT LAW AND ALTERNATIVES: CASES AND MATERIALS 333 (9th ed. 2011) (introducing cause-in-fact by noting that “[f]rom the outset of the book we have implicitly accepted the notion that a defendant . . . should not have to compensate an injured plaintiff unless the plaintiff’s injury is causally connected to the defendant’s negligent conduct”); cf. Mark Kelman, The Necessary Myth of Objective Causation in Liberal Political Theory, 63 CHI.-KENT L. REV. 579, 587–88 (1987) (arguing that dissociating causation from standard of care is irrational under efficiency theories).


determinate factual causes.\textsuperscript{7} Having shooed policy preferences into the back room of proximate cause, standard tort doctrine continues to insist that cause-in-fact reflects an objectively knowable reality—albeit knowable, ex post, only by counterfactual inference.

Causation-in-fact has traditionally been proven by persuading the fact finder that the familiar but-for test is satisfied: “Conduct is a factual cause of harm when the harm would not have occurred absent the conduct.”\textsuperscript{8} The first two restatements of the law of torts framed the required causal connection as a requirement that a cause be “a substantial factor in bringing about the harm.”\textsuperscript{9} They explained “substantial factor” as a limiting additional requirement distinguishing, from among all but-for causes, those that were legally cognizable.\textsuperscript{10} The Third Restatement of Torts, opining that this use of “substantial factor” led to inchoate results, abandoned the term entirely,\textsuperscript{11} completing

\begin{itemize}
\item \textsuperscript{7} Id. at 479–85 (describing the legal realist assault on the doctrine of objective causation and the rise of the distinction between “actual” and “legal” cause).
\item \textsuperscript{8}Restatement (Third) of Torts: Liab. for Physical & Emotional Harm § 26 (2010).
\item \textsuperscript{9}Restatement of Torts § 431 (1934); Restatement (Second) of Torts § 431 (1965).
\item \textsuperscript{10}Restatement of Torts § 431 cmt. a (1934) (“T]hat the harm would not have occurred had the actor not been negligent . . . is necessary but . . . is not of itself sufficient [to establish that negligence was a legal cause of harm].”); Restatement (Second) of Torts § 431 cmt. a (1965) (same); id. § 433 (listing considerations important to determining whether an act is a “substantial factor”); see Mahoney v. Beatman, 147 A. 762, 766–67 (Conn. 1929) (adopting this view of “substantial factor”); but see Restatement (Third) of Torts: Liab. for Physical & Emotional Harm § 26 cmt. j & reporters’ note (2010) (describing varied applications of “substantial factor”). Before publication of the First Restatement, some courts used “substantial factor” to justify the intuitively correct outcome in cases of multiple sufficient causes, in which a literal application of the “but-for” test would exonerate both causes. See Anderson v. Minneapolis, St. Paul & Sault Ste. Marie Ry. Co., 179 N.W. 45, 49 (Minn. 1920) (classic “two fires” case); Restatement (Second) of Torts § 432(2) (1965) (stating that in such circumstances “the actor’s negligence may be found to be a substantial factor” in bringing about harm); Joseph Sanders et al., The Insufficiency of the “Substantial Factor” Test for Causation, 73 Mo. L. Rev. 399, 416–17 (2008) (describing this use of “substantial factor”).
\item \textsuperscript{11}See Restatement (Third) of Torts: Liab. for Physical & Emotional Harm § 26 cmt. j, reporters’ note (2010) (noting that the term “substantial factor” seemed “to be doing scope-of-liability (proximate-cause) duty” in addition
(as asserted in the commentary) the conceptual separation of normatively driven proximate cause from (presumably objective) cause-in-fact.\textsuperscript{12}

The Third Restatement modeled a but-for factual cause as any necessary element of a set of causes that, together, are sufficient to bring about the result.\textsuperscript{13} This reflects an entirely deterministic conception of cause.\textsuperscript{14} Its philosophic underpinning is the theory that causal laws connect fully specified sets of causes to effects in an invariant, rather than probabilistic, way. Take away a true cause, in this model, and the effect always disappears, unless the effect is overdetermined—that is, unless some other sufficient causal set exists that does not include this cause. Leave the cause in existence, and the effect is inevitable, unless some other necessary element of the causal set is removed. It may be impossible ever to fully specify a likely infinite set of causal antecedents, but this model of causation posits that we can nevertheless make generalizations about the element of a set of antecedents that is of interest, e.g., the one alleged to be the cause of harm in a given tort case.\textsuperscript{15} Such a generalization reflects an experience-based imputation of a causal role ex ante that is applied ex post to make inferences about the antecedent’s determined causal role in the actual event.

\textsuperscript{12} Id. § 29 cmt. g. Although this separation is familiar to every American law student who has completed the first-year torts course, a distinguished British jurist recently questioned the use of cause-in-fact “as a kind of filter which you have to get through in order to qualify for the final round of being selected as legal causation.” Lord Hoffmann, \textit{Causation, in PERSPECTIVES ON CAUSATION}, supra note 1, at 3, 4–6 (2011).

\textsuperscript{13} \textbf{RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM} § 26 cmt. g (2010). As the Third Restatement explains, this model is generally consistent with the “NESS” (Necessary Element of a Sufficient Set) test. \textit{Id.} § 26 cmt. c, reporters’ note; see also Richard W. Wright, \textit{Causation in Tort Law}, 73 \textit{CALIF. L. REV.} 1735, 1788–1803 (1985) (explaining the NESS test). As Wright noted, however, but-for cause taken alone is not equivalent to the NESS test. Wright, \textit{supra} note 5, at 1021–22 nn.109–11.

\textsuperscript{14} \textit{See} \textbf{RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM} § 26 cmt. c (2010).

\textsuperscript{15} \textit{See} Wright, \textit{supra} note 5, at 1045.
This model also treats causation as a dichotomous, yes-or-no proposition: it either existed in a given case or it did not. Such treatment is not unique to causation, of course. In an adversarial truth-seeking system designed to choose between competing versions of events, many elements of claims or defenses are treated that way. But in today’s doctrine some of these elements, dichotomous as they are, nevertheless are commonly aligned on a continuous quantitative scale rather than being limited to quantum values of zero or one. Many jurisdictions, for example, ask fact-finders to apportion fault or, more broadly, “responsibility.”16 Causation, however, has proven mostly resistant to treatment as a continuous variable.17

B. The Lack of Fit in Toxic Injury Cases

A toxic tort plaintiff claims that exposure to some chemical, radiological, or biological agent caused a disease. A fundamental difficulty in proving such a claim is that exposure and disease usually do not correlate perfectly: some people get sick without exposure, and some people receive exposure without getting sick.18 In marked contrast to traumatic injury cases, the disease process itself is unobserved and unobservable as it occurs, and inscrutable afterward.19 What evidence then will permit an


17. See id. § 26 cmt. a (limiting apportionment based on causation to “separately caused” damages); id. § 26 cmt. c (prescribing that “the factfinder divides divisible damages into their indivisible component parts”); id. § 26 cmt. f (listing circumstances in which damages can be divided by causation); see also, e.g., James v. Bessemer Processing Co., 714 A.2d 898, 909 (N.J. 1998) (noting that despite legal scholarship advocating reforms, “courts have been resistant to novel models of causation”); Waste Mgmt., Inc. v. S. Cent. Bell Tel. Co., 15 S.W.3d 425, 433 (Tenn. Ct. App. 1997) (“Causation in fact is an all-or-nothing proposition.”).


inference that a plaintiff’s illness would not have occurred but for the exposure, to establish deterministically modeled causation-in-fact?

Courts initially demand proof of “general causation,” asking whether the exposure in question is ever a sine qua non for the plaintiff’s disease, or whether the existence of cases of disease after exposure is merely coincidental. To prove general causation, plaintiffs have attempted to rely on toxicology studies (either in vitro or in vivo in experimental animals), evidence of toxins’ biological mechanisms of action, simple inference from exposure dose and chronology, and epidemiologic studies (which investigate the relative risk of disease that is associated with exposure in samples of human populations). Many courts have insisted on the primacy of epidemiology for proof of general causation, in part because of situations in which many claims were brought despite large bodies of powerful epidemiologic studies that failed to find any association between the exposure and the disease the exposure allegedly caused.

Even strong proof of general causation may be unavailing, however. The existence of disease without exposure leads courts to further demand proof of “specific causation,” asking whether this plaintiff’s case of disease is one that would not have occurred but for the exposure. The difficulty lies in determining what if any additional proof a plaintiff must, may, or can adduce to prove specific causation.

---

20. General causation is sometimes expressed as whether an agent “is capable of” causing a disease, but more than theoretical capability is implied. Restatement (Third) of Torts: Liab. for Physical & Emotional Harm § 28 cmt. c (2010); Henricksen v. ConocoPhillips Co., 605 F. Supp. 2d 1142, 1158 (E.D. Wash. 2009); cf. Joseph Sanders, Proof of Individual Causation in Toxic Tort and Forensic Cases, 75 Brook. L. Rev. 1367, 1375 (2010) (“The general causation question is whether a substance... has been shown to harm any individuals.”).

21. See, e.g., In re “Agent Orange” Prod. Liab. Litig., 611 F. Supp. 1223, 1231 (E.D.N.Y. 1985), aff’d, 818 F.2d 187 (2d Cir. 1987) (“[E]pidemiological studies... are the only useful studies having any bearing on causation.”); Merrell Dow Pharm., Inc. v. Havner, 953 S.W.2d 706, 717–18 (Tex. 1997) (requiring a statistically significant epidemiologic result of relative risk greater than two).

22. E.g., Henricksen, 605 F. Supp. 2d at 1155.
For some courts, epidemiology may—or must—serve that function, but only if the epidemiologic studies find that exposure more than doubles the risk of disease, which would imply that exposure accounted for more than one-half of the incidence of the disease. From this statistic based on the incidence of disease in exposed and unexposed populations, courts have reasoned that it is "more likely than not" that an individual plaintiff's disease is a case associated with the exposure. Some commentators view this "doubling+" requirement as too stringent. James Robins and Sander Greenland, for example, showed that if the inference of general causation is accepted, then in most circumstances the epidemiologic relative risk is a lower bound of the probability that a given individual who has both exposure and illness is a case of "true" causation.

On the other hand, others have argued that because relative risk is undeniably a property of samples and populations, but specific causation is a property of an individual case, epidemiologic data cannot support inferences about specific causation at all. Some courts have effectively demanded that a plaintiff produce particularistic evidence—proof of some fact that will distinguish a plaintiff's case of disease from the cases that would have occurred even absent exposure. To try to satisfy

---


25. James M. Robins & Sander Greenland, The Probability of Causation Under a Stochastic Model for Individual Risk, 45 BIOMETRICS 1125, 1129–32 (1989); see also RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 28 cmt. c(4) (2010) ("[A]ny judicial requirement that plaintiffs must show a threshold increase in risk or a doubling in incidence in a group study in order to satisfy the burden of proof of specific causation is usually inappropriate.").

26. See Wright, supra note 5, at 1054 ("[P]articularistic evidence is necessary for causal explanation . . . ."); Wright, supra note 24, at 1318 (arguing that epidemiologic data provide no evidence that toxic exposure "actually caused" harm in individual case); Sienkiewicz v. Greif, [2011] UKSC 10, 157 (Lord Rodger), 170 (Lady Hale), 190–91 (Lord Mance).

27. See, e.g., Henricksen v. ConocoPhillips Co., 605 F. Supp. 2d 1142, 1162–
such demands, some plaintiffs have turned to medical experts who try to rule out other plausible causes, a method called differential diagnosis or differential etiology. Some plaintiffs have tried to rely on epidemiologic studies coupled with analysis of the plaintiff’s individual characteristics that put the plaintiff at a particularly heightened risk.28

The interaction of population-derived relative risk data and proof of specific causation under the traditional deterministic and dichotomous causation model received unusually explicit attention in two cases decided by the United States Court of Appeals for the Tenth Circuit. In each case, a group of plaintiffs alleged tortious exposure to radiation from uranium mining and milling.

In June v. Union Carbide Corp.,29 twenty-seven former residents of Uravan, Colorado, alleged that radioactive contamination from uranium and vanadium mining and milling operations in the company town had caused their thyroid disease or non-thyroid cancer. General causation was not in dispute: the defendants had to concede that ionizing radiation in general is capable of causing cancer, and that products of uranium and vanadium radioactive decay, such as iodine-131, cause thyroid cancer in particular. In an attempt to prove specific causation, plaintiffs relied on experts who estimated each plaintiff’s radiation dose and opined that such exposure was a “substantial factor” contributing to their various diseases.30 The defendants argued, and the district and appellate courts agreed, that such testimony could not suffice to prove causation because it did not assert that any plaintiff’s illness would not have occurred but for exposure to the Uravan radiation.

The Tenth Circuit panel analyzed the description of factual causation in both the Second Restatement of Torts, which used

---

28. See, e.g., Estate of George, 993 A.2d at 384 (Reiber, C.J., dissenting) (discussing the argument that a firefighter’s individual risk of contracting lymphoma from inhaled chemicals was higher than the average risk reported in epidemiologic studies).

29. June v. Union Carbide Corp, 577 F.3d 1234 (10th Cir. 2009).

30. Id. at 1237.
the “substantial factor” formulation, and the Third Restatement, which consciously deleted reference to “substantial factor.” The court concluded that despite the difference in language, the two restatements equivalently established but-for as the sine qua non for factual causation. The restatements admitted only one exception—situations involving multiple sufficient causes—and the court reasoned that even in such situations, proof of causation requires a showing that the alleged cause would have been a but-for cause in the hypothetical absence of the other sufficient cause(s).

In support of that reasoning, the court cited several illustrations given in the Third Restatement, including “the one most pertinent to the case before us,” a hypothetical product liability case in which a plaintiff whose daughter has a birth defect alleges that a drug manufacturer failed to warn of the drug’s teratogenicity. The plaintiff claims the drug caused the birth defect; the manufacturer claims that an independent genetic condition caused it. The Tenth Circuit noted that according to the Restatement, the plaintiff must show that the drug, acting alone, would have caused the birth defect, not merely that the drug could have caused the defect. This formulation echoes the distinction typically made between general causation and specific causation.

The majority held that the plaintiffs had “failed to present to the [district] court evidence, or even an argument” that radiation released by defendants constituted “either a but-for cause of . . .

31. Id. at 1241.
32. Id. at 1244 (citing Restatement (Third) of Torts: Liab. for Physical & Emotional Harm § 27 cmt. e, illus. 2 (2010)).
33. Id. Actually, the illustration was not particularly pertinent to June, because it illustrated the rule for multiple sufficient causes and therefore necessarily assumed that both the parents and the drug company presented sufficient evidence that either the genetic condition or the drug alone would have caused the injury. See Restatement (Third) of Torts: Liab. for Physical & Emotional Harm § 27 (2010). The real question lurking in June is whether demanding such evidence can be realistic in a case like June, and how to respond if it is not.
34. Many courts have said that general causation requires proof that an exposure can cause the plaintiff’s condition, and specific causation requires proof that the plaintiff’s exposure did cause the plaintiff’s individual case. E.g., Kilpatrick v. Breg, Inc., 613 F.3d 1329, 1334 n.4 (11th Cir. 2010).
or . . . a necessary component of a causal set that would probably have caused any of the plaintiffs’ medical conditions. Judge Holloway dissented in part, arguing that for those plaintiffs with thyroid disease, the testimony of one of plaintiff’s experts created a material issue of disputed fact with respect to but-for causation. Judge Holloway did not question the majority’s explication of “but-for” as the sole test for causation.

In Wilcox v. Homestake Mining Co., three plaintiffs alleged that their cancers had been caused by radiation exposure from another uranium mill in New Mexico. They submitted evidence similar to the plaintiffs’ evidence in June, but argued that New Mexico law, unlike Colorado law as interpreted in June, would permit a finding of causation based on the “substantial factor” test. The district judge dismissed the action on summary judgment. A Tenth Circuit panel that included none of the June judges unanimously affirmed—but in three separate opinions.

Judges Lucero and McKay agreed—often verbatim—with June’s embrace of mechanistic “but-for” causation as the only valid interpretation of causation-in-fact. They held that New Mexico law allowed for substantial factor causation in lieu of but-for causation only in two exceptional cases. Regarding the first exception, multiple sufficient causes, the court followed June, holding that “[o]nly a substance that would have actually (that is, probably) caused the cancer can be a factual cause without being a but-for cause.” Regarding the second, alternative liability, the court saw no “basis for alternative liability where only one

35. June, 577 F.3d at 1245 (10th Cir. 2009).

36. Id. at 1253–54 (Holloway, J., dissenting). The majority refused on procedural grounds to consider the plaintiffs’ arguments that persuaded Judge Holloway. Id.

37. Id. at 1252–54.

38. Wilcox v. Homestake Mining Co., 619 F.3d 1165 (10th Cir. 2010).

39. See id. at 1170–71 (Lucero, J., concurring) (describing testimony of plaintiffs’ experts, including one who testified in June, regarding the “assigned share” of risk and the lack of other known significant risk factors).

40. Id. at 1167.

41. Id. at 1168 (quoting June v. Union Carbide Corp., 577 F.3d 1234, 1243 (10th Cir. 2009)).
potential wrongdoer has been identified and the injury may simply have resulted from natural causes.”

Judge Lucero nevertheless disagreed with Judge McKay’s—and the June majority’s—application of the but-for standard. Judge Lucero embraced Judge Holloway’s partial dissent in June, but found the factual presentation in Wilcox insufficient to prove but-for causation, by contrast with June.

Of the six judges on the two panels, only Judge Holmes seemed willing to contemplate the possibility that “the but-for standard of causation” ought not apply to cases like June and Wilcox. But Judge Holmes did not decide that any other standard should or would apply, opining instead that plaintiffs had waived any argument for a different standard. Judge Holmes agreed that the plaintiffs had failed to satisfy the but-for test.

It is easy to understand the uneasiness reflected in the opinions of Judges Holloway, Lucero, and Holmes. In applying a deterministic model of causation, the Tenth Circuit effectively reinforced the specific causation requirement for toxic tort plaintiffs. The court demanded proof that each individual plaintiff would not have developed cancer absent exposure to the radiation released by the defendants in the respective cases. How could any plaintiff possibly do this?

The majorities, in very similar terms, denied that they had set the plaintiffs an impossible task. They rooted their

42. Id. at 1167. The court stated that alternative liability applies when “two or more defendants engage in simultaneous or nearly identical negligent acts but only one of these acts causes the injury complained of, thus making it difficult or impossible for the plaintiff to prove which defendant caused the harm” and distinguished asbestos cases in which “it is clear the plaintiff’s injury was caused by asbestos” but the defendant that caused the harm could not be specified. Id.

43. Id. at 1170–73 (Lucero, J., concurring).

44. Id. at 1173 (Holmes, J., concurring).

45. Id.

46. Id.

47. See id. at 1169 (“[W]e are not persuaded this requirement is so insurmountable . . . .”); June v. Union Carbide Corp., 577 F.3d 1234, 1246 (10th Cir. 2009) (stating that in holding that plaintiffs did not attempt to prove but-for causation, the court was “not being hypertechnical”).
explanations in the preponderance of the evidence standard: although a plaintiff must show that she or he “actually” would not have become ill absent a defendant’s tortious exposure of the plaintiff to a toxic substance, the plaintiff need not prove this to an “absolute certainty.” "Actually" actually means “probably” or “likely,” the majorities stated.

This question-begging explanation still leaves one to ponder what type of evidence might have satisfied the plaintiff’s burden. How could a plaintiff show it was “likely” that the defendants’ radiation releases “actually” caused his or her disease? Neither majority opinion says. Their discussions of but-for causation, however, are so mechanistic as to give the impression that only particularistic evidence—something about plaintiff’s disease that betrayed its origin, like the ballistics marks on a bullet—could suffice. Perhaps—the opinions don’t say—a sufficiently large population-based relative risk estimate derived from an epidemiologic study could also have satisfied the majorities. Neither of those options was available to the plaintiffs.

The plaintiffs in each of these cases did prove, however, that the defendants exposed them to radiation. The defendants did not deny that such exposure increased the likelihood that a person would develop the diseases from which the plaintiffs suffered. That increased risk is known partly because of an understanding of what ionizing radiation does to DNA, but also because of the observation that more disease is found in populations exposed to additional radiation than in populations not exposed. Nevertheless, even though the defendants had exposed the

48. June, 577 F.3d at 1243.
49. Wilcox v. Homestake Mining Co., 619 F.3d 1165, 1169 (10th Cir. 2010).
50. June, 577 F.3d at 1243.
51. See Wilcox, 619 F.3d at 1167 (characterizing plaintiffs’ position as an argument that causation could be proven “without regard to whether the injuries would likely have occurred in the absence of defendant’s action”); id. at 1169 (stating that a toxic tort plaintiff must demonstrate but-for causation “only to a reasonable degree of medical probability—not a certainty”).
plaintiffs to known risk factors for their diseases, no plaintiff could obtain any recovery.

*Ex hypothesi*, the defendants’ conduct was tortious—perhaps because reasonable precautions were available that would have avoided the haphazard broadcast of radionuclides in the community. The plaintiffs’ evidence fairly supported the inference that someone got sick who, if the mine and mill operators had behaved differently, would not have been sick. Yet the deterministic concept of causation demands that this plaintiff show that he or she is the “someone” whose disease would not have occurred absent the defendants’ wrongful conduct. In many cases, as in *June* and *Wilcox*, this will be impossible, resulting in denial of all recoveries against defendants who nevertheless wrongfully made people sick.

**III. Solutions Illusory and Real: The False—and True—Promise of Genomics**

The deterministic model of causation has performed poorly when confronted with the mechanistic opacity of toxic torts. That opacity is as intolerable to science and medicine as it is inconvenient for law. But science has developed tools to attack it. In just two decades, researchers have sequenced the human genome, developed information processing capacity to conduct statistical analysis of very large numbers of genetic variations, invented techniques to rapidly assay the effects of toxic substance exposure on highly variable genetic material, and learned new ways to detect disease-related changes at the sub-cellular or molecular level.

Observing all this biomedical research from the law school across campus, legal scholars understandably envision that one day the molecular impacts of carcinogens, mutagens, and other toxins will be as easy to detect and attribute as the impacts from wayward cricket balls, deformed automobile parts, and misdirected scalpels. If so, “an increasing number of people will have the tools necessary to prove *sine qua non* causation in toxic
tort litigation," making it easier for deserving plaintiffs to prevail and, by implication, allowing courts accurately to dismiss claims when those tools are available but do not provide the desired proof. At last, cause-in-fact in toxic torts will be as certain as in any other tort. Or will it?

A. A Bit of Genomics

The seed that grew into the Human Genome Project, interestingly, germinated in the soil of a toxic tort problem. In the early 1980s, veterans and their families demanded compensation for injuries allegedly resulting from service members’ deliberate experimental exposure to radiation by the government during the Cold War; in response, Congress commissioned a study of the feasibility of detecting low-level damage to DNA caused by environmental exposures. The researchers given the task realized that they would have to “scan many genomes’ worth of DNA to pick up what were sure to be small alterations in the mutation rate.” A scientist at the Department of Energy conceived “a more manageable goal, that of sequencing the entire human genome once rather than many times,” and the seed was planted. It would be about a half-decade before the first federal appropriation for the Human Genome Project sprouted.

53. Andrew R. Klein, Causation and Uncertainty: Making Connections in a Time of Change, 49 JURIMETRICS J. 5, 35 (2008); see also Gary E. Marchant, Genetic Data in Toxic Tort Litigation, 14 J.L. & POL’Y 7, 8 (2006) (“New genetic methods and data have the potential to fill some of the scientific uncertainties and data gaps in toxic tort litigation, thus making toxic tort litigation more accurate and fair.”); Sanders, supra note 20, at 1399 n.132 (“[T]he field of toxicogenomics offers the long run possibility that we will be able to ascertain causation at the level of the individual case, radically changing all specific causation testimony.”).

54. CHRISTOPHER WILLS, EXONS, INTRONS, AND TALKING GENES 72 (1991). According to Wills, other significant compensation claims of the era, including those of Vietnam veterans exposed to Agent Orange and of people living near the damaged Three Mile Island nuclear plant, also contributed to the political impetus for the study. Id. at 71.

55. Id. at 74.

56. Id. at 75.

57. Id. at 81.
Remarkably, it took only a bit more than twice that time to harvest the crop: a highly accurate sequential listing, coordinated with physical locations on the chromosomes, of about three billion nucleotide base pairs that constitute human nuclear DNA. The success of the Human Genome Project spawned a slew of other “omics” projects pursuing comprehensive indices of classes of human biochemicals. Some of this research promises better understanding of the genetic and environmental components of disease etiology, and thus may be directly relevant to the causation issue in toxic torts.

1. Genes and Disease

A gene is a segment of DNA that encodes information, which may be translated into a sequence of amino acids. The amino acid chain can then be formed into all or part of a protein. Biologists did not need to sequence the entire human genome to understand that a person’s inherited genotype could affect the phenotype—

58. Press Release, Nat’l Human Genome Research Inst., International Consortium Completes Human Genome Project (Apr. 14, 2003), available at http://www.genome.gov/11006929; see also Int’l Human Genome Sequencing Consortium, Initial Sequencing and Analysis of the Human Genome, 409 NATURE 860, 873, 875 (2001) (estimating the total number of bases in the genome). A nucleotide or “base” is one of the four different molecules that bond together to form a strand of DNA. ISRAEL ROSENFIELD ET AL., DNA 73 (2011). In the double helix of the DNA molecule, each base is “paired” with its complementary nucleotide. Id.


60. Two examples are the Environmental Genome Project and HuGENet (Human Genome Epidemiology Network). See generally Samuel H. Wilson & Kenneth Olden, The Environmental Genome Project: Phase I and Beyond, 4 MOLECULAR INTERVENTIONS 147 (2004) (describing Environmental Genome Project); Wei Yu et al., HugeWatch: Tracking Trends and Patterns of Published Studies of Genetic Association and Human Genome Epidemiology in Near-real Time, 16 EUR. J. HUMAN GENETICS 1155 (2008) (describing HuGENet and related projects).
the person’s observable traits—in important health-related ways. The most straightforward situations involve an inherited allele, a specific variation in a given gene, which changes the coded protein in a way that produces a particular disease or condition.61

The story of a genetic disease can be considerably more complicated, however, even for diseases caused by changes in a single protein coded by a single gene. Cystic fibrosis is a case in point. In 1989, researchers announced the sequencing of “the cystic fibrosis gene,” which codes for a protein involved in transporting molecules across cell membranes.62 A person who inherits from each parent a copy of a variant allele that results in the omission of one amino acid from the protein will have cystic fibrosis.63 But since that discovery, researchers have identified more than 1,600 different mutations that can produce cystic fibrosis.64 The mutations, located in various parts of the gene, affect the protein in various ways that produce disease of varying severity.65

Genes also relate to diseases in even more complex ways. For example, people with a particular variation of a gene called APOE are dramatically more likely to develop Alzheimer’s disease than people without that allele: ten times more likely if they inherited two copies of the variant gene, four times if they inherited one copy.66 But other genes also appear to play roles in

61. For example, alteration of one DNA nucleotide in a gene that codes for hemoglobin yields a substitution of one amino acid in that protein, which changes the molecule’s shape—producing the sickle-cell phenotype—and reduces its ability to carry oxygen. Muin J. Khoury & Janice S. Dorman, Genetic Disease, in MOLECULAR EPIDEMIOLOGY 365, 370 (Paul A. Schulte & Frederica P. Perera eds., 1993).


63. Id. at 188; Steven M. Rowe et al., A Breath of Fresh Air, SCI. AM., Aug. 2011, at 69, 71.

64. Rowe et al., supra note 63, at 72; see also WILLS, supra note 54, at 210 (stating that in initial experiments fewer than 70% of cystic fibrosis patients’ genes tested displayed the variation that deleted one amino acid).

65. Rowe et al., supra note 63, at 72; see also WILLS, supra note 54, at 212–13 (“[M]utations are found all over the gene.”).

Alzheimer's disease, albeit much smaller roles. There is no one “Alzheimer’s disease gene.”

The Alzheimer’s pattern is not unusual. Susceptibility alleles, rare variants of the \textit{BRCA1} and \textit{BRCA2} genes, significantly increase a woman’s risk of developing breast cancer. But genome-wide association studies—which examine large numbers of genes in persons with and without a disease to determine if particular DNA variations are statistically associated with higher disease incidence—identified half a dozen new susceptibility alleles; these occur relatively frequently in the population, but individually confer small increments of risk ranging from seven to twenty-six percent. Thus, the degree to which disease or other phenotypic change results from variation in a gene, which biologists call “penetrance,” varies from gene to gene. Sometimes, the risk conferred by a variation in a

\footnotesize{67. Pooled data from numerous studies revealed associations between five genes and the incidence of Alzheimer's disease; each of these alleles is associated with an increase in risk of 10 to 15%. Id.}

\footnotesize{68. See, e.g., Athina Christopoulou & John Spiliotis, The Role of \textit{BRCA1} and \textit{BRCA2} in Hereditary Breast Cancer, 10 GENE THERAPY & MOLECULAR BIOLOGY 95, 96 (2006), http://www.gtmb.org/pages/Vol10A/PDF/09. Christop&Spiliot,95-100.pdf (noting that women carrying mutations have a 40% to 85% lifetime risk, versus 12.5% in general population).}


\footnotesize{70. See Eleanor Raffan & Robert K. Semple, Next Generation Sequencing—Implications for Clinical Practice, 99 BRIT. MED. BULL. 53, 62–63 (2011) (defining variable penetrance as “the variable tendency of genetic mutations to translate into clinical disease;” noting that “every genome sequenced” includes numerous alleles implicated in disease); see also Elisabeth A. Lloyd, Normality and Variation: The Human Genome Project and the Ideal Human Type, in \textit{Are Genes Us?}, supra note 5, at 99, 100 (noting that most identified genetic differences are risk factors for, or provide a vulnerability to, development of a specific disease, rather than determinants of disease); Evgeny N. Imyanitov et al., Searching for Cancer-Associated Gene Polymorphisms: Promises and Obstacles, 204 CANCER LETTERS 3, 8 (2004) (“Low penetrance is often exemplified as something like 1.5-fold risk elevation.”); see generally SUZANNE H. REUBEN, PRESIDENT'S CANCER PANEL, REDUCING ENVIRONMENTAL CANCER RISK: WHAT WE CAN DO NOW 1 (2010) (“Single-gene inherited cancer syndromes are believed to account for less than 5 percent of malignancies in the United States.”).}
particular gene may depend on a person’s genotype at other genes or on the action of other biochemical constituents.71

2. Toxic Exposure and Disease—and Genes

Biologists also did not need to sequence the entire human genome to understand that different people may respond to the same dose of the same toxin in different ways. Even for acutely lethal poisons, although there may be a dose above which no person (or research animal) can survive, toxicologists typically determine what dose is lethal for half of the organisms. Because the toxin and dose are the same for all the exposed organisms, something—a variation from organism to organism, a variation in the environment from organism to organism, or pure random chance—must explain why the toxin is lethal to some but not to others.72

More extreme variations typify the chronic or latent toxicity ordinarily involved in toxic tort cases. Consider tobacco smoke,

71. See Raffan & Semple, supra note 70, at 63 (stating that variable penetrance “is usually ascribed either to environmental factors or to unspecified ‘genetic modifiers’”); Marc A. Schaub et al., Linking Disease Associations with Regulatory Information in the Human Genome, 22 GENOME RES. 1748, 1748 (2012) (noting that for the vast majority of DNA variations found to be associated with disease, “it is likely that the underlying mechanism linking them to the phenotype is regulatory” rather than a change in coded protein sequence). The effects of the breast cancer susceptibility alleles, for example, are modified by other genes that alter the body’s response to environmental insults such as radiation. Logan C. Walker et al., Use of Expression Data and the CGEMS Genome-Wide Breast Cancer Association Study to Identify Genes that May Modify Risk in BRCA1/2 Mutation Carriers, 112 BREAST CANCER RES. & TREATMENT 229, 229, 233 (2008). Cystic fibrosis is another example: some people have symptoms worse than would be suggested by the change in the protein coded by the “cystic fibrosis gene” because a group of several hundred helper proteins recognize the altered protein as defective and destroy it before it can be put into position to do its job. Rowe et al., supra note 63, at 72. See generally James R. Griesemer, Tools for Talking: Human Nature, Weismannism, and the Interpretation of Genetic Information, in ARE GENES US?, supra note 5, at 69, 82 (observing that genes “cause” inherited somatic characteristics but “[t]he body is a cause in inheritance . . . . Proteins and other cell components are causally responsible for the events of cell division and DNA replication”).

which is almost universally acknowledged to cause lung cancer.\textsuperscript{73} Many people with lung cancer seem to have acquired the disease by smoking.\textsuperscript{74} But some people smoke to a ripe old age,\textsuperscript{75} and only a small fraction of even the heaviest smokers ever contract lung cancer.\textsuperscript{76} On the other hand, though lung cancer was very rare before smoking became common, some people who never smoked did—and do—get lung cancer.\textsuperscript{77} What explains the differences?

The search for explanations commonly distinguishes extrinsic from intrinsic potential sources of inter-individual difference: environment versus genetics.\textsuperscript{78} Perhaps exposure to something other than tobacco smoke—say, diesel exhaust—causes lung cancer in some non-smokers, while another factor—say, a diet rich in just the right antioxidants—inhibits

\textsuperscript{73} As recently as 2005, however, Britain’s Imperial Tobacco successfully defended a tort claim by disputing that it had been “scientifically established” that cigarette smoking causes lung cancer. See McTear v. Imperial Tobacco Ltd., [2005] C.S.O.H. 69, 5.694–5.695, 6.149–6.17, 9.9 (stating that the burden of proof on causation was not satisfied); id. 9.10 (finding that epidemiologic data failed to prove individual causation).

\textsuperscript{74} See Graham G. Giles & Peter Boyle, Smoking and Lung Cancer, in TOBACCO AND PUBLIC HEALTH 485, 486 (Peter Boyle et al., eds. 2004) (estimating that 83% to 92% of lung cancer deaths in five developed countries are attributable to smoking).

\textsuperscript{75} See, e.g., George Davey Smith, Epidemiology, Epigenetics and the “Gloomy Prospect”: Embracing Randomness in Population Health Research and Practice, 40 INT’L J. EPIDEMIOLOGY 537, 537–38 (2011) (describing a healthy centenarian believed to have smoked 170,000 cigarettes).


\textsuperscript{77} See Asta Scesnaite et al., Similar DNA Methylation Pattern in Lung Tumours from Smokers and Never-smokers with Second-hand Tobacco Smoke Exposure, 27 MUTAGENESIS 423, 423 (2012) (acknowledging that 10% to 15% of lung cancer deaths occur in people who never smoked). Exposure to environmental tobacco smoke may explain some of those cases. See Giles & Boyle, supra note 74, at 495.

\textsuperscript{78} See generally Jennifer E. Below, Factors that Impact Susceptibility to Fiber-Induced Health Effects, 14 J. TOXICOLOGY ENVTL. HEALTH 246 (2011) (noting that only five percent of people exposed to asbestos develop mesothelioma and discussing genetic, nutritional, and other environmental factors that may affect susceptibility to the disease); Davey Smith, supra note 75, at 539 (describing studies that partition contributions to health outcomes from genetics, shared environment, and non-shared environment).
carcinogenesis in the lungs of some smokers.\textsuperscript{79} Or perhaps a fortunate genetic endowment protects some lifelong smokers from lung cancer, while a genetic mischance induces lung cancer in some non-smokers.\textsuperscript{80} Both environmental and genetic differences between individuals appear responsible for at least some of the variation in individuals' responses to toxic exposures.\textsuperscript{81} For the most part, it has been impossible (or at least impractical) to identify, quantify, and tease apart these possibilities using the investigatory tools of toxicology, environmental epidemiology, conventional biochemistry, and classical genetics.\textsuperscript{82}

3. The Tools of Toxicogenomics

Technologies developed during and after the Human Genome Project provide new tools of vast potential power. DNA microarrays and even more powerful next-generation sequencing technologies allow assays of many genes at once and make it

\begin{itemize}
\item \textsuperscript{79} See, e.g., Cassidy, supra note 76, at 2 (identifying dietary and other environmental factors that affect lung cancer risk); Giles & Boyle, supra note 74, at 487 (listing other environmental causes of lung cancer); Elizabeth A. Ward et al., Research Recommendations for Selected IARC-Classified Agents, 118 ENVTL. HEALTH PERSP. 1355, 1358 (2010) (describing meta-analyses showing diesel exhaust associated with slightly elevated risk of lung cancer (relative risks 1.33 (95% confidence interval 1.24–1.44) and 1.47 (1.29–1.67)).
\item \textsuperscript{80} See Ping Zhan et al., CYP1A1 MspI and exon7 Gene Polymorphisms And Lung Cancer Risk: An Updated Meta-Analysis and Review, 30 J. EXPERIMENTAL CLINICAL CANCER RES. 99, *15 (2011) (explaining that variations of two genes are associated with increased risk of lung cancer, and some variations particularly increase risk in smokers).
\item \textsuperscript{81} For example, the authors of a study that observed “[m]arked inter-individual variation in response to the same level of exposure” to air pollution concluded “that susceptibility might be due to genetic factors.” Shuang Wang et al., Methods for Detecting Interactions Between Genetic Polymorphisms and Prenatal Environment With a Mother-Child Design, 34 GENETIC EPIDEMIOLOGY 125, 131 (2010).
\item \textsuperscript{82} See David Altshuler et al., Genetic Mapping in Human Disease, 322 SCI. 881, 881 (2008) (describing limitations of classical genetics); Michael D. Waters et al., Toxicogenomic Approach for Assessing Toxicant-Related Disease, 544 MUTATION RES. 415, 415 (2003) (describing the potential of toxicogenomics to address previously “intractable” problems).
\end{itemize}
practicable to sequence an individual’s entire genome. The young but rapidly developing science of toxicogenomics marries the experimental techniques of toxicology to the analytical techniques of genomics. Using laboratory animals or cells or tissues cultured in vitro, researchers can expose genetic material containing many variations of many genes to a suspected toxin and observe any variations in response, or they can compare exposed and non-exposed genetic material and observe any differences. New technologies allow detection not only of variations (polymorphisms) in the DNA sequence of particular genes, but also rearrangements and other variations of the structure of genes along a chromosome.

Alternatively, researchers can study the genes of samples of actual human populations and determine whether observable genetic differences are associated with differential exposure to toxins or with differential toxic effects among those exposed.

83. Raffan & Semple, supra note 70, at 55; see also, e.g., S. Le Scouarnec & S.M. Gribble, Characterising Chromosome Rearrangements: Recent Technical Advances in Molecular Cytogenetics, 108 HEREDITY 75, 79 (2012) (describing how the Human Genome Project took more than a decade, but a whole human genome now can be sequenced in a few days); D.A. Wheeler et al., The Complete Genome of an Individual by Massively Parallel DNA Sequencing, 452 NATURE 872, 872 (2008) (reporting first sequencing of a human’s genome using “massively parallel” next-generation technology).

84. See generally Gerald T. Ankley et al., Toxicogenomics in Regulatory Ecotoxicology, 40 ENVTL. SCI. TECH. 4055, 4056 (2006) (documenting increasing numbers of toxicogenomics publications in each of the next five years following the first such publication, in 2000); Syril Pettit et al., Current and Future Applications of Toxicogenomics: Results Summary of a Survey from the HESI Genomics State of Science Subcommittee, 118 ENVTL. HEALTH PERSP. 992, 995 (2010) (reporting that toxicogenomics has improved understanding of biological mechanisms of toxicity but biological understanding of toxicogenomic data remains limited).

85. See Jamie A. Grodsky, Genetics and Environmental Law: Redefining Public Health, 93 CALIF. L. REV. 171, 190 (2005) (“DNA microarrays . . . permit thousands of genes to be monitored simultaneously to determine whether they have been activated or deactivated as a result of chemical exposure.”); see also Le Scouarnec & Gribble, supra note 83, at 76 (explaining how microarrays work).

86. Le Scouarnec & Gribble, supra note 83, at 75.

87. See e.g., Peter Soderkvist & Olav Axelson, On the Use of Molecular Biology Data in Occupational and Environmental Epidemiology, 37 J. OCCUPATIONAL ENVTL. MED. 84, 84–86 (1995) (describing study models); Wang et al, supra note 81, at 131 (discussing results of study of gene-environment
This extension of epidemiologic methods marks the field of molecular epidemiology.88

A goal of toxicogenomic and molecular epidemiologic studies is to identify biomarkers—biochemical characteristics that reveal a toxic relation of interest. A biomarker of exposure is an observable change that occurs with exposure but is otherwise absent. A biomarker of effect is an observable, medically significant, harmful change that occurs with exposure but is otherwise absent. A biomarker of susceptibility, by contrast, is an observable genetic variation that alters the extent to which an exposure causes toxic harm.89 Researchers have begun to find markers of each type.

For example, as stem cells divide and differentiate into various types of white blood cells, metabolites of benzene interact with DNA to cause errors during the copying of chromosomes.90 Research has revealed that blood cells of individuals occupationally exposed to benzene are more likely to have particular chromosomal aberrations associated with the development of certain forms of a group of cancers called acute myelogenous leukemia (AML).91 The observation suggests a biological mechanism of benzene carcinogenicity, tending to confirm classical epidemiologic studies that detected an association between exposure to benzene and incidence of AML. It also could provide a marker of benzene exposure or effect.92
Other studies have found potential biomarkers of susceptibility. For instance, epidemiologic studies strongly linked mesothelioma incidence to asbestos exposure, but estimates of relative risks varied. In a region of Turkey with high environmental exposure to a form of asbestos, researchers found that mesothelioma was not randomly distributed but rather clustered in certain families, suggesting that susceptibility to asbestos-induced mesothelioma has a genetic component. Although the full picture is far from clear, one study of asbestos-exposed people found higher susceptibility in those with certain variations in a gene that codes for an enzyme that catalyzes production of antioxidant molecules.

Another example involves the suspected link between tobacco smoke and breast cancer, which eluded conventional epidemiologic investigation. Genomic investigations observed that variations in a gene that codes a carcinogen-neutralizing enzyme dramatically influenced the breast cancer danger from smoking. Women whose genes coded for the most protective form selectively affects certain chromosomes).

---


94. R.M. Rudd, *Malignant Mesothelioma*, 93 BRIT. MED. BULL. 105, 108 (2010); see also Below, supra note 78, at 254 (suggesting that genome-wide association studies (GWAS) would likely reveal genetic susceptibility factors). Other factors, including random sampling error, could account for some of the variation in relative risk results as well. See Michael D. Green, *Second Thoughts About Apportionment in Asbestos Litigation*, 37 SW. L. REV. 531, 538 (2008) (describing gaps in knowledge of mesothelioma causation); see also McDonald, supra note 93, at 852–55 (discussing possible differences in forms of asbestos that study subjects were exposed to).

95. Aki Murakami et al., *Heme Oxygenase-1 Promoter Polymorphism is Associated with Risk of Malignant Mesothelioma*, 1 LUNG 333 (2012); see also Joseph R. Testa et al., *Germline BAP1 Mutations Predispose to Malignant Mesothelioma*, 43 NATURE GENETICS 1022, 1025 (2011) (reporting that mutations in a gene that codes for a tumor suppressor protein may be associated with heightened risk of several cancers even without asbestos exposure, but predominantly mesothelioma if asbestos exposure is present).

of the enzyme had no increased risk of breast cancer even if they smoked, but women smokers with less protective forms of the gene were eight times more likely to get breast cancer than were women with the same genotype who did not smoke. The gene variations thus could act as biomarkers of susceptibility.

These examples show the potential of toxicogenomics and molecular epidemiology but should not obscure the difficulties facing these sciences. Biologists did not need, or should not have needed, to sequence the entire human genome to understand that sequencing the genome alone would not tell the entire story of disease or of toxicity. The links between genes, toxic substances, and disease form a web far more complex than previously imagined.

Genome-scale studies test vast numbers of genes and alleles—so many that random chance would produce large numbers of coincidental associations between gene and disease (or between gene and toxicity). Although genomic researchers abide by a statistical significance convention orders of magnitude more stringent than the typical 95% level used in most scientific


99. So far, at least 16 million single nucleotide polymorphisms (SNPs)—variations that change one DNA base pair—have been identified. Simon N. Stacey et al., A Germline Variant in the TP53 Polyadenylation Signal Confers Cancer Susceptibility, 43 NATURE GENETICS 1098, 1099 (2011). That number is rapidly increasing. See Altshuler et al., supra note 82, at 833 (discussing how the number of known SNPs rose from 1.4 million in late 1990s to more than 10 million by 2008). At least two million variations involving small insertions or deletions of DNA have been found as well. Julienne M. Mullaney et al, Small Insertions and Deletions (InDels) in Human Genomes, 19 HUM. MOLECULAR GENETICS R131, R133 (2010). Some researchers suggest that these and other structural variations in DNA may be even more important than SNPs. Yingrui Li et al., Structural Variation in Two Human Genomes Mapped at Single-Nucleotide Resolution by Whole Genome de Novo Assembly, 29 NATURE BIOTECHNOLOGY 723, 723, 728 (2011).
investigation, false positive results are easy to obtain and difficult
to exclude.\textsuperscript{100} Associations must be investigated to assess their
biological reality.\textsuperscript{101} And the effect of a given allele on toxic
susceptibility may not be a fixed value; it may vary depending on
the alleles found at other genes or on other environmental
factors.\textsuperscript{102}

Furthermore, a gene’s base sequence is not its only
biologically relevant feature. All of a person’s cells have the same
DNA, except for mutations acquired in individual cells during
life,\textsuperscript{103} but they do many different things.\textsuperscript{104} Whether a gene is
expressed or not, and to what extent—that is, whether its DNA is
actively being transcribed into RNA and translated into
assembled proteins—is critical to the proper functioning of cells,
tissues, and organs.\textsuperscript{105} Deviations from normal gene expression

\textsuperscript{100} See Ian P.M. Tomlinson et al., Investigation of the Effects of DNA
Repair Gene Polymorphisms on the Risk of Colorectal Cancer, 27 MUTAGENESIS
219, 219 (2012) (describing a study that failed to support previously reported
gene-disease associations); Samuel P. Dickson et al., Rare Variants Create

\textsuperscript{101} Spitz & Bondy, supra note 88, at 130; see also Inês Barroso, Non-
Coding but Functional, 489 NATURE 54, 54 (2012) (“Association is not
causality, and identifying those variants which are causally linked to a given
disease or trait . . . has been difficult.”).

\textsuperscript{102} See Camille Limoges, Errare Humanum Est: Do Genetic Errors Have a
Future, in ARE GENES US?, supra note 5, at 113, 121 (noting that even in
Mendelian disorders caused by changes in a single gene, phenotype is often
subject to modification by other genes and environmental factors); Christopher
A. Maxwell et al., Genetic Interactions: The Missing Links for a Better
Understanding of Cancer Susceptibility, Progression and Treatment, 7
pdf/1476-4598-7-4.pdf (“The detection of these [gene-gene] interactions will be
invaluable to our understanding of cancer risk.”); Leonardo A. Pinto et al.,
Impact of Genetics in Childhood Asthma, 84 JORNAL DE PEDIATRIA S68, S68, S72
(2008) (noting that asthma involves many genes and results from interaction of
genetic and environmental factors); Christopher P. Wild, Environmental
Exposure Measurement in Cancer Epidemiology, 24 MUTAGENESIS 117, 117
(2009) (“Precise contribution of specific risk factors and their interaction, both
with each other and with genotype, continues to be difficult to elucidate.”).

\textsuperscript{103} See generally WILLS, supra note 54, at 91 (explaining that about one
DNA base per 500,000 is mutated during a lifetime). The gametes (sperm and
egg cells), of course, do not contain the same DNA as other body cells. Id.

\textsuperscript{104} See ROSENFELD ET AL., supra note 58, at 108–13.

\textsuperscript{105} Id. at 95–101, 108–13.
are important to many disease processes. Gene expression is controlled in part by regulatory genes and in part by an epigenome of other biochemical constituents that provides instructions that influence gene activity. The epigenome is easily altered by environmental factors, sometimes in ways that echo long after the exposure and sometimes in ways that (unlike a change to a gene) cannot be detected later.

All this complexity has implications for the genetic study of disease as well as for the study of toxicogenomics. Both the power and the limits of toxicogenomics will affect proof of specific causation in toxic torts.

B. The False Promise

A number of commentators have eagerly anticipated the day when science exposes the presumed deterministic mechanism of toxic causation for all to see. The hope is that biomarkers will

106. See generally Altshuler et al., supra note 82, at 881.

107. Raffan & Semple, supra note 70, at 14; see also Joseph R. Ecker, Serving Up a Genome Feast, 489 Nature 52, 52 (2012) (“T]he space between genes is filled with enhancers (regulatory DNA elements), promoters (the sites at which DNA’s transcription into RNA is initiated) and numerous previously overlooked regions that encode RNA transcripts that are not translated into proteins but might have regulatory roles.”).

108. Constituents of the epigenome include methyl groups that may be attached to DNA bases, histone proteins associated with DNA, non-coding DNA sequences, RNA sequences that interfere with or otherwise regulate the translation of DNA into protein, and other aspects of a cell’s biochemistry. See Mark A. Rothstein et al., The Ghost in Our Genes: Legal and Ethical Implications of Epigenetics, 19 Health Matrix 1, 5–8 (2009) (describing composition and function of epigenome); Ward et al., supra note 79, at 1356 (describing epigenetic effects).

109. See Kathryn Z. Guyton et al., Improving Prediction of Chemical Carcinogenicity by Considering Multiple Mechanisms and Applying Toxicogenomic Approaches, 681 Mutation Res. 230, 235 (2009) (noting that epigenetic effects are important “especially during critical developmental windows”).

mark the truth or falsity of an individual plaintiff’s causal allegation or a defendant’s suggested alternate cause. The image one gets is of tiny molecular flags waving from damaged DNA or proteins: red for benzene, green for X-rays, blue and yellow for polycyclic aromatic hydrocarbons in tobacco smoke, etc. At the dawn of the genomic era, it was easy for both scientists and law professors to anticipate the discovery of those flags.\textsuperscript{111} Deeper knowledge, however, has brought reason for doubt.

Scientists searching for biomarkers are concerned to ensure that the biomarkers are valid.\textsuperscript{112} Biomarker validity, from the scientific perspective, entails a number of technical requirements related to the marker’s intended use.\textsuperscript{113} Despite the large amount of research into potential biomarkers, validation of new markers remains frustrating.\textsuperscript{114}

To fulfill the hopes of tort scholars seeking relief from the puzzle of toxic causation, biomarkers must be valid both in a general sense and in a very particular way. In a general sense, the markers must be analytically valid so the results of a search for them can be trusted.\textsuperscript{115} With respect to the legal system’s needs, however, markers can only eliminate the indeterminacy of toxic causation claims if they can reliably distinguish between “true” and “false” causation claims in ill people. This implies, on the one hand, that a marker’s presence demonstrates the

\textsuperscript{111}. For an example from medical research literature, see Soderkvist & Axelson, \textit{supra} note 87, at 85 (suggesting that different patterns of DNA and protein adducts may be detected for different carcinogens). For an example from legal scholarship, by an author who is both a lawyer and a scientist, see Marchant, \textit{supra} note 110, at 109 (suggesting that biomarkers will help resolve “vexing causation issues”).

\textsuperscript{112}. See generally Paul A. Schulte & Frederica P. Perera, \textit{Validation, in Molecular Epidemiology}, \textit{supra} note 61, at 79.


\textsuperscript{115}. \textit{See} Bossuyt, \textit{supra} note 114, at 194 (discussing analytical validity).
suspected cause-and-effect mechanism, and on the other, that a marker’s absence demonstrates that the disease was caused by something other than the suspected cause.\footnote{116}{This discussion addresses only the potential of biomarkers to show that exposure to a particular toxic agent caused disease, as opposed to some other toxic agent, a genetic or other endogenous cause, or an unknown cause. There seems to be no particular reason to believe that biomarkers would be at all helpful in identifying a cause among multiple purveyors of a single indistinguishable substance, which would add its own layer of causal indeterminacy.}

For the presence of a marker to establish specific causation more deterministically than is possible with today’s evidence, the marker must link the exposure to the plaintiff’s disease and exclude other causes. A marker might unambiguously show that a particular disease is present. That won’t prove causation, however (except for the unusual disease that has been shown by other means to be a signature of a particular exposure, in which case the molecular marker would not be needed). A marker might even be able to distinguish that a particular case of disease was induced by an environmental toxin such as a carcinogen, instead of being “endogenous.”\footnote{117}{See Henricksen v. ConocoPhillips Co., 605 F. Supp. 2d 1142, 1149 (E.D. Wash. 2009) (noting a distinction between environmentally-caused cases of leukemia and “de novo,” “endogenous,” “primary,” or “idiopathic” cases).}

That might be definitive if human beings were lab rats living in controlled environments in which only the exposure of interest varied, but that is decidedly not the case.\footnote{118}{See CTRS. FOR DISEASE CONTROL & PREVENTION, FOURTH NATIONAL REPORT ON HUMAN EXPOSURE TO ENVIRONMENTAL CHEMICALS (2009), http://www.cdc.gov/exposurereport/pdf/FourthReport.pdf (presenting data on blood and urine levels of 212 chemicals in a sampling of the American population); David Ewing Duncan, The Pollution Within, NAT’L GEOGRAPHIC, Oct. 2006, at 116, 126 (describing 165 potentially toxic chemicals identified in samples of author’s body).}

Given the myriad of potentially harmful environmental agents to which we are all exposed, a biomarker’s presence alone will suffice to prove a plaintiff’s case only if the marker is specific to the exposure–disease combination.\footnote{119}{See Schulte & Perera, supra note 112, at 103 (explaining that a “specific” marker of exposure “attribut[es] negative results to a high percentage of unexposed persons”). For an example of the importance of marker specificity to a finding of causation, albeit in a different factual context, see Precourt v. Fairbank Reconstruction Corp., 856 F. Supp. 2d 327, 337 (D.N.H. 2012) (denying beef processor’s motion for summary judgment on crossclaim against beef}
Conversely, for the absence of a marker to disprove causation definitively, the marker must invariably be associated with disease caused by the suspect agent. This would imply that exposure produces the disease in question via a single biochemical pathway, that the pathway always produces the marker, and that the marker can always be detected after the disease has manifested. Given the myriad of metabolic and mutagenic pathways by which some substances can cause illness, a biomarker’s absence alone will suffice to disprove a plaintiff’s case only if the marker is perfectly sensitive to the exposure–disease combination.

In other words, the dream of certainty depends on the discovery of “signature” biomarkers that would connect harm to exposure in the same way that the handful of currently known “signature” diseases do. Biomarker studies ordinarily accept a trade-off between sensitivity and specificity. Biological and environmental complexity—of the human genome, the epigenome, metabolic and developmental processes, exposures, and their numbingly vast numbers of potential interactions—militate against the possibility that biomarkers of “signature” 

supplier because proof that bacteria from decedent was genetically identical to bacteria in supplier’s meat did not exclude the possibility that plaintiff was infected by genetically identical bacteria from another source).

120. See generally Carl F. Cranor, The Challenge of Developing Science for the Law of Torts, in PERSPECTIVES ON CAUSATION, 261, 262–63 (Richard Goldberg ed., 2011) (distinguishing induction period from latency period of a disease and explaining that either or both may be long).

121. See Schulte & Perera, supra note 112, at 103 (noting that “sensitive” marker of exposure “pick[s] up a high percentage of individuals in the exposed group”). For an example in which the sensitivity of a marker was questioned, see Declaration of Martyn T. Smith, Ph.D. ¶¶ 25–26, Milward v. Acuity Specialty Prods. Grp., 664 F. Supp. 2d 137 (D. Mass. 2009), rev’d, 639 F.3d 11 (1st Cir. 2011), cert. denied, 132 S. Ct. 1002 (2012) (No. 07CV11944), 2008 WL 7425049 (asserting that the failure to detect a chromosome abnormality after benzene exposure did not exclude possibility that benzene caused the same disease by a different mechanism).

122. See Grodsky, supra note 110, at 1707 (discussing possibility of signature biomarkers).

specificity and sensitivity are prevalent in our cells just waiting to be found.124

Some types of biochemical damage are thought to be common to numerous disease pathways or exposures. Oxidative stress is an example. Oxidation is thought to play a role in the genesis of cancer and other diseases,125 but finding valid oxidative biomarkers for particular substance–disease links has proven difficult.126 Oxidative damage from a given exposure may affect diverse biochemical components of human tissue;127 conversely, many different exposures may cause harm through an oxidative mechanism.128 Moreover, everybody’s DNA is oxidized during life, so any putative oxidative biomarker must be found against a background incidence of oxidative damage, which is highly variable from person to person.129 That background incidence and variability, whether it results from differences in genotype, environmental exposure, or both, makes it more difficult to find the type of specific and sensitive biomarkers needed for particularistic proof of toxic causation in individual cases.130

Some markers have proven to be less sensitive or specific than they at first seemed. For example, benzene is thought to cause certain leukemias by inducing relatively large-scale aberrations in chromosomes.131 But although studies have found

124. See Latterich & Schnitzer, supra note 114, at 602 (describing how study bias and genetic and epigenetic variability of patients, including gender, ethnicity, age, diet, and environmental factors, have contributed “to a level of biological complexity beyond the scope of what can be typically interrogated”).

125. See Cosetta Minelli et al., Interactive Effects of Antioxidant Genes and Air Pollution on Respiratory Function and Airway Disease: A HuGE Review, 173 AM. J. EPIDEMIOLOGY 603, 603 (2011) (describing that oxidative stress is a mechanism for air pollutants causing lung disease; Ward et al., supra note 79, at 1356 (stating that oxidative stress is believed to be carcinogenic mechanism)).


127. Ward et al., supra note 79, at 1356.

128. See Minelli et al., supra note 125, at 603 (describing oxidative stress as a “mechanism of action common to all pollutants”).

129. Ward et al., supra note 79, at 1356.

130. Id.

131. See Luoping Zhang et al., Use of OctoChrome Fluorescence in Situ Hybridization to Detect Specific Aneuploidy Among All 24 Chromosomes in
some chromosomal aberrations that occur at much higher frequencies in leukemias of patients with known occupational benzene exposure, they also occur in the control groups of these studies.132 A number of different aberrations have been associated with benzene exposure.133 The metabolism of benzene is complex;134 several metabolites may be involved in the carcinogenic effect of benzene exposure, and benzene-caused carcinogenesis is likely a multi-step process.135 Researchers do not know exactly how benzene causes chromosomal aberrations or exactly how those aberrations cause leukemia.136 Despite the progress in chromosomal and genetic study of leukemia in people exposed to benzene, the search for a “signature” biomarker continues.137

Similar issues have dogged the search for smaller toxic signatures in DNA, such as mutations of individual genes. Investigators examined the DNA of persons with a type of kidney cancer called renal cell carcinoma (RCC), which is known to begin with the mutation of a particular gene. They observed that the tumors of RCC patients who had been occupationally exposed to

---

132. See Zhang et al., supra note 91, at 6 (describing a study in which chromosome abnormalities were present in 100% of leukemia patients with benzene exposure but 54% of non-exposed patients).

133. See Luoping Zhang et al., Chromosome-Wide Aneuploidy Study (CWAS) in Workers Exposed to an Established Leukemogen, Benzene, 32 CARCINOGENESIS 605, 605 (2011) (finding that benzene-exposed subjects showed statistically significant increases in frequency of aberrations on eight chromosomes).


136. Zhang et al., supra note 133, at 610; see also Bird et al., supra note 134, at 3 (explaining that it is not known if chromosomal aberrations are random or which metabolites are responsible).

137. Zhang et al., supra note 133, at 610.
WHEN CERTAINTY DISSOLVES INTO PROBABILITY

trichloroethylene (TCE) carried a pattern of mutations not observed in people who had not been exposed or whose cancer resulted from inherited alterations in the gene. But fewer than half of the TCE-exposed study subjects showed that mutation pattern. Another group of researchers could not replicate the finding of the purported signature mutation.

RCC, moreover, is atypical in being traceable to mutations in a single gene. Much more commonly, cancer is characterized by many mutated genes, and sorting out causes from effects is difficult. The recently published genomic atlases of colon and rectal cancer and breast cancer, for example, identified a large number of genetic changes in the studied sample of tumors. In colon and rectal cancer, some of those changes were found in a large proportion of the tumors. That is good news from the perspective of treatment, because it suggests that a relatively

---

139. Id. at 859.
141. Wills, supra note 54, at 23 (“ Genetic mayhem at the gene and chromosome level can continue even after the cell becomes cancerous. This confuses the picture, covering up the traces of the original cancer-causing events.”); The Cancer Genome Atlas Network, Comprehensive Molecular Characterization of Human Colon and Rectal Cancer, 487 NATURE 330, 330 (2012) [hereinafter Human Colon and Rectal Cancer] (identifying thirty-two recurrently mutated genes and many other genetic alterations in colon cancer); Guyton et al., supra note 109, at 233 (noting it is unlikely that all 1,149 somatic mutations found in a group of breast and colorectal cancers are key events); Yih-Horng Shiao, Genetic Signature for Human Risk Assessment: Lessons from Trichloroethylene, 50 ENVTL. & MOLECULAR MUTAGENESIS 68, 70 (2009) (stating that only some genetic alterations found in tumors are tumorigenic “drivers” while others are “passengers”).
142. Human Colon and Rectal Cancer, supra note 141, at 330 (identifying “32 somatic recurrently mutated genes . . . in the hypermutated and nonhypermutated cancers”).
144. Human Colon and Rectal Cancer, supra note 141, at 333–35.
small number of pathways drive the development of these cancers, but it also suggests that finding pathways unique to individual carcinogens seems less likely. For breast cancer, the research described four major subtypes of the disease characterized by different sets of genetic and epigenetic changes, with some changes found frequently within a subtype and a few relatively common across subtypes.

Changes in gene expression, a common subject of study by DNA microarrays, also often are similarly multifarious and may be mediated both by mutations and by epigenetic effects. Their significance as markers is more limited than biologists first believed.

---

145. See Gina Kolata, In Gene Study, a Map to Fight Colon Cancer, N.Y. TIMES, July 19, 2012, at A3 (“The hope now is that the genetic alterations driving those 1,000 different tumors are operating through only a limited number of genetic pathways that can be targeted by a more manageable number of drugs.”).

146. Human Breast Tumours, supra note 143, at 61–62 (reporting that somatic mutations at three genes occurred in more than ten percent of all tumors studies, but numerous genes were frequently mutated in particular subtypes of breast cancer).

147. See Paolo Vineis & Miquel Porta, Causal Thinking, Biomarkers, and Mechanisms of Carcinogenesis, 49 J. CLINICAL EPIDEMIOLOGY 951, 955 (1996) (noting that once people are sick it is hard to distinguish altered expression that shows they are sick from altered expression that causes sickness); see also, e.g., Yoko Hirabayashi, p53-Dependent Gene Profiling for Reactive Oxygen Species After Benzene Inhalation: Special Reference to Genes Associated with Cell Cycle Regulation, 153–54 CHEMICO-BIOLOGICAL INTERACTIONS 165, 165 (2005) (determining that expression changes after benzene exposure may be masked by other genes); Sarah X.L. Huang et al., Role of Mutagenicity in Asbestos Fiber-induced Carcinogenicity and Other Diseases, 14 J. TOXICOLOGY ENVTL. HEALTH PART B 179, 214–19 (2011) (describing many gene expression changes in cancers of patients who had been exposed to asbestos).

148. Alan Dove, Biomarker Hunters Probe the Proteome, 329 SCI. 1373, 1373 (2010).

[Changes in the RNA transcription levels of a gene are mere hints of what’s actually happening in cells and tissues; the complex interactions between proteins downstream of the transcripts drive much of an organism's physiology. . . .] Validating a potential disease biomarker in the clinic turns out to be a much thornier problem than most investigators had realized.

Id. Very recently, an international consortium of researchers published findings showing that the large amount of human DNA that does not contain protein-encoding genes does include large numbers of DNA “switches” that are responsible for turning other genes on or off, and that may explain variable
Molecular complexes between possible carcinogens and DNA or proteins, known as addition products or adducts, are another potential toxic signature. Because of the long latency of many diseases that result from toxic exposure, an adduct would have to be long-lasting and detectable after disease has emerged in order to provide definitive proof in litigation. It also would have to be specific and sensitive: that is, a specific adduct would have to result only from the exposure at issue rather than from a metabolic pathway involving exposure to some other substance, and all metabolic pathways linking the accused exposure to disease would have to necessarily produce the adduct. Finally, perhaps most difficult, to satisfy this causal model the presence of the adduct would have to indicate causation and not merely exposure. The presence or absence of adducts will not

\[ \text{incidence of disease despite identity of genes. Gina Kolata, Study Discovers Road Map of DNA, a Key to Biology, N.Y. TIMES, Sept. 6, 2012, at A1. The consortium’s findings are expected to “force a rethink of the definition of a gene and of the minimum unit of heredity.” Ecker, supra note 107, at 52. Whether they herald biomarkers that will provide individualized proof in toxic tort cases, however, is far from clear. Legal use of any biomarkers that may be found in these genetic switches would face the same hurdles of specificity and sensitivity. The initial findings suggest that much complexity and interaction exist in the regulation of gene expression. See id. (describing findings of “poorly understood” regulation of genes by distant DNA regions); Wendy A. Bickmore, Expression Control, 489 NATURE 53, 53–54 (2012) (describing findings of more than 200,000 DNA “enhancers” per cell type and pairing of 500,000 enhancers with nearby target genes, which still “leaves more than 2 million putative enhancers without known targets”); Benjamin Vernot et al., Personal and Population Genomics of Human Regulatory Variation, 22 GENOME RES. 1689, 1689 (2012) (“We estimate that individuals likely harbor many more functionally important variants in regulatory DNA compared with protein-coding regions, although they are likely to have, on average, smaller effect sizes.”).}

149. See Frederica P. Perera, Molecular Epidemiology: On the Path to Prevention?, 92 J. NAT’L CANCER INST. 602, 603 (2000) (“[U]sing adducts as biomarkers has the theoretical advantage that they reflect chemical-specific genetic damage.”).

150. This may be true for some, but not necessarily all, adducts. In one study, DNA adducts of the carcinogen acetaldehyde had a half-life of thirty-five hours. Kimiko Hori et al., Stability of Acetaldehyde-derived DNA Adduct in Vitro, 423 BIOCHEMICAL BIOPHYSICAL RES. COMM. 642, 644 (2012).

151. See, e.g., Yongquan Lai et al., New Evidence for Toxicity of Polybrominated Diphenyl Ethers: DNA Adduct Formation from Quinone Metabolites, 45 ENVTL. SCI. TECH. 10,720, 10,726 (2011) (discussing a study that demonstrated adduct formation in vitro but acknowledged that in vivo
necessarily sharply distinguish cases of disease caused by a particular exposure from background cases caused by something else.152

What about viewing the gene–toxin interaction from the opposite direction: looking for genetic variations that increase susceptibility to the toxin, rather than seeking signs of the toxin’s effect on genetic material? Such research must confront a staggering amount of variability.153 Because of gene–gene or gene–environment interactions, individual genetic variants do not typically determine the occurrence of disease, either alone or in combination with exposure to toxic substances.154 A review of studies of genetic susceptibility to health effects of air pollution, for example, observed that the studies produced conflicting results. The review examined seven potentially relevant genes involved in antioxidant activity, but noted that many other implications of markers are unknown); Menglong Xiang et al., Chromosomal Damage and Polymorphisms of Metabolic Genes Among 1,3-butadiene-exposed Workers in a Matched Study in China, 27 MUTAGENESIS 415 (2012) (characterizing DNA alterations as markers of exposure).

152. See Stephen M. Rappaport et al., Protein Adducts as Biomarkers of Human Benzene Metabolism, 153–54 CHEMICO-BIOLOGICAL INTERACTIONS 103, 104 (2005) (noting “significant background levels” of adducts in persons in control group); Vineis & Perera, supra note 59, at 1954 (stating that adduct studies have shown “overall correlations” between adduct levels and exposures).

153. See supra note 99 and accompanying text (describing the large number of variations found in human DNA).

154. See Liam R. Brunham & Michael R. Hayden, Response, 337 SCIENCE 911, 911 (2012) (“[R]esults . . . appear to support [Nebert & Zhang’s] view as regards common diseases, as many studies have reported a large number of loci, each conferring a small risk of disease.”); D.W. Nebert & G. Zhang, Personalized Medicine: Temper Expectations, 337 SCIENCE 910, 910 (2012) (“[O]ne can infer that accurate statistical predictions of a complex trait require identification of many small-effect variants . . . . For most complex traits, this is an unachievable goal.”); Rajfan & Semple, supra note 70, at 10 (“[I]t is extremely difficult in an individual patient confidently to link a disease to only one or two mutations.”); Ward et al., supra note 79, at 1356 (asserting that the magnitude of genetic associations with toxic susceptibility “may be modest and involve multiple genes”); see also, e.g., Stacey et al., supra note 99, at 1098 (noting that the gene with the greatest increased risk of basal cell carcinoma had odds ratio of 2.36); Olivia Fletcher et al., Association of Genetic Variants at 8q24 with Breast Cancer Risk, 17 CANCER EPIDEMIOLOGY BIOMARKERS & PREVENTION 702, 705 (2008) (explaining that a study, although statistically equivalent to 12,000 samples, had insufficient statistical power to detect modest gene–gene interactions).
potentially important genes exist and concluded that because antioxidant mechanisms are complex, it is unlikely that any one polymorphic gene has a large effect on susceptibility. Nongenetic factors also play a role, so three-way interactions (among a given gene, pollutants, and other genes or environmental factors other than pollution) are real possibilities.155

More generally, the Environmental Genome Project has identified nearly 90,000 variations in more than 600 genes believed to be involved in response to environmental exposures.156 For carcinogens, “[a]ddressing the role of genetic susceptibility . . . is . . . important; however, the stable and reproducible associations are few.”157

In sum, many genes and epigenetic factors may affect toxic susceptibility, toxins may affect people in many ways, and many effects may result from more than one toxin.158 It would be foolish to predict that no signature biomarker will ever be found for any disease caused by any exposure, but the evidence so far does not seem to suggest that signature biomarkers are typical.159

155. See Minelli et al., supra note 125, at 603–05, 609–13 (describing studies of three-way interactions and the difficulties involved).

156. Mark J. Rieder, The Environmental Genome Project: Reference Polymorphisms for Drug Metabolism Genes and Genome-Wide Association Studies, 40 DRUG METABOLISM REV. 241, 244 (2008); see Nebert & Zhang, supra note 154, at 910 (noting that “most examples of pharmacogenomic traits (adverse drug reactions, as well as drug efficacy) resemble complex diseases and other multi-factorial traits [that] reflect contributions from innumerable low-effect genes”).

157. Ward et al., supra note 79, at 1356.

158. See Huang et al., supra note 147, at 213 (describing multiple carcinogenic mechanisms for asbestos); Latterich & Schnitzer, supra note 114, at 602 (“Many pathologies . . . are complex and have multiple etiologies, especially at the molecular level.”).

159. See Guyton et al., supra note 109, at 231–32 (describing the necessity of evaluating multiple modes of action of carcinogens); Latterich & Schnitzer, supra note 114, at 602 (stating that “very few single biomarkers are likely to have the high sensitivity and specificity necessary to make diagnosis and treatment decisions,” although combining multiple markers might improve sensitivity and specificity); Shiao, supra note 141, at 69 (“Some agents can produce more than one type of DNA damage.”); Ward et al., supra note 79, at 1360 (noting that genomic advances “are likely to increase the challenges and complexities of carcinogen testing and evaluation” in part because most carcinogenic mechanisms are not simple and carcinogens may act through diverse pathways).
Toxicogenomics and molecular epidemiology are unlikely to undo the Gordian knot of specific causation in toxic torts.160

C. The True Promise and Its Implications

Advancing scientific understanding can assist in legal fact-finding even if science will not provide law’s longed-for, conclusive post hoc answer to the question of what did make a particular plaintiff sick. But the law must understand how science can best contribute. That understanding begins with acceptance of the fact that bringing toxicological understanding to the molecular level will not bring causation to the individual level. Even toxicogenomics and molecular epidemiology produce data that ultimately are group-based, statistical, and probabilistic—much like the data available before genomics.

Thus, finding that a plaintiff does or does not have a genetic susceptibility to the disease-causing effect of a substance to which the plaintiff was exposed will provide probabilistic but not deterministic evidence of causation or its absence.161 Toxic

160. Such research, however, may be extremely probative with respect to general causation. Mechanistic insights can form part of the weight of the evidence supporting an inference of general causation. Milward v. Acuity Specialty Prods. Grp., 639 F.3d 11, 20–23 (1st Cir. 2011). A study of susceptible genotypes can provide evidence of general causation when classical epidemiology does not. Tomlinson et al., supra note 100, at 219.

161. As two researchers explained:

Particularly in the case of low-dose toxicants, the interactions of susceptibility genes with specific environmental factors are probably the dominant cause of any resultant human illness. However, the probability that an environmental exposure will cause illness is dependent on the capacity of the genetically-controlled metabolic machinery and repair mechanisms of the cell. . . . Thus, elucidating the cause of most chronic diseases will require an understanding of both the genetic and environmental contributions to their etiologies.

susceptibility genes do not determine that an individual of a particular genotype will contract a specified illness if subjected to a given exposure.\textsuperscript{162} Rather, “[t]hey modify risk.”\textsuperscript{163} So, for example, even though a particular genotype of the \textit{NAT2} gene makes it much more likely that a woman smoker will develop breast cancer, not all women of that genotype who smoke end up with breast cancer; some women who smoke develop breast cancer even though they do not have that genotype; some women develop breast cancer even though they neither smoke nor have that genotype.\textsuperscript{164} And multiple studies of toxic susceptibility genes are unlikely to give identical results because of the influence of other factors and of random chance.\textsuperscript{165}

Biomarkers of exposure or effect similarly provide probabilistic rather than deterministic evidence. For example, it would be relevant to know if a plaintiff’s kidney cancer had a mutation pattern that is found more often among patients who had been exposed to trichloroethylene, but the pattern could nevertheless be present without causation and vice versa.\textsuperscript{166} The same is true of the chromosome aberrations associated with benzene exposure in leukemia patients.\textsuperscript{167}

that susceptibility to most human diseases “is complex and multifactorial” and describing difficulty of assessing risk contribution in gene–environment interactions).

\textsuperscript{162} See, e.g., Bookman et al., \textit{supra} note 161, at 218 (giving example of how variations in the \textit{NAT2} gene alter relative risk of smoking-related bladder cancer).

\textsuperscript{163} Olden & Guthrie, \textit{supra} note 161, at 5.

\textsuperscript{164} See Ambrosone et al., \textit{supra} note 97, at 23 (providing a table with data showing that while cigarette smoking increases the risk of breast cancer in those with the \textit{NAT2} gene, possessing that gene is not the sole determining factor).

\textsuperscript{165} See, e.g., Xiang et al., \textit{supra} note 151, at 419 (discussing how a particular genotype showed a 2.28-fold increase in a specific type of DNA damage after exposure, and the results differed from earlier studies).

\textsuperscript{166} See Brauch et al., \textit{supra} note 138, at 856 (providing data showing that of kidney cancer patients in the study who had occupational TCE exposure, fewer than half displayed a putatively characteristic mutation pattern); see also Seesnaite et al., \textit{supra} note 77, at 426–27 (noting that various epigenetic changes were found more frequently in subjects exposed to tobacco smoke, but still in fewer than half of them).

\textsuperscript{167} See Vineis & Perera, \textit{supra} note 59, at 1956 (describing a study that found “increased . . . frequencies of aberrations . . . frequently seen in . . .
Biomarkers are thus most useful from a scientific perspective “at the population level.”\textsuperscript{168} Their specificity, sensitivity, and predictive value are population-based rather than individualized attributes.\textsuperscript{169}

Toxicogenomics and molecular epidemiology are producing evidence about suspected exposure-disease links at finer and finer scales of resolution,\textsuperscript{170} but they have not altered the essential nature of that evidence. What these sciences will do, however, is produce more such evidence. More agents will be investigated to see if they are associated with molecular consequences. More genes will be interrogated to see if they affect susceptibility to agents.\textsuperscript{171} The data will still be about relative risk, but risk will be parsed more and more finely. As research discriminates among genotypes, coexposures, and other variables that cannot be addressed with classical techniques, new associations will be detected or known associations will be disaggregated in new ways. This process has already begun even for causal connections that were already relatively well-accepted.\textsuperscript{172}

At the same time, the enormous number of possible combinations of potentially interacting causal factors—genes, epigenetics, other individual characteristics, and exposures—

\textsuperscript{168} Id.

\textsuperscript{169} See Bossuyt, supra note 114, at 197 (“Like sensitivity and specificity, predictive values are essentially group based measures.”).

\textsuperscript{170} See Ward et al., supra note 79, at 1360 (“Research gaps and opportunities have been identified that can help to resolve uncertainties. . . . We hope that this process will lead to well-planned epidemiologic and mechanistic studies for these agents . . . .”).

\textsuperscript{171} See id. (“Use of omics techniques will accelerate the understanding of the cellular and molecular basis for biological responses to environmental and occupational exposures, and high-throughput technologies will increase the number of agents that can be tested.”); Minelli et al., supra note 125, at 618 (“The study of genetic susceptibility can greatly improve our understanding of air pollution pathophysiologic mechanisms of action and allow identification of those pollution components with the highest potential for harm.”).

\textsuperscript{172} See, e.g., E. Brigitte Gottschall, Taking a Retrospective Look at Asbestos-Related Thoracic Disease Produces Interesting Results, 255 Radiology 681, 682 (2010) (discussing how polymorphisms in the gene for a certain enzyme appear to affect risk from asbestos).
WHEN CERTAINTY DISSOLVES INTO PROBABILITY  279

makes it extraordinarily unlikely that complete risk characterization will ever be possible at an individual level.173 In a scientific research paradigm that depends on ceteris paribus, there is real doubt over whether every important variable can be held constant—or even identified—for the foreseeable future, if ever. Scientists know that there is much they do not know, but the “unknown unknowns” may be even more important.174

With better data on multiple exposures,175 more frequent classification of risk by genotype,176 and increasing mechanistic knowledge, science will likely point toward multiple causal sets of genetic, epigenetic, and environmental factors that are associated with disease risk. Even if some of those combinations display strong associations with a particular toxic outcome, other combinations will display weaker ones.177 The better the science gets and the larger the data sets it can assemble, the easier it will be to identify relatively small incremental contributions to risk.

The reality is that the available proof in the new world of biomarkers, toxicogenomics, and molecular epidemiology will look a lot like the proof that has been available until now in the world of differential diagnosis, toxicology, and classical epidemiology. These sciences will dramatically increase the quantity of scientific information available to the legal system but will not represent a qualitative change in the nature of the information available to address the question of ex post causal attribution.


174. See Raffan & Semple, supra note 70, at 60 (concluding that new, next-generation gene sequencing technology “now allows researchers to take an unbiased approach to gene discovery, and thus to look for ‘unknown unknowns’”). (Apologies to Donald Rumsfeld).

175. See Lin, supra note 110, at 1470–72 (describing prospects for improved exposure assessment); Wang et al., supra note 81, at 126 (describing a study that used personal monitors to measure exposure to air pollutants).

176. See Muin J. Khoury, Genetic Epidemiology and the Future of Disease Prevention and Public Health, 19 EPIDEMIOLOGIC REV. 175, 176 (1997) (predicting that genotype information “will routinely be sought in almost every epidemiologic study”).

177. See, e.g., Pharoah et al., supra note 69, at 2797, 2801 (explaining that common genetic variations confer small incremental risks).
For the most part, increased knowledge of toxicity at the genomic and molecular levels will simply provide an increasingly detailed description of probabilistic associations—population-based frequencies rather than deterministic certainties. To seek a determined causal answer in this world would be like trying to determine the weather in Seurat’s *Sunday Afternoon* by looking very closely at a few points of color in a lady’s parasol.

At a molecular level, many of the processes associated with toxicity and disease are simply random. The “probabilistic description of the mutation process cannot be replaced by a deterministic one.” Experiments have shown that even genetically identical cells exposed to the same environmental conditions can display random variations in gene expression, leading to significant differences in the chemical and phenotypic characteristics of the cells. Such “random phenotypic noise, consequent on stochastic epigenetic processes,” could have substantial effects on biological outcomes. “Nowadays it is commonly stated that disease is either genetic or environmental, when in reality stochastic events are equally important.”

In the end, toxic causation questions dwell in a world with a substantial stochastic component that does not fit well with a deterministic causal model. It does not matter whether the connection between exposure and disease is really random or whether it only looks random because a truly deterministic pathway is too complex to be fully specified. What matters is that for the reasonably foreseeable future, science will not be able to

179. *Id.* at 39.
180. *See* Mads Kaern et al., *Stochasticity in Gene Expression: From Theories to Phenotypes*, 6 *Nature Revs. Genetics* 451, 451 (2005) (describing the study, in which “[s]pecial emphasis is given to stochastic mechanisms that can lead to the emergence of phenotypically distinct subgroups within isogenic cell populations”).
181. Davey Smith, *supra* note 75, at 548; *see also* Drmanac, *supra* note 161, at 911 (noting importance of stochastic effects in attempting to measure genetic contribution to disease risk).
give law a deterministic answer. Look closely enough, and certainty dissolves into probability.

IV. An Alternative Vision: Probabilistic Causal Contribution

Courts’ square-peg–round-hole frustration continues as they try to make causation judgments by applying an individual-based, deterministic model to population-level, frequency-based probabilistic evidence. The realization that reductionist science is not likely to discover the legal system’s way out of the toxic causation problem suggests that courts should be open to a different mental model of causation—one that treats cause and effect explicitly as probabilistic—and should reconsider alternative doctrinal approaches to suit. Adapting some ideas from earlier reform proposals offered by scholars, but largely ignored by courts, I propose that courts adopt an expressly probabilistic view of causation when the dominating evidence comprises population-based data of toxic effect. To frame the standard, an exposure should be considered a cause of disease if it was a contributing factor to the disease’s occurrence. To be a contributing factor, an exposure would be shown by a

183. See Merck & Co. v. Garza, 347 S.W.3d 256, 264 (Tex. 2011) (reiterating the view that “frequency data . . . cannot indicate the cause of a given individual’s disease [but the] use of scientifically reliable epidemiological studies and the requirement of more than a doubling of the risk strikes a balance between the needs of our legal system and the limits of science”) (quoting Merrell Dow Pharm., Inc. v. Havner, 953 S.W.2d 706, 718 (Tex. 1997)); Sienkiewicz v. Greif, [2011] UKSC 10, [2011] 170 A.C. 229, [170] (Lady Hale) (contrasting use of risk data to advise individual patient before getting sick with use of risk data to infer causation of existing illness); id. [190]–[92] (Lord Mance) (describing tension between use of statistical evidence and law’s “concern[] with the rights and wrongs of an individual situation” and expressing preference for use of epidemiology “in conjunction with specific evidence related to the individual circumstances and parties”).

184. One clarification is essential. Although genomic data will not likely provide definitive particularistic proof of causation, such data could still be relevant to specific causation even if courts do not adopt probabilistic causal contribution, along with other evidence deemed relevant in the absence of deterministic proof.

185. Many of these are discussed in Michael D. Green, The Future of Proportional Liability: The Lessons of Toxic Substances Causation, in EXPLORING TORT LAW 352, 357–70 (M. Stuart Madden ed., 2006).
preponderance of the evidence—not limited to any single favored
type of evidence—to have added incremental risk that the
plaintiff would develop a disease that the plaintiff in fact
developed. Damages should be apportioned to that contributing
factor in proportion to its contribution to the plaintiff’s risk.

A. A Metaphor

An illustration from the Third Restatement of Torts allows a
metaphorical comparison of the deterministic causation model
that fits typical cases and the probabilistic model that better fits
most toxic torts. The illustration posits three defendants—Able,
Baker, and Charlie—who negligently, independently, and
simultaneously lean against a car. Collectively they provide
enough force to send the car over a diminutive curb and down the
side of the mountain. Any two of the actors together would have
propelled the car over the edge, yet each actor alone is too weak
to budge the car. Thus, no single actor’s tort is either a sufficient
or a necessary cause of the harm.186

The illustration addresses the problem of multiple sufficient
causal sets, which are analogous only to a particular subset of
toxic tort claims.187 For present purposes, what matters is that

186. The hypothetical is given in Restatement (Third) of Torts: Liab. for
Physical & Emotional Harm § 27 cmt. f, illus. 3 (2010). See id. § 27 cmt. f (“In
some cases, tortious conduct by one actor is insufficient . . . to cause the
plaintiff’s harm. Nevertheless, when combined with conduct by other persons,
the conduct overdetermines the harm, i.e., is more than sufficient to cause the
harm.”).

187. The analogy is to several actors who contribute subthreshold doses of a
toxin that exerts toxic effects only after a threshold dose is reached. See id. cmt. g
(“Assuming that there is some threshold dose sufficient to cause the disease, the
person may have been exposed to doses in excess of the threshold before
contracting the disease. Thus, some or all of the person’s exposures may not have
been but-for causes of the disease.”). The Restatement’s quite proper rule, in both
cases, is that at least until the threshold is reached, each contributor is a cause of
the harm, even though causal sets that did not include that contributor would
also have been sufficient to produce the harm. See id. cmt. f (“When an actor’s
tortious conduct is not a factual cause of physical harm under the [but-for]
standard . . . only because one or more other causal sets exist that are also
sufficient to cause the harm at the same time, the actor’s tortious conduct is a
factual cause of the harm.”); id. cmt. g (“Nevertheless, each of the exposures prior
to the person’s contracting the disease . . . is a factual cause of the person’s
WHEN CERTAINTY DISSOLVES INTO PROBABILITY

the illustration works only because it fits an implicitly assumed mechanistic model of causation. Simple Newtonian physics describes the situation: we can compute the force required to overcome the car’s inertia and the static friction of its tires on the parking surface. If we could reconstruct the accident, we might find that 300 pounds of force were required to move the car and that Able, Baker, and Charlie each provided 200 pounds. This knowledge, at least in qualitative terms, is implicit in the illustration’s assumptions.

Suppose, however, that Able, Baker, and Charlie could not be described by Newtonian physics but only by quantum mechanics. On a mountaintop ringed with cars, the three charge around blindfolded. What is more, they are joined by undetectable sprites that also impart momentum to any object they strike. Sometimes Able, Baker, and Charlie hit a car, and sometimes the impact is powerful enough to tip the car down the hill. But this is a quantum world: if we know what they hit, we cannot tell how hard they hit it. And we can’t detect the sprite strikes at all. Every once in a while a car rolls down the hill. But the most science can tell us—if we can say whether Able, Baker, Charlie, or any combination of the three hit the car at some point before its descent—is the probability that they hit the car hard enough to make it move.

If Able, Baker, and Charlie represent independent risk factors for a disease, and the invisible sprites represent unknown causes, then the probabilistic metaphor fits a wide range of toxic tort cases. In fact, toxic tort plaintiffs with this much information—those who find some epidemiologic data that, despite being inherently group-based and probabilistic, connect their exposure with their disease—have been the lucky ones. Toxicogenomics and molecular epidemiology will make that type of information available in more cases and in a more tailored

---

188. This of course is a metaphorical modification of the uncertainty principle. See RUVINSKY, supra note 178, at 4 (explaining Heisenberg’s uncertainty principle).
way, but they will not change the nature of the appropriate model.

**B. Contributing Factor Causation**

The first step toward a different model of causation is to acknowledge frankly the irreducible indeterminism of post hoc causal assessments. A few courts that decided relatively early toxic tort cases, recognizing that the traditional “logical model . . . does not suit the toxic tort explanandum,” seemed to move toward such an alternative model. These courts used “substantial factor” as their anchor. Allen *v.* United States remains a leading example and stands in analytical counterpoint to the Tenth Circuit’s decisions in June and Wilcox.

---

190. The Third Restatement explained Elam as an instance in which “substantial factor” was invoked to deal with multiple sufficient causes. Restatement (Third) Torts: Liab. Physical & Emotional Harm § 26 cmt. c, reporters’ note (2010). Elam discussed multiple-sufficient-cause situations, but the issue on appeal was whether plaintiffs’ evidence proved that defendant’s chemical releases caused injury to plaintiffs at all. Elam, 765 S.W.2d at 174, 183. The court invoked “substantial factor,” but never suggested that it believed the evidence showed more than one sufficient cause had operated to produce plaintiffs’ conditions. See id. at 174 (determining that the substantial factor test suited toxic exposure cases where harm may result from a “confluence of causes”); id. at 187 n.63 (“[T]he substantial factor rule of causation . . . applies to [toxic tort] cases.”); id. at 195 (referring to many different possible causes).
192. Compare id. at 429–43 (applying the substantial factor rule to the plaintiffs), and Elam, 765 S.W.2d at 174–77 (distinguishing but-for from substantial factor), with June *v.* Union Carbide Corp., 577 F.3d 1234, 1241 (10th Cir. 2009) (determining that but-for causation was appropriate), and Wilcox *v.* Homestake Mining Co., 619 F.3d 1165, 1168 (10th Cir. 2010) (agreeing with June). See also James *v.* Bessemer Processing Co., 714 A.2d 898, 908–09 (N.J. 1998) (“To prove medical causation, a plaintiff must show that the exposure [to each defendant’s product] was a substantial factor in causing or exacerbating the disease.”) (citation omitted); id. at 913 (holding, without ever referring to “but-for” causation, that plaintiff, who worked at a drum reconditioning facility, had produced sufficient evidence of causation to withstand motion for summary judgment despite inability to prove the number or specific contents of drums from various defendants).
In *Allen*, twenty-four plaintiffs contended that radiation from above-ground nuclear weapons tests caused their cancers. As in *June* and *Wilcox*, general causation was indisputable, but specific causation was another matter. The plaintiffs' various "non-specific" cancers193 might have been caused by the accused bomb-test radiation, by radiation from other sources, or by something other than radiation. The impossibility of proof led the court to eschew but-for causation. Instead, *Allen* concluded that

> [w]here it appears from a preponderance of the evidence that the conduct of the defendant significantly increased or augmented the risk of somatic injury to a plaintiff and that the risk has taken effect in the form of a biologically and statistically consistent somatic injury, i.e., cancer or leukemia, the inference may rationally be drawn that defendant's conduct was a substantial factor contributing to plaintiff's injury.194

*Allen*’s creative move was the deft elision of risk augmentation with injury causation. The court’s inference from one to the other was “rationally drawn” because the scientific evidence needed to prove causation was, and could be, framed only in risk terms. Nevertheless, *Allen* broke new ground. 195 Its statement of the standard for proof of causation followed a lengthy discussion of creative common law solutions to other problems of causal indeterminacy, but none of those re-envisioned causation in probabilistic terms based on risk augmentation.196

The closest parallel is a sixteen-year-old opinion by the California Supreme Court in a worker’s compensation case, *McAllister v. Workmen's Compensation Appeals Board*.197 In

---

193. *Allen*, 588 F. Supp. at 406. Presumably, by “non-specific” the court meant that the cancers were not uniquely associated with a particular exposure, i.e., they were not "signature" diseases. *Id.*

194. *Id.* at 428. The court allowed the defendant to defeat the inference if “the facts [were] proven otherwise by sufficient evidence.” *Id.*

195. See *id.* at 415 (“A remedial framework can certainly be fashioned to meet the circumstances and requirements of the parties and issues now before this court in this action.”).

196. *Id.* at 406–10. *Allen* cited alternative liability cases, multiple sufficient cause cases, and idiosyncratic cases in which absence of evidence makes the post hoc counterfactual inference especially difficult. *Id.*

McAllister, a firefighter’s widow claimed that her husband’s on-the-job exposure to smoke caused his fatal lung cancer, although he had also smoked cigarettes for decades. The opinion prefigured the factual causation issues that would come to bedevil toxic tort litigation:

Given the present state of medical knowledge, we cannot say whether it was the employment or the cigarettes which “actually” caused the disease; we can only recognize that both contributed substantially to the likelihood of his contracting lung cancer. . . . Future scientific developments will tell us more about lung cancer. Ultimately it may be possible to pinpoint with certainty the cause of each case of the disease. But the Legislature did not contemplate years of *damnum absque injuria* pending such scientific certainty.

To avoid that result, McAllister invoked precedent holding that a worker could obtain benefits if the employment was a “contributing cause” of an injury. How could one tell whether exposure to smoke while firefighting was a “contributing cause” of lung cancer in a long-term tobacco smoker? By considering whether “the likelihood of contracting lung cancer from the smoking was so great that the danger could not have been materially increased by exposure to the smoke produced by burning buildings.”

Thus, in different contexts and with dramatically different amounts of evidence to work with, both McAllister and Allen recognized that scientific indeterminacy of specific causation justifies a risk-based, probabilistic reconceptualization of cause-

---

198. *Id.* at 314.
199. *Id.* at 319.
200. *Id.* McAllister went far beyond the precedent it relied on, which simply recognized that an event could have more than one cause. Emp’rs Mut. Liab. Ins. Co. of Wis. v. Indus. Accident Comm’n, 263 P.2d 4, 6 (Cal. 1953) (holding that injury could arise out of employment even if employment was not “the sole cause” but only “a contributory cause”).
201. *McAllister*, 445 P.2d at 319 (emphasis added). McAllister also held that there was sufficient evidence to establish what is now known as general causation, even though there was no proof of exactly how smoke from building fires caused cancer. *Id.*
in-fact. But McAllister remained obscure and Allen's causation analysis garnered little attention.\footnote{202}

To address a different but related problem of inherent causal indeterminacy, however, courts in both the United States and the United Kingdom have much more prominently linked evidence of risk creation to inferences of causation. In \textit{Rutherford v. Owens-Illinois, Inc.},\footnote{203} the California Supreme Court confronted the "indeterminate defendant" problem typical of many asbestos cases. Mr. Rutherford had died of lung cancer after being tortiously exposed to asbestos fibers by numerous defendants, including his former employer and the manufacturers of asbestos-containing products that he had used at work.\footnote{205} Science could not interrogate the cancer to determine the source of the asbestos fiber or fibers that caused the malignancy.\footnote{206} The trial court instructed the jury on alternative liability, shifting the burden of proof on causation to the defendants.\footnote{207} The California Supreme Court held that doctrine inapplicable, but not because the

\begin{itemize}
\item \footnote{202}{Just a handful of California worker’s compensation cases and two dissenting opinions outside of California have cited McAllister. Only Elam v. Alcolac, Inc. cited Allen in the course of adopting a causation standard anything like Allen’s. See Elam v. Alcolac, Inc., 765 S.W.2d 42, 174 (Mo. Ct. App. 1988) (“The substantial factor standard . . . is particularly suited to injury from chronic exposure to toxic chemicals where the sequent manifestation of biological disease may be the result of a confluence of causes.”).}
\item \footnote{203}{Rutherford v. Owens-Ill., Inc., 941 P.2d 1203 (Cal. 1997).}
\item \footnote{204}{See id. at 1218 (“[A]sbestos-related cancer would, under the single-fiber theory of carcinogenesis, be an example of alternative causation, i.e., a result produced by a single but interminable member of a group of possible causes.”).}
\item \footnote{205}{Id. at 1207.}
\item \footnote{206}{See id. at 1206 (implying that plaintiff could not “prove with medical exactitude that fibers from a particular defendant’s asbestos-containing products were those, or among those, that actually began the cellular process of malignancy”).}
\item \footnote{207}{See id. at 1206–08. (“This instruction shifts the burden of proof to defendants in asbestos cases tried on a products liability theory to prove that their products were not a legal cause of the plaintiff’s injuries, provided the plaintiff first establishes certain predicate facts . . . .”). Mr. Rutherford also had been a smoker, adding the possibility that the cancer had really been caused by tobacco smoke rather than asbestos. See id. at 1209 (“Undisputed evidence indicated that smoking sharply increases the risk of lung disease, including lung cancer, and works ‘synergistically’ with asbestos exposure to enhance the severity of resulting damage to the lungs.”); infra notes 268–69 and accompanying text.}
\end{itemize}
plaintiff could (and therefore must) establish which of the tortfeasors had delivered the fiber that actually caused his cancer. Rather, the court held, “no insuperable barriers” prevented the plaintiff from proving causation without relying on the alternative causation doctrine—if causation were appropriately understood in the circumstances of the case.

The most important of those circumstances was the “irreducible uncertainty” of determining which defendants’ asbestos fibers actually contributed to the cellular development of cancer. Despite that uncertainty, the court observed, every exposure to asbestos increased plaintiff’s risk of disease. The court therefore conceived all of the exposures as concurrent rather than alternative causes. Left implicit in this conceptual shift was rejection of the deterministic model of causation as inappropriate. Implementing that rejection, the court held that a particular product would be a “substantial factor in causing or bringing about the disease” if “it was a substantial factor

208. See Rutherford, 941 P.2d at 1223 (“In an asbestos-related cancer case, the plaintiff need not prove that fibers from the defendant’s product were the ones, or among the ones, that actually began the process of malignant cellular growth.”).

209. Id. at 1206.

210. Id. at 1218.

211. See id. at 1209 (“[The defendant’s] own medical expert . . . testified . . . that if a worker had occupational exposure to many different asbestos-containing products, each such exposure would contribute to the risk of contracting asbestos-related lung cancer . . . .”). As Michael Green pointed out to me, if any exposure occurred after the malignant transformation, that exposure did not in fact contribute to the risk of developing the tumor that plaintiff actually developed. Determining which exposures occurred before and after the malignancy began, however, is also impossible. Moreover, it is at least possible that continued exposure presents an additional carcinogenic risk even after a particular tumor has begun to grow.

212. Id. at 1220–21, 1223.

213. Cf. id. at 1218 (stating that the mechanism of cancer causation by asbestos remained a debated scientific issue, the resolution of which could affect legal conclusions regarding causation in cases of exposures from multiple sources). But cf. Jane Stapleton, The Two Explosive Proof-of-Causation Doctrines Central to Asbestos Claims, 74 BROOK. L. REV. 1011, 1029 (2009) (stating that Rutherford’s concurrent causation approach is based on “a fiction[] that every asbestos fiber was involved in the cancer mechanism”).
contributing to plaintiff’s or decedent’s risk of developing cancer.”

A year after *Rutherford*, the New Jersey Supreme Court issued an opinion that was consistent with *Rutherford’s* reasoning, although it did not explicitly equate risk augmentation and causation. Mr. James worked in a drum reprocessing facility where he was exposed to different types of chemical and petroleum residues in drums from many companies. His widow alleged that these residues caused Mr. James’s fatal stomach and liver cancer. Witnesses testified about which companies’ drums were processed during Mr. James’s employment, and expert witnesses testified that residues these companies shipped were carcinogenic. But a lack of records made it impossible to prove how much of which residues from which companies Mr. James had come in contact with. The court nevertheless decided that the evidence of causation against some defendants was adequate to survive summary judgment. To establish “medical causation,” the New Jersey Supreme Court held, a plaintiff like Mr. James must prove two things: (1) sufficient exposure to defendant’s products and (2) “medical and/or scientific proof of a nexus between the exposure and the plaintiff’s condition.” This is another way of saying that the defendant augmented the plaintiff’s risk of disease.

216. *Id.* at 901.
217. *Id.* at 904–06.
218. *Id.* at 903, 913.
219. *Id.* at 913–14.
220. *Id.* at 911. To test sufficiency of exposure the court adopted the criterion of “frequent, regular and proximate exposure” borrowed from many asbestos cases. *Id.*
221. The close relation of risk augmentation to the *James* description of medical causation is evident in the court’s description of the way a defendant might avoid joint and several liability under New Jersey’s Comparative Negligence Act. Once “deemed [a] substantial factor” in causing James’s cancer, the court held, a defendant would bear the burden of proving a basis for apportioning fault. *Id.* at 916. “[E]ach defendant may seek to reduce its individual percentage of fault by submitting proof that its products . . . were less carcinogenic . . . or that James’s exposure to its products was more limited . . . .” *Id.*
Courts in the United Kingdom reached a similar result by a path that began with a decision in a somewhat odd case, *McGhee v. National Coal Board*. Mr. McGhee became covered in dust and sweat while cleaning out brick kilns and had to bicycle home in that condition because his employer did not provide washing facilities. He developed dermatitis and sued his employer. The medical experts agreed that the presence of the irritating dust on skin made vulnerable by perspiration caused Mr. McGhee’s dermatitis. The rub was that only the lack of a shower that extended the exposure to the dust, and not the working conditions that first deposited the dust on Mr. McGhee’s skin, was held negligent. Longer exposure increased the risk, but no expert could say whether Mr. McGhee would have avoided dermatitis had he showered at his workplace. Therefore, the trial and intermediate appellate courts concluded, Mr. McGhee could not prove by a preponderance of the evidence that the employer’s negligence had caused his dermatitis. The losing plaintiff appealed to, and prevailed in, the House of Lords.

Four of the five members of the panel reasoned that in the circumstances of the case, to establish causation, it was sufficient that Mr. McGhee had shown that his employer’s negligence had

223. *Id.* at 3–4 (Lord Reid).
224. *Id.* at 3.
225. *Id.*
226. *Id.*
227. *See id.* at 5 (Lord Wilburforce) (“The experts could [determine the cause of dermatitis], but had to admit that they knew little of the quantity of dust or the time of exposure necessary to cause a critical change.”).
228. *See McGhee*, 1 W.L.R. at 3–4 (Lord Reid) (“It was held in the Court of Session that the appellant had to prove that his additional exposure to injury . . . caused the disease in the sense that it was more probable than not that this additional exposure to injury was the cause of it.”).
229. The British cases refer to the party seeking relief as the pursuer or claimant, and to the party from whom relief is sought as the defender. For simplicity and ease of reading, I use the familiar American terms plaintiff and defendant (including survivors of decedents).
230. *See id.* at 13 (“Appeal allowed.”).
materially increased his risk of disease. Lord Salmon’s speech was representative:

[W]hen it is proved, on a balance of probabilities, that an employer has been negligent and that his negligence has materially increased the risk of his employee contracting an industrial disease, then he is liable in damages to that employee if he contracts the disease notwithstanding that the employer is not responsible for other factors which have materially contributed to the disease. . . . In the circumstances of the present case, the possibility of a distinction existing between (a) having materially increased the risk of contracting the disease, and (b) having materially contributed to causing the disease may no doubt be a fruitful source of interesting academic discussions between students of philosophy. Such a distinction is, however, far too unreal to be recognised by the common law.231

The House of Lords revived McGhee’s reformulation of the concept of causation when it confronted a case similar to Rutherford, in which each claimant or decedent had developed mesothelioma after being exposed to asbestos by multiple defendants.232 In Fairchild v. Glenhaven Funeral Services Ltd.,233 the Lords allowed the appeals of three claimants who had lost below for failure to establish but-for causation.234 The committee members differed in their analyses and in the limits they would impose on the doctrine being announced.235 They were

---

231. Id. at 12–13 (Lord Salmon); see also id. at 5 (Lord Reid) (“From a broad and practical viewpoint I can see no substantial difference between saying that what the respondents did materially increased the risk of injury to the appellant and saying that what the respondents did made a material contribution to his injury.”); id. at 8 (Lord Simon of Glaisdale) (“‘Material reduction of the risk’ and ‘substantial contribution to the injury’ are mirror concepts in this type of case.”).

232. McGhee’s formulation needed to be revived because intervening cases had cast doubt on the interpretation of McGhee. See Sanders, supra note 1, at 23–28 (tracing the discussion of the relation between risk contribution and causation in United Kingdom cases).


234. See id. [1] (Lord Bingham of Cornhill) (“[I]t was announced that these three appeals would be allowed.”).

235. See id. [35]; [45] (Lord Nicholls); [118] (Lord Hutton); [170] (Lord Rodger).
unanimous, however, on the implication of the “rock of uncertainty” that made it impossible for each claimant to prove which asbestos fiber(s) had invaded the cell from which the fatal mesothelioma grew. In light of that scientific indeterminacy, a material contribution to risk should be considered a “sufficient degree of causal connection” to support a finding of causation. The panel members noted the parallels with Rutherford, and all but one openly acknowledged that Fairchild embraced a legal standard of causation different from the deterministic but-for model.

These cases, from McAllister to Allen to Rutherford to James to Fairchild, groped toward a redefinition of specific causation in cases in which scientific indeterminacy rendered it impossible to prove by a preponderance of the evidence that a particular exposure was a but-for cause of a plaintiff’s harm. Treating risk creation as causal contribution is appropriate when a plaintiff can prove that tortious conduct “materially increased the risk of him contracting a particular disease and the disease occurred, but where in the state of existing medical knowledge he is unable to prove by medical evidence that the [conduct] was a [but-for] cause of the disease.”

236. Id. [7] (Lord Bingham of Cornhill).

237. See id. [7] (Lord Bingham), [49] (Lord Hoffmann), [77] (Lord Hutton), [120] (Lord Rodger) (acknowledging that current medical knowledge is insufficient to determine which asbestos fiber caused the cancer).

238. Id. [42] (Lord Nicholls of Birkenhead); accord id. [34] (Lord Bingham); id. [47] (Lord Hoffmann); id. [108]–[09] (Lord Hutton); id. [168] (Lord Rodger of Earlsferry).

239. See id. [31] (Lord Bingham), [73] (Lord Hoffmann), [105] (Lord Hutton), [161] (Lord Rodger) (noting that the court was asked to apply the rule from Rutherford that “the causal requirements of the tort were satisfied by proving that exposure to a particular product”).

240. See id. [9] (Lord Bingham), [41], [43] (Lord Nicholls), [56], [63] (Lord Hoffmann), [168] (Lord Rodger) (variously characterizing the Fairchild holding as a variation, relaxation, or policy-based deviation from traditional rules of factual causation); but see id. [109] (Lord Hutton) (taking the view that in McGhee and Fairchild the plaintiffs succeeded based on judicial inference that but-for causation had been established more likely than not, though acknowledging that this interpretation makes “little practical difference”).

241. Id. [108] (Lord Hutton). This formulation, by the one member of the Appellate Panel who declined to deviate expressly from traditional but-for causation, is almost indistinguishable from the way the other members of the
It is immediately apparent that this general statement of principle need not be restricted to asbestos cases or even to indeterminate-defendant cases, despite the courts’ tendency to do so. The same type of indeterminacy afflicts a wide range of toxic tort claims that have two salient characteristics: (1) general causation is reasonably well-established (for example, by appropriately confirmed epidemiologic data showing increased risk associated with exposure); and (2) science cannot specify the cause of an individual case of disease. As explained above, molecular science is likely to increase the prevalence of the first characteristic without greatly decreasing the prevalence of the second. Thus, increasingly, a probabilistic conception of causation based on risk contribution will be needed to conform legal doctrine to the realistically available evidence.

Two asserted bases for limiting this view of causation to asbestos indeterminate-defendant cases merit discussion. The first argues that this concept of causation should apply only if the competing possible causes of plaintiff’s illness all contributed to risk in the same way, e.g., by exposing a plaintiff to the same substance (i.e., asbestos). The second argues that this concept panel expressed the rule. See also Barker v. Corus UK Ltd., [2006] UKHL 20, [2006] 2 A.C. 572, [77] (Lord Rodger) (noting that McGhee held “that, in the particular circumstances, by proving that the defenders had materially increased the risk of injury, the pursuer had proved that they had materially contributed to his injury”).


243. Sanders, supra note 1, at 32–33 (describing reticence of courts, including California courts, to extend the Rutherford “risk rule” to contexts beyond asbestos); but cf. Green v. Alpharma, Inc., 284 S.W.3d 29, 35–40 (Ark. 2008) (using “substantial factor” formulation, frequency-regularity-proximity test from asbestos cases, and testimony about leukemogenic risks of arsenic, to reverse summary judgment for poultry producers, each of which contributed an unknown amount of plaintiff’s exposure to arsenic-laced chicken litter).

244. See Fairchild, [2002] UKHL 22, [115], [118] (Lord Hutton) (arguing that risk contribution model is inappropriate if multiple possible causal agents exist); Green, supra note 94, at 546 (“Any risk contribution scheme should be limited to a single toxic substance whose risk profile is established (asbestos is surely that) and in which multiple defendants contributed to the risk but plaintiff is unable to prove which one(s) is the actual cause of her harm.”).
of causation should apply only if all of the sources of plaintiff's exposure to enhanced risk are tortious.\textsuperscript{245} From a theoretical perspective, neither argument is persuasive.

The first argument seems rooted in a concern for being reasonably certain of the substance that caused the disease and for being able to compare various risk contributions. It fails because exposure to an additional risk factor, and not only exposure to more of a single risk factor to which the plaintiff was already exposed, may increase a plaintiff's risk. As Lord Hoffmann noted in \textit{Fairchild}, “what if [a plaintiff] had been exposed to two different agents—asbestos dust and some other dust—both of which created a material risk of the same cancer and it was equally impossible to say which had caused the fatal cell mutation? I cannot see why this should make a difference.”\textsuperscript{246} Lord Hoffmann later tweaked this view, explaining that the \textit{substance} to which a plaintiff was exposed need not be the same, but the \textit{mechanism} of disease causation must be the same for all of the exposures at issue.\textsuperscript{247} The latter distinction, however, is equally untenable. The justification for the “risk rule”\textsuperscript{248}—the inability to tell which of multiple exposures caused a given case of illness—is just as strong when the exposures increase the risk of a disease in different ways, if after the plaintiff is sick it is impossible to tell which mechanism operated to cause the illness.

Further, if identity of mechanism were crucial, then \textit{defining} mechanism would become critical. That definition would depend on both the state of scientific knowledge and on judicial line-

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{245} See Barker, [2006] UKHL 20, [14]–[16] (Lord Hoffmann) (acknowledging though disagreeing with the argument that a risk-based conception of causation should not apply if one of the sources of exposure was not tortious); \textit{id.} [128] (Baroness Hale) (opining that rationale for imposing liability on those responsible for all asbestos exposures is absent or weakened if not all exposures were tortious).
\item \textsuperscript{246} \textit{Fairchild}, [2002] UKHL 22, [72] (Lord Hoffmann).
\item \textsuperscript{247} See Barker, [2006] UKHL 20, [24] (Lord Rodger) (“It may have been different in some causally irrelevant respect, as in Lord Rodger’s example of the different kinds of dust, but the mechanism by which it caused the damage, whatever it was, must have been the same.”) \textit{Cf. id.} [17] (referring ambiguously to exposure to “the same risk”); \textit{see also} Sanders, \textit{supra} note 1, at 29 (calling Lord Hoffmann’s revision “a delightful piece of self-reinterpretation”).
\item \textsuperscript{248} The nomenclature is Joseph Sanders’s. Sanders, \textit{supra} note 1, at 11.
\end{itemize}
\end{footnotesize}
WHEN CERTAINTY DISSOLVES INTO PROBABILITY

drawing. For Lord Hoffmann, asbestos and some other hypothesized “dust” were similar enough, but asbestos and tobacco smoke were not. For the California Supreme Court in 1968 (when the mechanism of lung carcinogenesis was completely mysterious) smoke was smoke, whether it came from cigarettes or burning buildings. If two substances are shown to cause the same histological form of cancer, should the law treat them as acting by the same mechanism? What if it can be shown that both are capable of altering DNA? Or must each be shown to increase cancer risk by turning off a tumor suppressor gene? Must it be the same tumor suppressor gene? What if both of them are shown sometimes to turn off the same tumor suppressor gene, and other times to turn off another tumor suppressor gene, and both genes are turned off in the patient’s cancer? If the mechanisms by which two or more substances cause a disease were distinct, specific, and determinable after the fact—the false promise—then specific causation would be scientifically knowable, and there would be no need to conceptualize causation in probabilistic terms. In any other situation, however, mechanistic understanding at smaller scales will make it harder for two exposures to look the “same” but will not solve the indeterminacy problem that led to the Fairchild solution.

The second argument, that a probabilistic, risk-augmentation view of causation should be limited to cases in which all the contributors of increased risk are tortious, seems rooted in the view that the only justification for modifying causation rules in

249. See id. at 36 (noting that such limits are both arbitrary and subject to change as science learns more).

250. See Barker, [2006] UKHL 20, [24] (Lord Rodger) (“I do not think that the exception applies when the claimant suffers lung cancer which may have been caused by exposure to asbestos or some other carcinogenic matter but may also have been caused by smoking . . . .”).

251. See McAllister v. Workmen’s Comp. Appeals Bd., 445 P.2d 313, 318–19 (Cal. 1968) (“We cannot doubt that the more smoke decedent inhaled—from whatever source—the greater the danger of his contracting lung cancer . . . . Given the present state of medical knowledge, we cannot say whether it was the employment or the cigarettes which ‘actually caused the disease . . . .’”).

252. It is true, however, that exposure to multiple risk factors may present formidable factual complexities. See infra notes 274–75 and accompanying text (discussing the possibility of the same solvent in materials produced by two different manufacturers).
the face of indeterminacy is the relative moral position of an “innocent” plaintiff and multiple “wrongdoers.”

The fact that each claimant in *Fairchild* had been exposed to asbestos by multiple employers distinguished each claim from what would have been an easy case if only one entity had been responsible for the exposure. But what distinguished these claimants from the general population of the United Kingdom was the fact that they were exposed to asbestos on the job. The source of the other exposure did not logically affect the determination of whether the tortious exposure contributed materially to the risk. The House of Lords recognized this in *Barker v. Corus UK Ltd.*, which applied *Fairchild* to hold defendant employers liable to a claimant even though one of the claimant’s material exposures to asbestos was entirely his own fault.

The same reasoning would apply regardless of the competing source of exposure or risk. Thus, the Supreme Court of the United Kingdom reaffirmed this aspect of the logic of *Barker* in *Sienkiewicz v. Greif (UK), Ltd.*, which involved two women with mesothelioma who had been exposed to relatively small amounts of asbestos at their respective workplaces. The defendants


255. See *id.* at [17] (Lord Hoffmann) (“These distinctions may be relevant to whether and to whom responsibility can also be attributed, but from the point of view of satisfying the requirement of a sufficient causal link between the defendant’s conduct and the claimant’s injury, they should not matter.”); *id.* [58]–[59] (Lord Scott) (asking how the *Fairchild* principle would apply to a case involving a claimant’s exposure to multiple sources and determining that this would make no difference); *id.* [97] (Lord Rodger) (“Having reserved my opinion on the point in *Fairchild*, I would now hold that the rule should apply in that situation.”); *id.* [117] (Lord Walker) (agreeing, except still believing that “the *Fairchild* principle [should apply] to cases where less than 100% of the risk has been caused by employers or occupiers guilty of breaches of duty”); *id.* [128] (Baroness Hale) (explaining that because *Barker* imposed only several liability based on risk contribution, “[t]he victim’s own behaviour is only relevant if he fails to take reasonable care for his own safety during a period of tortious exposure by a defendant.”).


257. See *id.* [2]–[4], [59]–[60], [115], [117]–[19], [124]–[25] (describing the
argued that each woman’s exposure to asbestos in the ambient air exceeded her incremental exposure at work, and therefore it was more likely than not that the ambient exposure, rather than the workplace exposure, had caused her disease. The Supreme Court unanimously rejected the argument, holding that so long as the negligent employer’s contribution to the risk of mesothelioma had been “material,” that negligence would be considered a cause of the claimant’s mesothelioma. Similarly, it should not matter if the competing exposure was not created by human agency at all, such as exposure to radiation by naturally occurring radioisotopes as in Allen, June, and Wilcox.

Allen shows that the traditional “substantial contributing factor” formulation of the causal connection opens the possibility of reformulating causation in risk-creation terms to address specific causation in toxic torts. The treatment of asbestos indeterminate-defendant cases in California and the United Kingdom shows that the reformulation can be credibly applied. “At the very least,” it proves “that it is not necessarily the hallmark of a civilised and sophisticated legal system that it treats cases where strict proof of causation is impossible in exactly the same way as cases where such proof is possible.”

C. Risk as Cause, Not as Harm

A probabilistic model of specific causation in toxic torts, as opposed to deterministic but-for causation, is best supported by

258. See id. [60]–[61] (Lord Phillips) (“[T]he judge . . . heard expert evidence which quantified this exposure and compared it to the environmental exposure that would be experienced by everyone. . . . It was on the basis of this finding that the judge held that the claimant’s case on causation had not been made out.”).

259. See id. [107]–[09] (Lord Phillips) (determining what constitutes a “material risk” and applying that determination to this case).

260. See Sanders, supra note 1, at 40 (“[A]ll of the distinctions designed to rein in the reach of the risk rule feel arbitrary.”).

existing and developing scientific evidence—but not in the sense that some particularistic trait of an individual’s disease allows a direct estimate of the likelihood that his or her case was caused by a particular exposure.\textsuperscript{262} Rather, an individual’s combination of genetic makeup and exposure increases the likelihood of the disease relative to genetically similar individuals not exposed.\textsuperscript{263} The exposure, accordingly, is considered a risk factor for the disease. The courts that decided Allen, Rutherford, McGhee, and Fairchild all recognized that in light of causal indeterminacy, it is appropriate to treat proof of contribution to risk as proof of contribution to cause.

In Barker, however, the House of Lords adopted a different conceptual framework to explain its Fairchild holding. According to the Barker majority, the Fairchild defendants were held liable not because the tortious augmentation of a plaintiff’s risk was a contributing factor to the plaintiff’s harm of mesothelioma, but because the tortious exposure of a plaintiff to asbestos caused the plaintiff’s harm of increased risk of developing mesothelioma.\textsuperscript{264} This reformulation extended Fairchild considerably with respect to the harm element of a tort.\textsuperscript{265} With respect to causation

\textsuperscript{262} The reasoning process known as differential etiology (usually—though inaccurately—called differential diagnosis) may seem to approach a direct estimate of a probability of causation in an individual case. Differential etiology attempts to prove that but-for causation is more likely than not, by ruling out other possible causes of the plaintiff’s disease. It could be fully determinative if a set of deterministic causes of plaintiff’s disease were fully characterized (e.g., if an infection were caused only by a known virus or a known bacterium, and exposure to one were ruled out). More often, however, the disease in question is characterized by several risk factors and a residuum of disease incidence unaccounted for by any known risk factor. Ruling out some risk factors may allow a more refined estimate of risk, but that estimate still would be derived ultimately from population-based studies.

\textsuperscript{263} Other traits besides genetics and exposure may be pertinent, such as exposures to additional toxins, environmental factors (e.g., nutrition or weather), or personal factors (e.g., age).

\textsuperscript{264} See Barker v. Corus U.K. Ltd., [2006] UKHL 20, [2006] 2 A.C. 572, [35]–[36] (Lord Hoffmann), [59]–[62] (Lord Scott), [126] (Baroness Hale) (noting that the defendants in Fairchild were not found to have ultimately caused the harm but that they were liable for creating the risk).

\textsuperscript{265} As Lord Rodger noted in dissent, the majority took pains to limit the extension, without convincing rationale, to the context of multiple asbestos exposures and mesothelioma. Id. [85] (Lord Rodger).
WHEN CERTAINTY DISSOLVES INTO PROBABILITY

doctrine, however, it implicitly retreated to the comfortable confines of but-for: each material exposure to asbestos, tautologically, was a but-for cause of the increment of risk associated with that exposure.

Although the increased risk caused by a tortious toxic exposure could be considered sufficient harm to support a cause of action, reconceptualization of causation-in-fact as proposed here has advantages over recognizing liability for a new tort based on creation of risk alone. The latter would invite the objection that many negligent or otherwise wrongful acts create risk of harm, but tort law ordinarily does not, and practically

266. See id. [71] (Lord Rodger) ("[T]he majority of the House proceeded on the simple basis that the creation of a material risk of mesothelioma was sufficient for liability."). The Barker majority’s attempt to reframe Fairchild was unconvincing. The majority labeled the equation of risk creation with contribution to injury a “fiction” and cited passages from Fairchild rejecting reliance on legal fictions, but those passages rejected a different fiction—that proof of material contribution to risk permitted an inference of but-for causation. Compare id. [31]–[34] (Lord Hoffmann) (arguing that the Fairchild Court did not rely on the fiction that creation of a material risk constitutes material contribution to contraction of disease), with id. [80]–[83] (Lord Rodger) (countering that Lord Hoffmann’s view that Fairchild “did not proceed on the basis that a defendant who had created a material risk of mesothelioma was deemed to have caused . . . the disease” was true only of Lord Hoffmann’s own speech in Fairchild but not true of the Fairchild majority). This reframing was used to justify the conclusion that each defendant should be liable only severally for its proportionate share of risk contribution, rather than for the plaintiff’s full damages as United Kingdom law would otherwise have required. See id. [2], [31] (Lord Hoffmann), [60]–[62] (Lord Scott), [112]–[13] (Lord Walker) (concluding that if tort consisted of exposure to risk then several liability proportionate to risk would be appropriate). Barker held that fairness demanded that result. See id. [127] (Baroness Hale) ("It seems to me most fair that the contribution [that defendants in a Fairchild-type situation] should make [to plaintiff’s compensation] is in proportion to the contribution they have made to the risk of that harm occurring."). Parliament promptly reversed that policy judgment. Compensation Act, 2006, c. 29, § 3(1)(a), (c), (2) (U.K.). The Supreme Court, with some justices holding their noses, later adhered to the precedent that material contribution to risk was sufficient causal connection despite the statutory requirement of joint and several liability. E.g., Sienkiewicz v. Greif, [2011] UKSC 10, [2011] 2 A.C. 229 [167]–[68] (Baroness Hale). In the United States, risk contribution as a tort is the premise underlying some claims for medical monitoring. E.g., Donovan v. Philip Morris USA, Inc., 914 N.E.2d 891 (Mass. 2009). This Article takes no position on whether claims for medical monitoring should be permitted upon proof of toxic exposure even absent manifest illness.
could not, compensate every person exposed to such risks.\textsuperscript{267} The \textit{Barker} majority answered this objection, despite holding that the harm caused by tortious asbestos exposure was creation of a risk of mesothelioma, by holding that tortfeasors would be liable only if the harm of mesothelioma had materialized. It imposed this limit by assertion rather than by reasoning, however.\textsuperscript{268}

By contrast, if risk augmentation is recognized as a cause of injury within a model of probabilistic causal contribution, a

\textsuperscript{267} Another problem soon led the United Kingdom’s Supreme Court to repudiate the risk-as-harm formulation. A group of employers’ liability insurers balked at covering liabilities owed to employees who developed disease long after their workplace exposures to asbestos. Cf. Jeffrey W. Stempel, \textit{The Insurance Policy as Statute}, 41 \textit{McGeorge L. Rev.} 203, 209–10 (2010) (recounting history of similar litigation in United States). The insurers argued that their policies covered only liability for disease that became manifest during the policy period, not for disease that became manifest after the policy period even if it had been caused by exposure during the policy period. BAI (Run Off) Ltd. v. Durham, [2012] UKSC 14, 3 (Lord Mance). The justices unanimously rejected this argument. \textit{Id.} 76 (Lord Clarke) (“Like other members of the Court, I agree . . . [that] for the purposes of the . . . policies, mesothelioma is ‘sustained’ or ‘contracted’ when the process that leads to the disease is initiated” by exposure to asbestos). Lord Phillips, however, would have held nevertheless that the insurance was not triggered. He argued that \textit{Barker} correctly interpreted \textit{Fairchild} as imposing liability for risk creation alone, and that because no one could tell which employer’s fiber actually caused a plaintiff’s mesothelioma, liability based on risk creation did not satisfy the insurance policies’ requirements of causation and injury. \textit{Id.} 124, 134–35. The majority, however, disagreed. \textit{Id.} 65 (Lord Mance) (“[I]t is impossible, or at least inaccurate, to speak of the cause of action recognized in \textit{Fairchild} and \textit{Barker} as being simply ‘for the risk created by exposing’ someone to asbestos.”); \textit{id.} 82 (Lord Clarke) (\textit{Barker} “cannot have intended to hold, without more, that the basis of liability was the wrongful creation of the risk . . . because there would be no liability at all but for the subsequent existence of the mesothelioma.”) I am indebted to Sandy Steel for calling my attention to the insurance trigger litigation.

\textsuperscript{268} See \textit{Barker}, [2006] UKHL 20, [48] (Lord Hoffmann) (“Although the \textit{Fairchild} exception treats the risk of contracting mesothelioma as the damage, it applies only when the disease has actually been contracted.”); \textit{id.} [61] (Lord Scott) (“If, in the event, the victim does not contract the disease, no claim can be made for the trauma of being subjected to the risk.”); Chris Miller, \textit{Liability for Negligently Increased Risk: The Repercussions of Barker v. Corus UK (PLC)}, 8 \textit{Law, Probability & Risk} 39, 42 (2009) (emphasizing distinction between liability for creating risk of harm and liability for creating risk that eventuates in actual harm). Courts that accept “lost chance” doctrines similarly limit their application by pronouncement. See generally David A. Fischer, \textit{Tort Recovery for Loss of a Chance}, 36 \textit{Wake Forest L. Rev.} 605 (2001) (discussing possible limiting principles for application of lost chance doctrine).
plaintiff would still need to establish an injury.\textsuperscript{269} The mechanistic view of causation deems a cause-in-fact of that injury to be any necessary element of a sufficient causal set. A stochastic model of causation would imply that in any particular toxic tort case it is impossible to tell whether even a known risk factor fits that description. The causal set that matters, instead, is a set of factors that increased the probability of the plaintiff's injurious result. Each element of this set should be considered a cause-in-fact of the harm the plaintiff experienced.

Jane Stapleton has called such reasoning a “fiction” in asbestos cancer cases because, so far as is known, asbestos-related lung cancer or mesothelioma does not result only if aggregate asbestos exposure exceeds some threshold and is not made more severe by additional exposure. Thus, she reasoned, a tortfeasor who contributed part of the plaintiff's total exposure can in no factual way be said to have “caused” just a “part” of the plaintiff's disease.\textsuperscript{270}

Professor Stapleton’s careful distinctions among cumulative, threshold, and non-threshold mechanisms of toxic injury are valuable and informative. But probabilistic causal contribution is a fiction only in relation to assumed scientific\textsuperscript{271} and legal models.

\textsuperscript{269} Jamie Grodsky argued that injury itself could be reconceptualized to include detectable sub-clinical cellular or biochemical changes. Grodsky, supra note 110, at 1671–74. Although this Article takes no position on that suggestion, there is nothing about Professor Grodsky’s suggestion that is inconsistent with this Article’s proposal.


\textsuperscript{271} I do not mean to suggest that mesothelioma may be a cumulative disease like asbestosis or that asbestos can only cause mesothelioma after a threshold exposure is achieved. At bottom, however, the assertion that contributing-factor causation is a legal fiction depends on the scientific model that assumes that the interaction of one asbestos fiber with one cell causes mesothelioma. This may well be true, although, curiously, the Supreme Court of the United Kingdom asserted, without citation, that “[t]he single fibre theory has . . . been discredited.” Sienkiewicz, [2011] UKSC 10, 102 (Lord Phillips). It is clear, at least, that more than one molecular change is necessary for a mesothelium cell to become malignant. Huang, et al., supra note 147, at 180–81.
of causation. The deterministic legal model assumes that one and only one source of asbestos exposure should be treated as “the” cause of the plaintiff’s entire illness: that as a matter of historical fact, the physical tumor in the plaintiff’s body originated with a cellular alteration initiated by a particular defendant’s fiber, which could be identified if only science were omniscient. Even if reality matches this assumption, the counterfactual inference required by the but-for test does not necessarily follow. We cannot say that if only plaintiff had been protected from inhaling this one fiber, plaintiff would not have mesothelioma today, because it is quite plausible that if this fiber had not turned this cell malignant, some other fiber would have—or would have enabled another cell, already mutated multiple times, to evade the body’s defenses and take the last step to cancer. A probabilistic view of causation acknowledges this uncertainty. Moreover, the lack of omniscience alters the balance between historical fact and legal fiction. We can know as a fact that a particular defendant’s tortious exposure of the plaintiff to asbestos increased the risk that the plaintiff would develop cancer. We can make some type of estimate of that risk contribution as compared to other risk contributions. We cannot know, as a fact, which fiber is the one without which plaintiff’s

---

272. I say “defendant’s fiber” for ease of reference. The party responsible for the exposure in a given case might not be a defendant for any of a variety of reasons (insolvency, worker’s compensation bar, inapplicability of any theory of tort liability, inability to identify the source of an ambient exposure).

273. See Noel F.C.C. de Miranda et al., Role of the Microenvironment in the Tumourigenesis of Microsatellite Unstable and MUTYH-associated Polyposis Colorectal Cancers, 27 MUTAGENESIS 247, 247 (2012) (explaining that after accumulating genetic alterations, various clones of tumor cells compete against one another for growth and space). In theory, a similar argument could be made about any tort: how can we say that, if the pedestrian plaintiff’s leg had not been broken by impact with the negligently driven car, it would not have been broken by something else? The difference is that in the ordinary case no reason exists to suppose that at the time of plaintiff’s injury some other cause created a material risk of the same injury.
WHEN CERTAINTY DISSOLVES INTO PROBABILITY

303
disease would not have developed. To preclude recovery as a result would embrace a legal fiction that the defendant did no harm.

To think of causation in risk-creation terms requires a shift in the legal mindset, despite the centrality of risk creation to many parts of tort theory. Lawyers customarily think of risk as “a forward-looking concept” while “[c]ausation usually looks backwards.”274 But even in traditional counterfactual causal inference, the causal generalizations that support the inference derive from experience that associates the event we call a cause with the occurrence that is the result of interest. We do not speculate that the pedestrian plaintiff’s broken leg might have resulted from something other than impact with the negligently driven car because common experience tells us that such impacts break legs that otherwise normally remain intact. In other words, impact with vehicles is an extremely powerful risk factor for broken legs.275 The difference in toxic torts is that scientific indeterminacy makes the reasoning from ex ante risk creation to ex post injury causation both more explicit and more necessary. Despite the doubts some courts have expressed about causal proof based on population-based estimates of relative risk,276 such measures are relevant to individual cases—even though they do not directly measure the probability of causation in an individual case.277 Because such measures will continue to be the type of


275. The inference is strengthened because of the typical absence of competing strong risk factors that might have caused the broken leg. Yet I strongly suspect that even a plaintiff with a diseased leg that might spontaneously fracture at any time would be able to obtain a factual finding of causation, and certainly would be able to reach a jury.

276. This skepticism has manifested in various ways. In Sienkiewicz, some members of the Supreme Court of the United Kingdom questioned whether population-based relative risk data were at all useful to causal judgments about an individual case. Id. The Texas Supreme Court, by contrast, acknowledged that epidemiologic data could be relevant, but its skepticism led it to impose the doubling+ rule and other unrealistic requirements. See Merrell Dow Pharm., Inc. v. Havner, 953 S.W.2d 706, 719 (Tex. 1997) (justifying doubling+ rule and other requirements as “stri[k]ing a balance” allowing acceptance of epidemiologic proof).

evidence that science can provide, it is time for a corresponding probabilistic contributing-factor model of causation.

D. The Question of Substantiality

Probabilistic causal contribution begins with the idea that when causal evidence is population-based, the law should treat risk factors as contributing causal factors despite the inability to infer but-for causation. In the United States, however, the prevailing verbal formulation of causation in tort has been the “substantial factor” test of the Second Restatement.278 In light of the Third Restatement’s decisive rejection of “substantial factor” as vague and analytically inchoate, it may seem contrarian to propose a “contributing factor” concept of causation. Probabilistic causal contribution can avoid the biggest problems of “substantial factor” simply by jettisoning the requirement of substantiality.

That Allen v. United States relied on the then-current Second Restatement’s “substantial factor” formulation is hardly surprising, but this reliance exemplified some of the confusion that led to the next Restatement’s rejection of “substantial factor.”279 Allen held that to be deemed a “substantial factor” in causing a plaintiff’s illness, the defendant must have “significantly” increased that plaintiff’s risk.280 The court’s interpretation of “significance,” therefore, had a dispositive effect. In that regard, Allen considered but rejected adopting the doubling+ rule.281 Yet, apparently because of a heuristic view of “significant” risk contribution, the outcome of each plaintiff’s case

---

278. RESTATEMENT (SECOND) OF TORTS § 431(a) (1965).
279. See RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 26 cmt. j (“The substantial-factor test has not, however, withstood the test of time, as it has proved confusing and been misused.”).
281. Id. at 416–18.
closely mirrored what might have been expected with a doubling+ requirement.\textsuperscript{282}

Thus, Allen’s use of “substantial factor” served two purposes. First, the court treated “substantial factor” as an alternative test of factual causation that better fit the nature of the available evidence. Second, the court used “substantial factor” to narrow the alternative test it had created. The latter use, more policy judgment than factual determination, is problematic because it so closely tracks the doubling+ version of but-for causation, when doubling+ always was an inappropriate test of cause-in-fact and will become more so as risk characterization improves.

Rutherford, too, used the word “substantial” to describe the risk contribution that would be required for an exposure to be considered a cause of a plaintiff’s illness.\textsuperscript{283} Rutherford cautioned, however, that the “substantial factor standard . . . requir[es] only that the contribution of the individual cause be more than negligible or theoretical.”\textsuperscript{284} Similarly, throughout the development of the mesothelioma case law in the United Kingdom, only exposures creating “material” increases in risk have been considered causative,\textsuperscript{285} but in that context “material” simply means “not de minimis.”\textsuperscript{286} In an analogously difficult situation requiring proof of causation, a federal district court\textsuperscript{287} applied a “contributing factor” standard to a claim for damages arising under the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA).\textsuperscript{288} The government plaintiffs alleged that

\textsuperscript{282} See id. at 429–40 (describing the specific details and outcome of each plaintiff’s case).

\textsuperscript{283} See Rutherford v. Owens-Ill., Inc., 941 P.2d 1203, 1223 (Cal. 1997) (holding that “a substantial factor contributing to plaintiff’s . . . risk” would be deemed a “substantial factor in causing or bringing about the disease”).

\textsuperscript{284} Id. at 1219.


defendants were liable for damages for “injury to, destruction of, or loss of natural resources” caused by defendants’ releases of PCBs to New Bedford Harbor. The defendants claimed, however, that some releases had been “federally permitted” and were therefore exempt from liability. Commingled in the harbor, the PCBs from allegedly exempt and non-exempt releases could not be distinguished in situ. The court considered the question: what would the government need to prove to establish that the non-exempt releases caused injury?

The court held that the non-exempt releases would result in liability if shown to be “a contributing factor to an injury to natural resources.” The court specifically considered and rejected requiring a showing of a “substantial contributing factor,” noting appellate precedent holding that defendants could be liable—even jointly and severally liable—for CERCLA response costs despite a lack of proof that their contribution to hazardous substance releases had been “substantial.”

An injury to natural resources provides a useful analogy: it is essentially a toxic tort committed on an ecosystem. In that


290. *Acushnet*, 722 F. Supp. at 897; see also 42 U.S.C. § 9607(j) (2006) (“Recovery . . . for response costs or damages resulting from a federally permitted release shall be pursuant to existing law in lieu of this section.”).


292. Id. at 897–98.

293. Id. at 897.

294. Id. at 897 n.8 (citing *O’Neil* v. *Picillo*, 883 F.2d 176, 179 n.4 (1st Cir. 1989)). In *O’Neil*, the trial judge held that defendants, who asserted they had arranged for disposal of only a relatively small portion of the hazardous substances at a dump, were jointly and severally liable for the full cleanup costs. *O’Neil* v. *Picillo*, 682 F. Supp. 706, 730–31 (D.R.I. 1988); see also 42 U.S.C. § 9607(a)(4)(A) (creating a cause of action for recovery of government response costs). On appeal, defendants contended that Restatement (Second) of Torts § 433B required the government to prove that each defendant was “a ‘substantial’ cause of the harm.” *O’Neil*, 883 F.2d at 179 n.4. The court of appeals rejected that contention, observing that to require proof of substantial contribution would impose an “almost impossible task” on the government, in conflict with statutory objectives. Id.

295. The *Acushnet* court announced a legal standard for proof of causation but did not have to determine whether the evidence satisfied the standard. The particular harm allegedly inflicted on New Bedford Harbor—the concentration
context and in indeterminate-defendant toxic tort cases, courts have successfully eliminated the substantiality requirement for contributing factors. With only a sensible exclusion for *de minimis* contributions,296 “contributing factor” defined by increases in risk fits well with the causal world that toxicogenomics and molecular epidemiology are in the process of revealing.

**E. Discounted Recoveries**

The preceding discussion showed that a “contributing factor” standard of causation, based on risk contribution, can be used to implement a stochastic, probabilistic model of causation that better fits the realities of proof of specific causation in many toxic tort cases—even, perhaps especially, in light of the anticipated results of continued research at the genomic and molecular scales. This proposed standard will face an easily anticipated objection: that tortfeasors will overpay if held liable for full damages for all harms of which they created incremental risk, no

---

296. Rejecting causal attribution for *de minimis* risk contributions would reduce some of the complexity that a probabilistic causal contribution model could introduce to toxic tort litigation. It would take some care, however, to apply this sensible limit sensibly. A common scenario in CERCLA cases involves many contributors of small fractions of a hazardous substance release that, as a whole, requires response action. An analogous situation in toxic torts—many exposures, each of which adds a small increment of risk—is plausible and may become increasingly apparent with improved ability to characterize both risk and exposures. Eliminating liability for each contributor as *de minimis* would allocate the loss entirely to the victim. Thus, *de minimis* ultimately must be a matter of relative contribution rather than an arbitrarily fixed absolute proportion of risk contribution. See also infra notes 319–20 and accompanying text.
matter how small their (non-\textit{de minimis}) risk contribution. The objection impinges, in different ways, on both corrective justice and efficiency rationales for tort liability. And, in different ways, the objection may be overcome by a carefully designed system of discounted recoveries in proportion to the degree of risk created by the tortious exposure, a suggestion that has recurred in the literature in various forms since the early 1980s.\textsuperscript{297}

The corrective justice or fairness argument is that contributing factor causation would unjustly hold some defendants liable for illnesses they did not cause. That criticism tautologically depends on the deterministic model of causation and the assumption that specific causation is knowable. If specific causation is not deterministically knowable, but population-based data are available, the deterministic all-or-nothing model would allow two outcomes that are different from contributing factor causation. The first possible outcome would be to deny recovery in every case because measures of relative risk are insufficient to establish specific causation. That would offend corrective justice and fairness just as much as contributing factor causation, but in the opposite direction. The second possible outcome would be to take specific causation as true only if the relative risk exceeded some threshold (e.g., doubling\textsuperscript{+}). That, however, is internally inconsistent with the deterministic model’s insistence that each affirmative finding of causation represents a belief about what would probably have happened to the particular plaintiff had the plaintiff not been exposed. The doubling\textsuperscript{+} rule simply glosses over the fact that each group of exposed and sick plaintiffs—whether they prevail or not—necessarily includes some whose litigation outcomes would have been different but for the irreducible scientific indeterminacy.

The efficiency argument takes a population perspective from the outset, contending that contributing factor causation would make tortfeasors overpay—and thus be over-deterred—by holding defendants liable for an aggregate amount in excess of the total amount of harm the defendants could have prevented. Discounting recoveries to reflect risk contribution is, at least

\textsuperscript{297} For a thorough survey of prior proportional liability proposals, see Green, \textit{supra} note 185, at 357–70.
theoretically, a complete response to that objection. The claim that such discounting would produce optimum deterrence has figured prominently in arguments for proportional liability schemes.298

From the perspective of a probabilistic model of causation, discounted recoveries would be a form of causal apportionment that would cohere conceptually with the available evidence on specific causation in many toxic torts. Consider a relatively simple yet paradigmatic hypothetical case. The plaintiff worked for many years at a job that entailed daily inhaling vapors of some solvent. After the appropriate latency period, the plaintiff’s doctors diagnosed a particular cancer. Undisputed, peer-reviewed scientific research shows that: (1) cells of this cancer almost always have mutations in a particular set of genes; (2) DNA treated with the solvent in vitro is statistically significantly more likely than untreated DNA to contain mutations in at least some of those genes; (3) the frequency of this cancer among exposed individuals with a specific susceptible genotype is slightly, but statistically significantly, greater than among unexposed individuals of similar genotype; and (4) no other specific risk factors for this cancer have yet been identified, so the vast majority of cases are considered unexplained or idiopathic. This hypothetical evidence of both biological mechanism and statistical association should suffice to support a reasonable factual finding of general causation.

Assume that genetic testing showed that the plaintiff’s tumor has mutations in the usual genes and the plaintiff has the susceptible genotype. Under the contributing factor standard of probabilistic causal contribution, the evidence of specific causation would be sufficient for submission to a fact-finder. The fact-finder would be asked to determine by a preponderance of the evidence whether the exposure caused the cancer; cause would be defined as a non-trivial increase in plaintiff’s risk of developing cancer. Upon an affirmative answer, the fact-finder

would estimate, as a percentage, the extent to which the solvent exposure contributed to the plaintiff's risk of the cancer. The fact-finder would also determine the plaintiff's total damages under the jurisdiction's damages rules. The court would enter judgment for the determined percentage of the total damages.

The preceding hypothetical deliberately did not specify a value for the relative risk estimated by molecular epidemiologic investigation, in order to avoid the suggestion of automatic equivalence between observed relative risk and causal apportionment in a particular case (i.e., relative risk of 2 equates to 50% risk contribution, relative risk of 1.5 equates to 33.33% risk contribution, etc.). A fact-finder might rationally apply such reasoning, but also might, for various reasons, reach a different result. The biological assumptions implicit in such reasoning might be disputed. The significance and weight to be assigned multiple studies that reported different relative risk values might be disputed. The fact-finder might understand the reported relative risk(s) not as point values but as estimates of a parametric value as indicated by a statistical confidence interval, and might find a central tendency in overlapping confidence intervals of multiple studies. Or facts particular to the case at hand—characteristics of the plaintiff, the exposure, or both, as compared to the study subjects—might prompt the fact-finder to make a causal apportionment greater or less than implied by the study's relative risk. Adjustments for these reasons are appropriate and are, or should be, part of the fact-finding process under the but-for mechanistic model of causation as well.

The hypothetical can be made more complicated by assuming that two or more parties contributed to the solvent exposure. The analysis would depend on the nature of the multiple contributions and on what is known about the effect on risk of increasing amounts of exposure. A solvent manufacturer that failed to warn that its product should not be inhaled and a company that used the product without taking reasonable precautions to prevent inhalation, for example, would both be

299. See Greenland & Robins, supra note 277, at 325 (discussing, as an example, the independence-of-background assumption and explaining that many plausible disease mechanisms do not satisfy this assumption).
but-for causes of the same exposure under traditional causation analysis; therefore each should be treated as fully causally responsible for the share of damages attributed to the solvent exposure.300 By contrast, if the plaintiff was exposed to the same solvent in products made by two different manufacturers, the question of causal attribution between the manufacturers would resemble the mesothelioma indeterminate-defendant cases, with the important difference that the causal link between plaintiff’s cancer and the overall solvent exposure would be treated probabilistically.301 Assuming that the amount of risk created by the solvent exposure was somehow proportional to the amount of exposure,302 the logic of probabilistic causal contribution implies division of the solvent’s causal contribution in proportion to each source’s contribution of solvent exposure. In some cases this rule would invite fine parsing of whether multiple exposures truly involved the same agent; that assessment would best be tailored to the available evidence in particular circumstances.

The next complication adds other proven, non-*de minimis* contributing risk factors. As noted above, conceptually it does not matter whether the plaintiff’s exposure to additional risk factors required human agency or not, whether it was tortious, or whether it was created by a defendant or the plaintiff or a third party. If multiple risk factors apply to a plaintiff, a court’s first

---

300. This illustration assumes that all other requisites of liability are satisfied for both these parties. Applicable comparative fault and scope-of-liability rules would determine whether judgment would be entered against each tortfeasor for the full amount of this apportioned share.

301. In the mesothelioma cases, by contrast, it was not in dispute that asbestos, rather than something else, caused the disease.

302. This critical biological assumption might be well established in some cases (e.g., asbestos-mesothelioma) and in other cases might reasonably be treated as an acceptable approximation. It should be subject to factual dispute, however. For example, the available data might show any of the following: increased risk only after exposure exceeded a threshold amount; increased risk relatively insensitive to dose (perhaps after dose exceeded a threshold); increased risk in some way proportional to dose up to a limit beyond which increasing dose adds no increment of risk. Perhaps the most likely scenario, even in molecular epidemiology studies, would be insufficient data to assess the assumption, i.e. studies that simply compared relative risk among “exposed” and “unexposed” groups. In that scenario the reasonableness of the proportionality assumption could be a subject for expert debate.
instinct might be to ask the fact-finder to assign percentages to all of them, but strictly speaking it would not be necessary to quantify the proportion assigned to any individual non-defendant risk factor. The crucial assessment would be what causal share to attribute to each of the exposures for which any defendant was liable. The non-defendant risk factors could be treated, collectively, as part of that plaintiff’s “background” risk, although of course they could elevate that plaintiff's background risk value above the background risk faced by somebody not affected by any known risk factors.

Proof that the plaintiff had been exposed to multiple risk factors would present significant factual complications, however. If each risk factor independently added incremental risk, a fact-finder could relatively easily assign causal contributions by comparing the risk increments to one another. But risk contributions from diverse factors might not be independent of one another, and if they were not, they could interact in a variety of ways. 303 The best-known example is the synergistic effect of tobacco smoking and occupational asbestos exposure on the risk of lung cancer. 304 Exposure to both carcinogens increases lung cancer risk far more than the sum of the risk contributions of each acting alone. 305 Although some courts have tried to divide the additional risk resulting from the synergy between the two contributing factors, there is no entirely satisfying way to do so. 306 An alternative would be to assign the additional risk resulting from a synergistic interaction to each of the interacting risk factors. The latter approach would mean that, as compared to a

304. Huang et al., supra note 147, at 203.
305. Sanders, supra note 1, at 35.
306. See, e.g., Dafler v. Raymark Indus., Inc., 611 A.2d 136, 145–56 (N.J. Super. 1992), aff’d, 622 A.2d 1305 (N.J. 1993) (affirming a jury verdict that appeared to apportion the synergistic risk in proportion to the amount of risk added by exposure to each carcinogen separately and independently); see also V. McCormack et al., Estimating the Asbestos-Related Lung Cancer Burden from Mesothelioma Mortality, 106 BRIT. J. CANCER 575, 575 (2012) (“Quantifying the asbestos-related lung cancer burden is difficult in the presence of this disease’s multiple causes.”). But see Green, supra note 94, at 545 (“[V]arious plausible, yet complicated, methods for determining risk contribution exist . . . .”).
case of independent and additive risk factors, in a case of synergy the proof of plaintiff's exposure to an additional risk factor would have less effect on the causal share assigned to the risk factor contributed by defendant(s). 307

The quality of evidence concerning potential interaction of multiple risk factors will likely vary from case to case. In some cases the effect of combined exposure will itself have been the subject of scientific study designed to assess the independence or interaction of risk contributions. Even then, factual issues such as the extent of exposure to the various risk factors might affect the causal attribution. In other cases each risk factor might have been studied independently, with the attendant possibility of confounding by the other risk factor(s), leaving the experts to battle over whether an assumption of independence or of synergy best fit the available data. Toxicogenomic and molecular evidence about the mechanism of toxicity for the various risk factors could inform that debate.

In a probabilistic causal contribution model, a plaintiff seeking to prove a risk contribution need not exclude other risk factors, unlike a plaintiff using differential diagnosis in a deterministic causation model. Thus, defendants should be assigned the initial burden of production of evidence that plaintiff was exposed to additional risk factors and of general causation with respect to those risk factors. There would be no justification, however, for shifting the ultimate burden of persuasion: plaintiff must persuade the fact-finder by a preponderance of the evidence.

---

307. Susan Poulter used the asbestos-tobacco-lung cancer example to show how under this approach, if the relative risk jointly created by two exposures is the product rather than the sum of the relative risks of each acting individually, the fraction of cases attributed to each exposure in the population exposed to both is exactly the same as the fraction attributed to each in the population exposed only to it. See Poulter, supra note 303, at 233–34, n.215. (This handy result obtains only if the synergistic risk exactly equals the product of the individual risks, but applying this approach to any synergistic interaction will result in attributable shares closer to the individual shares than would be the case if the risks were additive.) Because this approach assigns overlapping parts of the synergistic risk to both exposures, the sum of the assigned risk fractions would exceed one. See also Susan R. Poulter, Genetic Testing in Toxic Injury Litigation: The Path to Scientific Certainty or Blind Alley?, 41 JURIMETRICS 211, 223–31 (2001) (describing possible interactions among genetic and environmental causes of disease).
that defendant increased plaintiff’s risk of disease with evidence that will permit the fact-finder to make some estimate of a percentage causal contribution (even if, as will usually be true, the scientific evidence does not produce a single point value for that contribution).

Undeniably, in these cases the proof, and the fact-finders’ task, will be complicated. One can imagine cases as complex as one cares to: exposures to many different risk factors, from tortious, non-tortious human, plaintiff, and natural sources, with each tortious exposure including contributions of multiple tortfeasors. A premise of this Article is that toxicogenomics and molecular epidemiology will make these cases more, rather than less, complex. Nonetheless, it is possible to exaggerate the concern about complexity. The adversary system, the human decision-making process, the reduced importance of additional risk factors in cases of synergy, and the de minimis limitation on risk factors that “count,” all will tend to focus litigation on the most salient tortious conduct and alternative causes. In any event, complexity of the facts is not a good excuse for oversimplification of the law.

Here, too, an analogy from CERCLA is instructive, demonstrating that causal apportionment is feasible even if factually difficult. A defendant liable for the government’s response costs incurred under CERCLA is jointly and severally liable, unless the defendant shows, on a purely causal and not equitable basis, that the environmental harm to which it contributed is reasonably capable of apportionment. The statute makes liable several qualitatively different classes of actors. Information about their relative contributions is often

308. Burlington N. & Santa Fe Ry. Co. v. United States, 556 U.S. 599, 613–14 (2009) (following RESTATEMENT (SECOND) OF TORTS § 433A (1965)). The decision expresses what is required to justify apportionment in at least three ways: whether the harm is “capable of apportionment,” whether “a reasonable basis for apportionment exists,” and whether the harm is “a single, indivisible harm.” Id. at 1881. The Court emphasized that the analysis is purely causal, invoking none of the equitable factors courts use to allocate responsibility among jointly and severally liable parties. See id. at 1882 n.9; United States v. Hercules, Inc., 247 F.3d 706, 718–19 (8th Cir. 2001); United States v. Twp. of Brighton, 153 F.3d 307, 318–19 (6th Cir. 1998).

scant. Apportionment may require substantial amounts of assumption and inference. Nevertheless, the Supreme Court upheld a trial court’s use of data that were extremely limited—in type, in clarity of causal significance, and in precision—to apportion liability.310

The Maryland Court of Special Appeals required causal apportionment in another different yet analogous context. The plaintiffs in CSX Transportation, Inc. v. Bickerstaff311 sued their employer railroad under the Federal Employers’ Liability Act (FELA).312 They alleged that they suffered osteoarthritis of the

including the owner and operator of a vessel or facility, any person who owned or operated a facility at the time of disposal of a hazardous substance, any person who arranged for disposal, treatment, or transport for disposal or treatment of hazardous substances at a facility owned by another person, and any person who accepted substances for transport to a disposal or treatment facility selected by that person).

310. See Burlington Northern, 556 U.S. at 619 (concluding that “the District Court reasonably apportioned the Railroads’ share of the site remediation costs at 9%”). The trial court apportioned liability among three parties that owned or operated a facility when hazardous substances were disposed of there: an agricultural chemical distributor that spilled toxic chemicals during its operations and two railroads that jointly owned a piece of property that the distributor leased. Id. at 1874–76. By the time the United States sued to recover more than eight million dollars in response costs, the distributor was defunct. Id. at 1876. The railroads’ parcel was only a portion of the land the distributor used, was used for only a portion of the time during which the distributor operated, and was used to handle only two of the three chemicals that were the subject of the cleanup. Id. The district court concluded that the product of these three fractions represented a reasonable apportionment of the harm, and then increased the computed share by a 50% fudge factor to account for uncertainty. Id. at 1882. The Supreme Court acknowledged that the district court had received little evidence on the apportionment issue, that one could question the assumption that the harm was proportional to geographic area and years of operation, and that little record evidence supported the conclusion that the two chemicals handled on the railroads’ parcel were responsible for two-thirds of the harm. Id. at 1882–83. Nonetheless, the Court upheld the apportionment as reasonable with only one justice dissenting. Id.; see also United States v. Bell Petrol. Servs., 3 F.3d 889, 903 (5th Cir. 1993) (reversing judgment of joint and several liability because, although it was “not possible to determine with absolute certainty the exact amount of chromium each defendant introduced,” the evidence sufficed for “reasonable and rational approximation of each defendant’s individual contribution to the contamination”).


knee because CSX negligently used mainline ballast—relatively large rocks—as a walking surface in the rail yards where they worked. The trial court instructed the jury on comparative negligence, and the jury returned verdicts assigning some fault to the plaintiffs but larger shares to the defendant. The defendant contended that in addition to the parties’ negligence, “other significant causes such as obesity, smoking, other pre-existing medical conditions, and age,” as well as genetics, had contributed to causing the plaintiffs’ injuries. The appellate court agreed that the trial court should have treated such alleged causes separately from the plaintiffs’ own negligence.

On appeal, the defendant did not argue that any of these other factors was a but-for cause of plaintiffs’ injuries and the ballast was not. Rather, CSX relied on evidence that osteoarthritis is “multifactorial, meaning that there are many factors involved in contributing to it.” A defense expert had defined a “risk factor” as “something that would cause or lend someone to have a certain problem or condition.” The appellate court treated these risk factors as concurrent causes and held that the trial court should have instructed the jury to apportion any fraction of damages the jury found “attributable” to these “other causes and factors.”

So, in addition to the House of Lords’s application of causal apportionment to asbestos indeterminate-defendant cases in Barker, American courts have held that causal apportionment

313. Bickerstaff, 978 A.2d at 770.
314. Id. at 770–71, 793.
315. Id. at 793.
316. A defense expert testified that a “small percentage of osteoarthritis” is hereditary, and argued (without presenting genomic evidence) that one plaintiff’s inherited bowleggedness was a significant cause of that plaintiff’s injury. Id. at 797, 799.
317. Id. at 799.
318. The trial judge seemed to think this the only proper use of evidence of such non-negligent causes. See id. at 794–95 (summarizing the transcript of the trial judge’s ruling).
319. Id. at 798 (quoting plaintiffs’ expert witness).
320. Id. at 797 (quoting defendant’s expert witness).
321. Id. at 799 n.22.
322. Some American courts have used comparative fault to achieve results
can work in a CERCLA case in which multiple defendants were liable in different ways and in a FELA case in which plaintiffs bore non-negligent risk factors for their injuries. These examples show that causal apportionment is a practicable approach that has received judicial imprimatur, even when facts are complex and data limited. Causal apportionment, via a probabilistic causal contribution model as described here, can work in toxic tort cases more generally. Although imperfect, it is the best available solution to the problem of liability disproportionate to the contribution to risk.

To apply probabilistic causal contribution correctly, however, courts will need to take care to avoid duplicate reductions of plaintiff’s damages through comparative responsibility regimes.

similar to Barker’s. See, e.g., Barnes v. Owens-Corning Fiberglas Corp., 201 F.3d 815, 822–23 (6th Cir. 2000) (applying Kentucky comparative fault statute); Green, supra note 185, at 353 (distinguishing apportionment based on relative culpability from determining and apportioning liability based on probabilistic assessments).

Each of these holdings may be criticized for undermining a joint and several liability scheme that fosters core objectives of the statute involved. See Norfolk & W. Ry. Co. v. Ayers, 538 U.S. 135, 165–66 (2003) (holding that under FELA, “an employee who suffers an ‘injury’ caused ‘in whole or in part’ by a railroad’s negligence may recover his or her full damages from the railroad”); United States v. NCR Corp., No. 12-2069, 2012 WL 3140191, at *8 (7th Cir. Aug. 3, 2012) (agreeing with defendants that under Burlington Northern, “apportionment calculations need not be precise. To the contrary, the [Supreme] Court upheld a district court’s rather rough, sua sponte calculation of apportionment. . . . But we do not agree that . . . lower courts must always take such an approach”). This Article takes no position on whether Burlington Northern and Bickerstaff correctly applied CERCLA and FELA, respectively.

See, e.g., Loui v. Oakley, 438 P.2d 393, 396–97 (Haw. 1968) (requiring instruction that if the jury “is unable to determine by a preponderance of the evidence how much of the plaintiff’s damages can be attributed to the defendant’s negligence” in causing one of two entirely separate auto accidents that injured plaintiff, the jury “may make a rough apportionment”); Campione v. Soden, 695 A.2d 1364, 1375 (N.J. 1997) (holding that a comparative negligence statute required apportionment of damages for injuries caused in part by each of two auto accidents occurring in rapid succession, although the “absence of conclusive evidence concerning allocation . . . will necessarily result in a less precise allocation”). See generally RESTATEMENT (THIRD) OF TORTS: APORTIONMENT OF LIA, § 26 cmt. h (discussing the difficulty of apportioning harm on causal grounds). The Restatement endorsed placing the burden of proof on a party seeking to avoid responsibility for an entire injury, but relaxing such a party’s burden of production of evidence. Id.
Consider three hypothetical workers with lung cancer, each of whom was exposed to an asbestos product made by a manufacturer that negligently failed to warn of the product’s dangers. Each worker contributed to his or her elevated risk of lung cancer in a different way. Worker #1 negligently failed to use a protective respirator provided by the worker’s employer. Worker #2 negligently inhaled fibers from a similar product (made by an unidentifiable manufacturer) during a home renovation project. Worker #3 negligently smoked cigarettes. Worker #1 would be causally responsible for the same increment of lung cancer risk as the manufacturer; the only apportionment between them would be based on comparative fault. Worker #2 would be causally responsible for an apportioned share of the fraction of lung cancer risk causally attributed to asbestos exposure; after the causal apportionment, further reduction in the plaintiff’s damages based on a share of fault would be inappropriate, as the plaintiff bore no fault for the portion of risk attributed to the manufacturer. Worker #3, for the same reason, would recover damages subject to causal apportionment (with appropriate treatment of the asbestos/tobacco synergy), but not further reduced by comparative fault.

Critics of apportionment (whether causal or fault-based) argue that joint and several liability appropriately places the risk that some tortfeasors will be insolvent on other tortfeasors rather than on injured plaintiffs. There is much to this argument, particularly as applied to situations like asbestos, where the total harm caused was so great that many liable parties entered bankruptcy. In the controversy over specific causation in toxic torts, however, the real choice frequently will not be between joint and several liability for full damages or only several liability for apportioned damages—it will be between liability based on probabilistic causal contribution and no liability at all based on deterministic models of but-for causation. Moreover, as the


326. Sanders, supra note 1, at 15.
variety of state comparative fault regimes demonstrates, acceptance of apportionment need not dictate the allocation of the risk of insolvency.\textsuperscript{327} Courts or legislatures could, despite the tension with the theoretical rationale for probabilistic causal contribution, make the policy choice that a tortious risk contributor should bear all or part of the portion of risk contributed by another tortfeasor that is insolvent. That choice would make a party liable for a risk the person did not cause, but not necessarily for a harm the person did not cause (which cannot be determined).

F. Boundaries

Commentators on the components of probabilistic causal contribution—treating risk contribution as causation and discounting recoveries in proportion to relative contribution of risk—often question whether, once accepted in one class of cases, these doctrines can sensibly be restricted to that class.\textsuperscript{328} Limiting doctrinal reform to asbestos mesothelioma cases, for example, has been criticized as arbitrary.\textsuperscript{329}

I propose probabilistic causal contribution as a response to a particular problem of causal indeterminacy that will be exacerbated rather than solved by scientific advances. Although it is tempting to say that if the concept is appropriate in some cases it could be applied in every case, the boundaries of the solution need not extend beyond the boundaries of the problem.\textsuperscript{330}

\textsuperscript{327} See \textit{Restatement (Third) of Torts: Apportionment of Liab.} § 10 cmt. a (describing five variously adopted "tracks").

\textsuperscript{328} See Jonathan Morgan, \textit{Causation, Politics and Law: The English—and Scottish—Asbestos Saga}, in \textit{Perspectives on Causation}, supra note 1, at 57, 64–65 (despairing that judicially-imposed limits can withstand "common law's ineluctable method of reasoning by analogy"); see also Sanders, supra note 1, at 39 (expressing concern that the scope of risk rule could be difficult to limit).

\textsuperscript{329} See Sanders, supra note 1, at 36 (expressing concern that the scope of the risk rule could be difficult to limit).

\textsuperscript{330} See generally BIC Pen Corp. v. Carter, 346 S.W.3d 533, 545 (Tex. 2011) (distinguishing toxic tort cases and refusing to allow plaintiff to apply doubling+ rule and statistical evidence of increased risk to support claim that defect in cigarette lighter caused injury).
And the boundaries of the problem can be defined, if not precisely, then reasonably well.331

Once again a metaphor from the physical sciences is helpful. At a quantum scale, the behavior of elementary particles is inherently random and unpredictable.332 But the behavior of millions of particles in a beam is not.333 Quantum mechanics did not render Newtonian physics useless for describing the motion of billiard balls. But a probabilistic view is required to describe the motion of the electrons of the atoms of which those balls are made.334

Causation is always an inference, and all evidence of causation can be conceived as probabilistic.335 But in most cases a fact-finder can comfortably fit that evidence to an inferential process grounded in the deterministic model, deciding whether the proof leads it to a level of belief in but-for causation that satisfies the standard of persuasion.336 The mechanistic model fails when proof of causation rests on evidence derived from

331. Probabilistic causal contribution would hardly be the first doctrinal adjustment to recognized difficulties of proof. Professor Green gave as examples res ipsa loquitur, market share liability, the rationale for strict products liability, see Escola v. Coca-Cola Bottling Co. of Fresno, 150 P.2d 436, 461–63 (Cal. 1944) (Traynor, J., concurring) (citing plaintiff's inability to refute evidence of due care as grounds for strict liability), and the sufficiency standard for the magnitude of damages caused by a defendant's wrongdoing. Michael D. Green, Pessimism About Milward, 3 WAKE FOREST J.L. & POL'Y (forthcoming 2013) (on file with author).

332. RUVINSKY, supra note 178, at 4–6.

333. Id. at 6.

334. See id. at 4 (explaining that Heisenberg's uncertainty principle means that position and speed of elementary particles “is always characterized by a probability distribution”).

335. See Rosenberg, supra note 298, at 870 (suggesting that particularistic evidence is no different than statistical evidence in that it “offers nothing more than a basis for conclusions about a perceived balance of probabilities”); see also Gold, supra note 277, at 384 n.42 (agreeing with Rosenberg that “all evidence . . . involves inference from observed probability patterns,” but acknowledging the “power of particularistic proof to generate belief probabilities regardless of known fact probabilities”).

336. In a civil case “preponderance of the evidence” is routinely described as proof leading the fact-finder to believe that the fact is more likely than not, but Professor Wright has argued that to find a fact proven, legal fact-finders actually require a degree of belief much higher than belief that the fact and its negative are equally likely. Wright, supra note 19, at 201–02.
population-based data on the association of disease and exposure, or of disease and genotype and exposure, or of exposure and disease-related biomarkers. In such cases the fact-finder must test its belief in a frequentist-probability value supported by evidence of risk contribution. Probabilistic causal contribution is therefore the appropriate model to apply. Advances in toxicogenomics and molecular epidemiology will make such evidence available more often than it has been in the past.

Sometimes, of course, such evidence will not be available although other evidence exists that might support an inference of general causation. To require relative risk data as proof of specific causation in every case is to ask too much. Compelling cases may exist in which such data are unavailable for perfectly good reasons. Zuchowicz v. United States, in which epidemiologic study was implausible because both the disease and the exposure were exceptionally rare, is an example. Or a plaintiff may have adequate mechanistic evidence of general causation, as in Milward v. Acuity Specialty Products Group, Inc. Most likely the best that can be done in such cases is to muddle through under the all-or-nothing, but-for causation rule, as courts have been doing. It is ironic that a plaintiff who succeeds without relative risk data would receive a full recovery while a plaintiff with “better” evidence would recover only a discounted award, but then again, the absence of relative risk data would in most cases make success substantially more difficult for the plaintiff.

337. Zuchowicz v. United States, 140 F.3d 381 (2d Cir. 1998).
338. Id. at 384–85, 391 (affirming a verdict that negligent prescription of endometriosis drug at a dose that “very, very few women” had ever received had caused plaintiff’s “very rare” disease).
339. See Milward v. Acuity Specialty Prods. Grp., Inc., 639 F.3d 11 (1st Cir. 2011), cert. denied, 132 S. Ct. 1002 (2012) (reversing summary judgment for defendant because the trial court abused its discretion in excluding plaintiff’s expert testimony on general causation). Because general causation was addressed before other issues in the case, id. at 13, the decision does not address how plaintiffs intended to prove specific causation.
340. See generally Sanders, supra note 20, at 1375–80 (discussing admissibility standards for specific causation testimony in toxic tort cases in which the plaintiff presents adequate evidence of general causation).
V. Probabilistic Causal Contribution and Tort Goals

Toxicogenomics and molecular epidemiology will strengthen the argument for reenvisioning specific causation in toxic tort cases, but the indeterminacy at the core of that argument was understood even in the pre-genomic era. Nevertheless, American courts, as in June and Wilcox, have generally continued to insist on rules that assign all-or-nothing liability depending on whether there is more-likely-than-not proof of but-for causation. Michael Green has noted that this insistence seems to buck two tides: a general trend away from an all-or-nothing rule in tort law and a "striking consensus" among academic writers (from which Professor Green dissented) in support of proportional liability. Although courts have not typically explained their resistance by reference to the goals of tort law recognized by scholars, it is worth assessing probabilistic causal contribution against those goals.

A. Deterrence Goals: Is Probabilistic Precision Probable?

Legal economists have argued that using probabilities to discount recoveries, across a population of exposed and sick claimants, would produce an economically efficient deterrence signal in which the aggregate amount of awarded damages matched the value of the harm actually caused. Professor

341. The view in other countries is less monolithic. See Ken Oliphant, Uncertain Factual Causation in the Third Restatement: Some Comparative Notes, 37 WM. MITCHELL L. REV. 1599, 1624–30 (2011) (comparing the tort principles of several European nations).

342. Green, supra note 185, at 352 (noting that modern tort law is “more receptive” to apportioning liability); id. at 357 (scholarly consensus); id. at 397 (“[T]here is good reason for the courts’ unwillingness.”); see also Michael D. Green, Introduction: The Third Restatement of Torts in a Crystal Ball, 37 WM. MITCHELL L. REV. 993, 995 (2011) (saying he has been “critical of the proportional liability literature”).

343. See, e.g., Green, supra note 185, at 354–55 (noting that proportional liability to acknowledge probabilistic uncertainty in fact-finding, “the argument goes, would, through more accurate outcomes, provide more fine-tuned deterrence incentives”); Mario J. Rizzo & Frank S. Arnold, Causal Apportionment in the Law of Torts: An Economic Theory, 80 COLUM. L. REV. 1399, 1427–28 (1980); Rosenberg, supra note 298, at 867.
Green questioned how well epidemiologic reality would match economic theory. He pointed out that when epidemiologists repeatedly investigate a suspected association between exposure and disease, the resulting relative risk values generally vary considerably. This real-world variability derives in part from random sampling error but also from the inherent difficulty of epidemiologic study design. With the possible exception of very large and expensive studies, any two epidemiologic investigations are likely to be conducted in ways that create different biases and confounding factors that can affect the results. Because the notion of a single, knowable quantitative measure of risk contribution is illusory, Professor Green noted, so too is the notion of an accurately modulated optimum deterrence signal based on risk contribution.344

The improbability of probabilistic precision is indisputable. Significant variation in measured relative risk characterizes even some of the strongest and most accepted substance-disease links uncovered by classical epidemiology.345 Even if repeated studies could be designed to be perfectly comparable in avoiding bias and confounding, sampling error would frustrate any hope of replicating relative risk values. For claims involving rare diseases, uncommon exposures, or small toxic effects, an epidemiologic study’s a priori statistical power will be relatively low, and the statistically computed sampling error relatively high.346

345. See, e.g., W.C. HUEPER, OCCUPATIONAL AND ENVIRONMENTAL CANCERS OF THE URINARY SYSTEM 118–19, 156 tbl.48 (1969) (describing studies reporting relative risk of bladder cancer ranging from 8.7 to 17 in aniline dye industry workers). Estimates of the fraction of mesothelioma cases that occur absent known exposure to asbestos range from less than ten percent to upward of thirty percent. Compare Mark A. Behrens, What’s New in Asbestos Litigation?, 28 REV. LITIG. 501, 527 (2009) (asserting that “there is wide agreement that a significant number (by some estimates, twenty to thirty percent) of mesotheliomas are not asbestos-induced”), with Troyen A. Brennan, Environmental Torts, 46 VAND. L. REV. 1, 15 (1993) (noting that “well over ninety-percent” of mesothelioma deaths are attributable to asbestos exposure). See also Becker v. Baron Bros., Coliseum Auto Parts, Inc., 649 A.2d 613, 618 (N.J. 1994) (reporting that plaintiff’s witness testified that fifteen percent of cases “have no known cause,” and defendant’s witness testified that twenty to forty percent of cases have “unknown causes”).
Shifting epidemiologic analysis to the molecular level will improve this situation. In past epidemiologic studies, undetected genetic differences contributed to the variability of results (because the research subjects were not all equally susceptible to the toxic effect under study or not equally at risk for the disease even without exposure). Assessing relative risk by genotype will reduce that variability. Biomarkers of exposure will enable finer calibration of the very rough distinctions between “exposed” and “unexposed” samples to which epidemiologists have previously been limited. Biomarkers of effect or pre-clinical biomarkers of disease (even if not associated with toxic exposure) will similarly allow more precise delineations.

Nonetheless, toxicogenomics and molecular epidemiology will not produce quantitative precision for the legal system’s convenience. Their results will inherently be probabilistic, statistical, and variable. Molecular epidemiologists will be able to account for more causes of variation than classical epidemiologists could, but they will not be able to control for all sources of bias and confounding. The legal system must understand that neither classical nor molecular epidemiology is likely to satisfy a norm of mathematical precision.

But where does that norm come from? Why should it be the norm? And why does it apply to causal findings but not to other elements of a tort claim?

To a large extent, the precision norm arises directly from the economic efficiency rationale. Critics ask: If optimal deterrence is the sole rationale for abandoning all-or-nothing causation rules, how can toxicogenomics and molecular epidemiology be expected to produce such precision? The norm of mathematical precision has its roots in the economic theory of deterrence. Critics ask: If optimal deterrence is the sole rationale for abandoning all-or-nothing causation rules, why should the legal system expect toxicogenomics and molecular epidemiology to provide such precision?


347. See Marchant, supra note 53, at 7–8 (“[E]ach person is unique in his or her susceptibility to toxic agents, further complicating the inquiry into what caused illness in that individual.”).

348. See id. at 18–27 (discussing potential use of biomarkers to prove or disprove exposure, as well as general and specific causation); Grodsky, supra note 85, at 181–87 (defining biomarkers of exposure, susceptibility, and effect).

349. Paolo Boffetta, Biomarkers in Cancer Epidemiology: An Integrative Approach, 31 CARCINOGENESIS 121, 125 (2010) (asserting that molecular epidemiology studies “fit into the same framework” as classical epidemiology).
but optimization is impossible because of inadequate data, why forgo the perceived accountability advantages of traditional *sine qua non* causation?²³⁵⁰

The argument that probabilistic precision is illusory, however, applies also to factual determinations under the but-for test. This is especially true in jurisdictions that apply bright-line doubling+ relative risk standards for specific causation.²³⁵¹ Even absent bright line rules, the all-or-nothing but-for causation model forces a fractional statistical probability to either 1 or 0. As Professor Green acknowledged, no one can say that those errors balance each other out.²³⁵² Probabilistic causal contribution may not be able to discriminate finely between, say, a 40% risk contribution versus 35% or 45%. But even with limited accuracy, its results would more closely match the available risk evidence than the all-or-nothing test of but-for causation. It would avoid the under-deterrence problem that results when claims against tortious exposures to known risk factors fail because the but-for test cannot be satisfied, and it would provide some compensation and accountability in such cases.²³⁵³ At the same time, it would

---

²³⁵⁰. See Green, *supra* note 185, at 395–96 (doubting that proportional liability would enable “tort law to provide a liability signal precisely calibrated to the costs of accidents worth avoiding”); Klein, *supra* note 53, at 44 (asserting that although class-based, proportional liability schemes theoretically might provide optimal deterrence, Agent Orange experience suggests otherwise).

²³⁵¹. The arbitrary effects of the doubling+ rule have been felt in real cases. See, e.g., Merck & Co. v. Garza, 347 S.W.3d 256, 267 (Tex. 2011) (rejecting plaintiff’s expert’s reliance on epidemiologic study with reported relative risk of 1.92); Estate of George v. Vt. League of Cities & Towns, 993 A.2d 367, 382 (Vt. 2010) (affirming the exclusion of plaintiff’s expert testimony when epidemiologic results varied, some relative risks exceeded 2.0, and plaintiff’s expert testified to “summary risk estimate” of 1.51).

²³⁵². Green, *supra* note 185, at 396. The net effect of all-or-nothing causation in toxic tort cases is a subject of debate. Compare Estate of George, 993 A.2d at 396 (“[T]here is reason to believe that the current system does not systematically underdeter.”), with Lin, *supra* note 110, at 1442 (stating that causation proof problems and litigation costs cause “systematic undercompensation of environmental tort victims and the systematic underdeterrence of polluters”).

²³⁵³. Professor Green also argued that epidemiology cannot detect relatively small increases in risk as accurately as needed for proportional liability, and questioned the legal system’s ability to assess the causal significance of a weak association between exposure and disease. Green, *supra* note 185, at 375–87, 392–95. Molecular epidemiology and toxicogenomics are likely to alleviate
reduce the over-deterrence of the “all” part of all-or-nothing rules. 354

Large swaths of tort law are, to put it charitably, highly approximate. It is somewhat ironic to criticize causal apportionment of damages as imprecise when the amount of damages itself contains a large heuristic component and when significant components even of seemingly objectively calculable economic damages are computed by informed probabilistic estimates. Furthermore, comparative fault regimes that divide up liability and thus affect deterrence signals have become nearly universal. Of course, apportioning among liable parties “on the basis of a normative judgment about relative culpability or responsibility is different from employing probabilistic assessments of the existence of an element of a prima facie case to determine the liability of a party and the damages for which the party is liable.”355 But both processes ask fact-finders to divide liability for a harm into pieces. Probabilistic causal contribution, although an imprecise fraction based on imprecise data, would at least be based to some degree on quantitative evidence. It is incongruous to let fact-finders hear all the evidence about the parties’ conduct and then carve up the liability based

(though not completely eliminate) the problem of weak associations by disaggregating them into relatively strong associations for susceptible individuals and very weak or nonexistent associations for resistant individuals. Ofer Shpilberg et al., The Next Stage: Molecular Epidemiology, 50 J. CLINICAL EPIDEMIOLOGY 633, 637 (1997); see also Perera, supra note 149, at 608 (noting that “new molecular epidemiologic and other data invalidate the assumption of population homogeneity” with respect to toxic susceptibility). The need to decide whether an epidemiologic association is causal, which addresses general rather than specific causation, arises in every case in which a plaintiff relies on epidemiologic evidence, even under but-for causation. The decision requires consideration of the “aspects” of an association that epidemiologists use to assess the causal significance of an association, only one of which is the association’s strength. See Hill, supra note 18, at 295. Probabilistic causal contribution would be applied to specific causation only if the evidence were sufficient to support an inference of general causation.


355. Green, supra note 185, at 353.
on a normative judgment, but prevent them from using relative risk as an apportionment tool because epidemiology is imprecise.

In the CERCLA context, the Supreme Court accepted the division of millions of dollars of liability based on very indirect measures of causal contribution. Tort law tolerates the imposition of full liability based on a belief that each element of the claim is “more likely than not.” It tolerates the extremely rough justice of comparative fault. Why should it demand a higher level of accuracy for causal apportionment? Getting it right is important, but the presence of scientific evidence should not fool us into thinking the law can or must get it exactly right—especially when the alternative is acknowledged over-deterrence in some cases and under-deterrence in many others. Detailed toxicogenomic and molecular information, to the extent available, will improve a fact finder’s ability to assign a causal share to a defendant’s toxic substance as opposed to all other contributors to the risk of a plaintiff’s disease. We should recognize this as an improvement—without feeling a need to pretend that such information will allow courts to impose theoretically perfect amounts of liability.

---

356. See Wright, supra note 325, at 1144–45 (stating that comparative responsibility is assessed “not according to any detailed formula but rather through rational commonsense judgment”).

357. Burlington N. & Santa Fe Ry. Co. v. United States, 556 U.S. 599, 612 (2009) (affirming liability allocation under CERCLA even though “the evidence adduced by the parties did not allow the court to calculate precisely the amount of hazardous chemicals contributed by the Railroad parcel to the total site contamination or the exact percentage of harm caused by each chemical.”).

358. Martin v. Owens-Corning Fiberglas Corp., 528 A.2d 947 (Pa. 1987), illustrates the tensions inherent in all-or-nothing rules. The court rejected a jury verdict apportioning a plaintiff’s disability between his asbestos exposure and his tobacco smoke exposure. Id. at 950–51. Two dissents argued that apportionment should have been permitted using “rough approximation,” id. at 951 (Nix, J., dissenting) without requiring “[m]athematical exactitude,” id. at 954 (Hutchison, J., dissenting). The first complained that apportionment was essential to avoid the “unconscionable” result of no recovery (presumably because the plaintiff’s own actions vitiated but-for causation by the defendant). Id. at 951 (Nix, J., dissenting). The second complained that apportionment was essential to avoid the “unjust” result of full liability (presumably because plaintiff’s and defendants’ acts would be treated as multiple sufficient causes). Id. at 954 (Hutchison, J., dissenting).
Another deterrence-based critique argues that causally-based apportionment of damages will provide inadequate deterrence because defendants will have incentives to hunt for extra risk factors to blame—and finding them would, perversely, reduce the liability of each wrongdoer as more wrongdoers (or at least risk factors) are found. Yet defendants have every incentive to conduct this hunt under an all-or-nothing but-for model of causation as well. The difference is that under the current regime a defendant that successfully posits additional risk factors as independent alternate causes has a good chance of avoiding liability altogether, while under probabilistic causal contribution the defendant would bear at least some liability even if many cases of the plaintiff’s disease have no known cause. Probabilistic causal contribution would also shift the onus with respect to alternative risk factors. Under current law, plaintiffs often must attempt to “rule out” other possible causes to leap the “more likely than not” hurdle and obtain a finding of but-for causation. Under probabilistic causal contribution, if a plaintiff proves that the exposure for which a defendant is responsible increased the plaintiff’s risk, a defendant seeking to use another risk factor to reduce its own liability would have to justify that reduction by showing that the plaintiff was exposed to the other risk factor and consequently faced a greater risk than would have been the case with only the exposure for which the defendant was responsible.

Moreover, the premise of the under-deterrence argument, that the tortfeasor caused the whole injury but is being held liable for only part of it, fits poorly with the reality of

359. See, e.g., David A. Fischer, Proportional Liability: Statistical Evidence and the Probability Paradox, 46 Vand. L. Rev. 1201, 1211–14 (1993) (arguing that the “probability paradox” means that “probabilistic causation requires people to use the least care when the world is the most dangerous”).

360. If the evidence shows that the risk factors interact to create risk, the defendant is less likely to prevail under an all-or-nothing regime but also likely to pay a larger share under probabilistic causal contribution. See supra note 273 and accompanying text. And in an all-or-nothing regime, even in the case of a synergistic interaction, a defendant might be able to convince a fact-finder that only the other risk factor was a “more likely than not” cause, depending on the strength of the interaction and the risk characteristics of the individual exposures.
indeterminate specific causation. Because that indeterminacy is the rationale for probabilistic causal contribution, exposure to multiple independent risk-contributing factors should matter: the additional risk factors reduce the relative role of a defendant’s risk factor. The deterrence signal of a discounted recovery, even if less than that of a full recovery, would still be higher—and therefore theoretically would produce greater investment in safety—than the signal that would result if but-for causation rules precluded any recovery. And if the hunt for alternative risk factors leads to legitimate scientific research, it will help society learn the truth and perhaps make better decisions about what products create risks that are not worth their benefits and vice versa.

361. David Fischer’s contrary argument mis apprehended the nature of epidemiologic data. Professor Fischer hypothesized seven known carcinogens, exposure to each of which posed a thirty percent risk of developing lung cancer. Fischer, supra note 359, at 1213. He argued that under proportional liability a plaintiff tortiously exposed to just one of them would probably receive a 90% recovery because there would be no other plausible cause, while if each of seven equally culpable defendants exposed the plaintiff to a different carcinogen, each defendant would be liable only for one-seventh of the damages. Id. The premise of a 90% recovery from the single tortfeasor, however, assumed that the seven known carcinogens are the only possible causes, so if the plaintiff was exposed to only one, that exposure must have been the cause—which is inconsistent with the hypothesized fact that *ceteris paribus*, exposure to that one carcinogen only explained 30% of the incidence of the disease. A court applying all-or-nothing but-for causation and the doubling+ rule would not let the hypothesized single-tortfeasor case be tried. Under probabilistic causal contribution, the single tortfeasor would appropriately be liable for 30% of the damages. The seven-defendant case would be more complicated, because the hypothetical treated the carcinogens independently but the hypothesized facts ruled out the possibility of independent additive risk contributions: the seven carcinogens could not together account for 210% of the risk of lung cancer. Based on equal relative risk values for each carcinogen, a fact-finder might attribute a one-seventh share to each, but evidence of the degree of confounding in the various studies, the similarity or differences of the carcinogens’ biological mechanisms, or studies of the interactions of multiple exposures might alter that result.

B. Corrective Justice Goals: Is Causation Different?

The Third Restatement of Torts acknowledged widespread adoption of comparative responsibility and apportionment principles. It embraced a theoretical distinction, however, between apportionment based on cause and apportionment based on fault or other norm-based conceptions of responsibility.\footnote{363} Apportionment based on cause, according to the Third Restatement, may be appropriate in limited circumstances when “legally culpable conduct . . . was a legal cause of less than the entire damages for which the plaintiff seeks recovery” and a fact-finder can determine “the amount of damages separately caused by that conduct.”\footnote{364} Other than straightforward enhanced-injury cases, however (as where a defendant negligently causes an accident that injures a plaintiff who is then further injured by medical malpractice), the Third Restatement did little to illustrate when such a situation might arise, noting that very little case law provides clear analysis of the problem.\footnote{365}

The Third Restatement thus took a clear position that but-for causation is a prerequisite for any fault-based apportionment. The basis for this choice was the principle that “[n]o party should be liable for harm it did not cause.”\footnote{366}

\footnote{363. Restatement (Third) of Torts: Apportionment of Liab. § 26 (2010).}
\footnote{364. Id. § 26(b); see also, e.g., Restatement (Third) of Torts: Liab. for Physical & Emotional Harm § 28 cmt. d (2010) (discussing circumstances under which harm is “causally divisible”); id. cmt. l (distinguishing circumstance when each one of multiple actors causes all of a plaintiff’s harm from circumstances in which each actor causes part of a plaintiff’s harm). Under the Third Restatement, liability for what the Second Restatement called distinct “harms,” Restatement (Second) of Torts § 433A (1)(a) (1965), would be assigned to each defendant whose conduct was a but-for cause of the separate harms, with any apportionment to be determined by the jurisdiction’s scope of liability rules. Restatement (Third) of Torts: Apportionment of Liab. § 26(b) (2010). The Third Restatement appeared to abandon the Second Restatement’s concept of single harms for which there is a reasonable basis for apportionment, or “divisible” harm. Compare Restatement (Second) of Torts § 433A (1)(b) & cmt. d (1965), with Restatement (Third) of Torts: Apportionment of Liab. § 26 cmt. a, c (2010).
}
\footnote{365. Restatement (Third) of Torts: Apportionment of Liab. § 26 cmt. a (2010) (“[F]ew if any cases provide much analysis.”).}
\footnote{366. Id.}
Andrew Klein relied heavily on the same principle in his argument for a strong requirement of *sine qua non* causation even in toxic torts. Professor Klein argued that requiring but-for causation is compelled by corrective justice principles. He also noted that to hold defendants liable for damages they did not “cause” undermines confidence in the legal system and bolsters calls for replacing tort law altogether. The powerful intuitive notion that cause is a yes-or-no phenomenon that cannot be divided up thus fits tightly with corrective justice theories of tort.

The principle that persons should be liable only for damages they cause, depends, however, on the meaning of “cause.” In the uncertain molecular world of toxic causation, if a substance is a risk factor for disease then that substance is causing injury to someone, but science does not let us identify that person in the routine, familiar way. Reliably attributing a particular plaintiff’s harm to a particular defendant’s conduct—were it possible—would be more satisfying than basing causation on contribution to risk, realized injury, and discounted recovery. But in the face of irreducible uncertainty, a discounted recovery provides at least some redress and is a better measure of corrective justice than no recovery at all. Probabilistic causal contribution thus can restore some degree of accountability that has been eroded by narrow judicial interpretations of the

---

367. See Klein, supra note 53, at 9–13 (discussing the value of *sine qua non* causation in terms of corrective justice principles).

368. Id. at 30.

369. See Sanders, supra note 1, at 26 (noting that mesothelioma is an indivisible injury not susceptible to apportionment, “[a]t least not apportionment based on causation”).


372. See John Makdisi, Proportional Liability: A Comprehensive Rule to Apportion Tort Damages Based on Probability, 64 N.C. L. Rev. 1063, 1073–74 (1989) (“While it is true that those who are actually injured by the tortious conduct do not receive their compensation in full and the others receive compensation to which they are not entitled under traditional tort theory, a fairer distribution of the damages is not possible given the uncertainty of causation.”).
admissibility and sufficiency of causation evidence.\textsuperscript{373} It is also better than allowing full recovery, which in the aggregate (assuming sufficient claiming) would force a defendant to pay for more harm than its actions created.

The theoretical purity of but-for causation has already been diluted to some extent by comparative responsibility regimes. Fact-finders must, and do, weigh the incommensurate all the time in comparative responsibility cases, which may entail a quantitative comparison of behavior found blameworthy (i.e., negligent) to behavior that is fault-free but nevertheless gives rise to liability.\textsuperscript{374} Who is more responsible for the truck that backed up into its driver: the strictly liable auto manufacturer that designed the defective transmission that mis-shifted, or the negligent driver who exited the car without disengaging the ignition?\textsuperscript{375} It is fanciful to think jurors view such questions through polarized lenses that transmit no light on relative causal contribution. Judgments of blameworthiness incorporate causal judgments too. When a drunk driver runs a stop sign and injures another driver who could not avoid the crash because a mechanic erred when repairing her brakes, the drunk driver gets a greater share of liability not just because the conduct is worse, but also because of the sense that the brakes would not have been needed absent the drunk driving, even though the bad brakes were a but-for cause of the accident.


\textsuperscript{374} New Jersey’s statute is exemplary: “In all negligence and strict liability actions[,]” the fact-finder must determine the “extent, in the form of a percentage, of each party’s negligence or fault.” N.J. STAT. ANN. § 2A:15-5.2.a (West 2002). The statute provides for only several liability for any party less than 60% “responsible” for the damages, subject to a partial exception for environmental tort actions. \textit{Id.} §§ 2A:15-5.3.b, 5.3.d.

\textsuperscript{375} See Gen. Motors Corp. \textit{v. Sanchez}, 997 S.W.2d 584, 598 (Tex. 1999) (holding that the plaintiff's actual recovery would be reduced by the “jury’s finding of fifty percent comparative responsibility”).
Current doctrine requires fact-finders to engage in *sub silentio* causal attribution in some toxic tort cases as well. As noted above, data show that exposure to both tobacco smoke and asbestos synergistically increases the risk of lung cancer above the sum of each product’s independent risk enhancement. Yet manufacturers of tobacco and asbestos products routinely name each other as the cause of lung cancer in plaintiffs exposed to both products. Even if fault dominates fact-finders’ thinking in such cases, it seems extraordinarily unlikely that they pay no attention to the risk data and other causal issues, such as the relative amounts of exposure, when they assign comparative responsibility.

A related view, prominently espoused by Richard Wright, relies on a sharp distinction between ex ante risk and ex post causation. The argument holds that statistical data derived from studies of populations provide insufficient information to satisfy corrective justice norms for a causal finding in an individual case; instead, particularistic evidence should be required to support a specific causal inference. If, however, particularistic evidence means evidence of some characteristic that distinguishes an exposure-caused case of disease from any other, this would imply that no plaintiff relying on epidemiologic or similarly population-based evidence could ever prevail, even in the post-genomic era. Professor Wright has acknowledged the normative appropriateness of forms of proportional liability to avoid this harsh result, at least in some circumstances in which it is impossible to prove “who actually caused the plaintiff’s injury.”

---

376. See *supra* Part III.A.2.


378. See, e.g., Wright, *supra* note 19, at 196, 206–08 (“Legal fact-finders are not told that they merely need to place a bet on the existence of some fact, but rather are instructed that they must determine whether the fact actually existed.”); Wright, *supra* note 24, at 1312–14; Wright, *supra* note 5, at 1049–50.

379. Wright, *supra* note 5, at 1072; see also Green, *supra* note 185, at 362–63 (discussing Professor Wright’s views on proportional liability); Wright, *supra* note 19, at 214 (approving of market-share liability in DES cases and the proportional liability imposed in *Barker*, though disagreeing with *Barker*’s risk contribution theory).
When the causation evidence inherently must be statistical, however, probabilistic causal contribution is not just a second-best approach; it is the approach that best fits knowable truth. The indeterminacy of toxic tort causation is different from the familiar example in which the knowledge that one company operates a large majority of buses on a route will not suffice to hold that company liable when an unidentified bus runs a car off the road. The intuitive discomfort with holding the bus company liable arises not because of the statistical nature of the evidence that it was the most probable owner of the errant bus. Rather, we do not really believe that the statistic accurately reflects the probability. To say that the bus company probably caused the accident is to say that one of its drivers was probably at fault and we are loath to do that without more information. Despite reductionist neurobiology, we don’t treat momentary lapses of attention as the result of random molecular movements in the driver’s brain. The statistical frequency of motor vehicle accidents per vehicle mile traveled is well known, yet most people think they can avoid accidents—indeed most people think they are above-average drivers. Instinctively we believe that road accidents, especially accidents caused by bad driving, are not

380. See Makdisi, supra note 372, at 1100–01 (when a causal link “is believed by a factfinder to be probable rather than actual, the willingness to consider this probability and allocate proportional relief reflects a more complete and accurate notion of cause”). But cf. Wright, supra note 19, at 214 (describing the market-share scheme as a “second best liability doctrine”).


382. See Makdisi, supra note 372, at 1079 (arguing that Smith v. Rapid Transit is partly explained because it is normatively “more offensive to attribute to a potentially innocent defendant a wrongdoing harm that he may not have caused”).

383. See generally, e.g., Leilani Greening & Carla C. Chandler, Why It Can’t Happen to Me: The Base Rate Matters, But Overestimating Skill Leads to Underestimating Risk, 27 J. APPLIED SOC. PSYCH. 760 (1997); Mark S. Horswill et al., Drivers’ Ratings of Different Components of Their Own Driving Skill: A Greater Illusion of Superiority for Skills That Relate to Accident Involvement, 34 J. APPLIED SOC. PSYCH. 177 (2004).
randomly distributed, and that bad drivers may not be randomly distributed across bus companies either.\footnote{384}

There is no morality to the molecule. In toxic torts, once a fact-finder concludes that general causation has been established by the available evidence,\footnote{385} no reason exists to suspect anything other than randomness for specific causation.\footnote{386} Or, at least, science usually won’t be able to distinguish the “actual” cause in a given case from the probability distribution of causes.\footnote{387} In such cases, insisting upon particularistic evidence to protect

\footnote{384. When causation turns on statistical data but issues of fault are resolved without statistical evidence, as in alternative and market-share liability cases, courts have been less likely to require particularistic evidence of causation. But cf. RESTATEMENT (THIRD) OF TORTS: LIAI. FOR PHYSICAL & EMOTIONAL HARM § 28 cmt. j (2010) (noting that courts are less willing to apply alternative causation when evidence connecting particular defendants to even the possibility of harm is absent, perhaps because exculpation is too difficult or plaintiffs are insufficiently diligent).

385. Professor Wright agrees that evidence of statistical associations may bear on general causation. Wright, supra note 19, at 206, 215.

386. See RUINSKY, supra note 178, at 153 (“[Q]uestions such as what is the cause for this mutation . . . do not have answers.”).

387. Michael Green pointed out that the distinction between particularistic ex post proof of causation and population-based ex ante proof of risk collapses if an epidemiologic study is the only evidence available. He wondered how Professor Wright would respond to a hypothetical in which a plaintiff proved exposure to an agent presenting a very high statistical risk of disease but a witness “claims to have observed the plaintiff contract the disease from an alien bite.” Green, supra note 185, at 362 n.38, 363. Professor Wright responded that the hypothetical was fallacious because it did not really involve competing causes but instead depended on an implicit assumption that the alien bite could not cause disease: “If . . . there is no proven or accepted alien-bite causal generalisation and no other possibly applicable causal generalisation, then” the toxic agent is “the only possibly applicable causal generalisation [the known statistical risk] with at least some particularistic instantiation [the known exposure] . . . which fact could support the formation of a belief that it was the causal process actually at work in the particular situation.” Wright, supra note 19, at 210 n.67. If proof of exposure could satisfy the demand for particularistic proof, that demand would present only a small obstacle to reliance on population-based data (even though proving exposure is sometimes difficult). But if exposure constitutes sufficient particularistic proof only when “no other possibly applicable causal generalisation” exists, then it will never be sufficient. No one can ever exclude every possible alternative cause, see Stubbs v. City of Rochester, 124 N.E. 137, 140 (N.Y. 1919), and frequently the competing cause is not a specific risk factor but rather “unknown causes.” See, e.g., Henricksen v. ConocoPhillips Co., 605 F. Supp. 2d 1142, 1149 (E.D. Wash. 2009).}
defendants from paying for harm they did not cause rings hollow; probabilistic causal contribution requires them to pay for the harm that they did cause.\footnote{See Makdisi, supra note 372, at 1073–74.}

Toxicogenomics and molecular epidemiology will continue to undermine the appropriateness of the deterministic model of but-for cause in toxic tort cases. Corrective justice would be better served by using a more appropriate causation model. With but-for and toxic torts, if it doesn’t fit, abandon it.

\textit{C. Truth-Seeking Goals: Wrong All the Time?}

In their helpful catalogue of “vices” of proportional recovery, Ariel Porat and Alex Stein include an epistemological problem: although all-or-nothing rules may produce erroneous results in some cases (either wrongly attributing causation where it did not actually exist or wrongly failing to attribute causation where it did actually exist), a rule that assigns a fractional recovery will produce a result at variance with historical truth in each individual case, because it will never be true that a tortious exposure caused \(x\)% of a plaintiff’s harm but not the remainder.\footnote{Ariel Porat & Alex Stein, \textit{Liability for Future Harm}, in \textit{Perspectives on Causation, supra} note 1, at 221, 227–28. Porat and Stein, who are generally sympathetic to probability-based compensation, argued that probabilistic recovery for the chance of future harm avoids this and some of the other “vices” of probability-discounted recovery for past harms. \textit{See generally id.} at 226–33 (describing how different characteristics of the causal indeterminacy problem for past and future harms affect the strengths and weaknesses of using proportional liability to address the indeterminacy).}

This is indisputable for a non-cumulative disease in a deterministic causal model when only population-based evidence is available.\footnote{To say that there is an \(x\)% probability that a randomly-selected exposed individual with a disease will be a “true” causation case, is not the same as saying that the individual’s case is \(x\)% probable of having “true” causation, even though many courts equate the two statements. \textit{See Gold, supra} note 277, at 390 n.72.} Any all-or-nothing rule, however, also accepts the existence of some erroneous adjudications, and further accepts that the aggregate result across a population of cases will be
wrong.\textsuperscript{391} It is not clear why having a lower count of errors is epistemologically superior to reducing the overall magnitude of the consequences of the errors.\textsuperscript{392} More fundamentally, if we adopt a probabilistic view of causation at the individual level—even if that model is only a concession to the limits of scientific understanding—then probabilistic causal contribution gets it roughly right in every case, and the epistemological problem that Porat and Stein identified disappears.

\textbf{D. Administration of Justice Concerns: Too Much Information?}

In a thoughtful essay, Joseph Sanders emphasized a number of “practical, administration-of-justice” problems that he concluded outweighed the “persuasive substantive arguments” in favor of the “risk rule” of \textit{Rutherford} and \textit{Barker}.\textsuperscript{393} Professor Sanders expressed concern about the level of complexity that risk-based causal attribution would bring to toxic tort litigation,\textsuperscript{394} although he acknowledged that toxicogenomics could actually alleviate some of that concern by providing a more finely-grained understanding of causal mechanisms.\textsuperscript{395}

\textsuperscript{391} Not everybody who has experienced both illness and exposure will sue, which adds a further element of unpredictability to the fact-finding accuracy of all-or-nothing rules. If the probability of claiming were strongly positively correlated with the probability of “true” causation (e.g., because only people with obvious exposures or known genetic susceptibility are likely to sue), but the applicable all-or-nothing rule bars recovery, the frequency of false negative errors among cases actually brought could approach 100%.

\textsuperscript{392} See \textit{Makdisi}, supra note 382, at 1065, 1074 (arguing that proportional liability is more accurate and fair).

\textsuperscript{393} Sanders, \textit{supra} note 1, at 34, 36.


\textsuperscript{395} Sanders, \textit{supra} note 1, at 36.
The concern is well-founded. Probabilistic causal contribution does not have the virtue of simplicity. Courts will be challenged to administer cases in reasonable time, at reasonable cost, and with appropriate levels of scrutiny of expert testimony, which will be offered by all parties. But even if courts adhere to a deterministic “more-likely-than-not but-for” causation model, they will still be confronted with toxicogenomic and molecular epidemiologic data and all of the complexity such evidence will entail. As I’ve said elsewhere, that train is coming, and courts cannot get off the tracks.\footnote{Gold, supra note 3, at 397.}

The complexity courts will face will reflect the complex world that science reveals. If toxicogenomics, molecular epidemiology, and related sciences fulfill their potential, they will show us an elaborate landscape of toxic risks and variable genetic susceptibility to those risks, many of them relatively small but still significant. If, in the name of simplicity and reduced administrative expense, courts impose on that landscape the same all-or-nothing rules now in effect, the law will willfully blind itself to the truth, will too frequently fail to redress harm truly caused, and will impose too much liability for the harm it does redress. Such outcomes would be neither efficient nor just.

VI. Conclusion

Paradoxically, the emergence of toxicogenomics and molecular epidemiology promises to increase our understanding of the effects of toxic exposures while simultaneously highlighting the fundamental uncertainty of any individual claim. These scientific advances strengthen, rather than weaken, the case for reform of causation doctrine.

The deterministic model of but-for causation does not fit toxic tort cases that depend on population-based data—even if the data come from studies at a molecular scale. Worse, it produces results that are inefficient and unjust. Commentators have recognized this for a quarter-century, but with few exceptions the courts have not responded. In the post-genomic era, it is time for the
courts to adapt,\textsuperscript{397} and to adopt a probabilistic causal contribution model in toxic tort cases.

\textsuperscript{397} See Donovan v. Philip Morris USA, Inc., 914 N.E.2d 891, 901 (Mass. 2009) (recognizing a medical monitoring claim for increased risk resulting from toxic exposure).
Information Privacy and Data Control in Cloud Computing: Consumers, Privacy Preferences, and Market Efficiency

Jay P. Kesan*
Carol M. Hayes**
Masooda N. Bashir***

Abstract

So many of our daily activities now take place “in the cloud,” where we use our devices to tap into massive networks that span the globe. Virtually every time that we plug into a new service, the service requires us to click the seemingly ubiquitous box indicating that we have read and agreed to the provider’s terms of service (TOS) and privacy policy. If a user does not click on this box, he is denied access to the service, but agreeing to these terms without reading them can negatively impact the user’s legal rights.

As part of this work, we analyzed and categorized the terms of TOS agreements and privacy policies of several major cloud services to aid in our assessment of the state of user privacy in the cloud. Our empirical analysis showed that providers take similar approaches to user privacy and were consistently more detailed when describing the user’s obligations to the provider than when describing the provider’s obligations to the user. This asymmetry, combined with these terms’ nonnegotiable nature, led us to

* Professor and H. Ross & Helen Workman Research Scholar, University of Illinois College of Law.
** Research Associate, University of Illinois College of Law; Fall 2010 Fellow in the Christine Mirzayan Science and Technology Policy Graduate Fellowship program at the National Academy of Sciences.
*** Assistant Director for Social Trust Initiatives, Information Trust Institute, University of Illinois. The authors also wish to acknowledge the excellent research assistance of Robert Zielinski in preparing this work.
conclude that the current approach to user privacy in the cloud is in need of serious revision.

In this Article, we suggest adopting a legal regime that requires companies to provide baseline protections for personal information and also to take steps to enhance the parties’ control over their own data. We emphasize the need for a regime that allows for “data control” in the cloud, which we define as consisting of two parts: (1) the ability to withdraw data and require a service provider to stop using or storing the user’s information (data withdrawal); and (2) the ability to move data to a new location without being locked into a particular provider (data mobility). Ultimately, our goal with this piece is to apply established law and privacy theories to services in the cloud and set forth a model for the protection of information privacy that recognizes the importance of informed and empowered users.

Table of Contents

I. Introduction ..................................................................... 344

II. Cloud Computing Fundamentals .................................... 347
   A. Background Technology ............................................ 347
      1. The Internet ......................................................... 349
      2. Mobile Computing ............................................... 351
      3. Security ............................................................... 352
      4. Related Regulations............................................... 353
      1. Defining Cloud Computing .................................. 355
      2. Growth of Cloud Computing ................................ 356
      3. Uses of Cloud Computing .................................... 358
      4. Types of Cloud Computing Services .................... 360
   C. Advantages and Disadvantages of Cloud Computing ......................................................... 362
   D. Cloud Computing Legal Issues ................................ 365
      1. Privacy ................................................................. 366
      2. Jurisdiction .......................................................... 368
   E. Calls for Action in the Cloud .................................... 371
      1. Transparency and Control .................................... 372
   F. Cloud Services in Different Industries ..................... 373
II. Privacy Fundamentals .................................................... 375
   A. Privacy Theories ........................................................ 375
      1. Warren and Brandeis ........................................ 380
      2. Prosser ............................................................. 381
         a. Prosser’s Privacy Torts and Information
            Privacy .............................................................. 383
      3. Modern Informational Privacy Theory ................ 384
         a. Concepts of Privacy ........................................ 386
         b. The First Amendment Critique ..................... 388
         c. Privacy as a Commodity .............................. 390
   B. Privacy Law ............................................................... 392
      1. Steps Toward Regulation of Privacy ................ 393
      2. Federal Privacy Statutes and State Laws .......... 395
         a. Electronic Communications Privacy Act .......... 399
            (1) Stored Communications Act ..................... 401
            (2) Applying the SCA to the Cloud ............... 405
      3. Case Law ............................................................ 407
         a. Fourth Amendment ..................................... 408
         b. Stored Communications Act ......................... 414
         c. Contracts and Privacy ................................ 416
      4. European Privacy Law ........................................ 418
         a. The Safe Harbor Framework ....................... 419

IV. Companies, Customer Data, and Customer-Company Interactions ........................................ 421
   A. Companies and Customer Data ............................ 421
      1. Terms of Service Agreements .......................... 421
         a. TOS Agreements as Contracts of
            Adhesion ...................................................... 424
      2. Privacy Policies .............................................. 425
         a. Sharing Information with the Government .... 427
      3. Effects of Security Breaches .............................. 430
      4. Protecting Consumer Data—Who Watches the Watchers?.............................. 432
      5. Tracking Technologies and Behavioral
         Marketing .......................................................... 436
      6. Personally Identifiable Information and
         “Anonymous” Information .................................. 440

V. Empirical Analysis of Agreements and Policies in the Cloud ............................................. 443
I. Introduction

What price for your privacy? As social interactions and business activities have shifted online, or into “the cloud,” personal information has become a currency with an undervalued exchange rate. What data are consumers willing to trade in exchange for convenience and services online? Would they be as willing to engage in this trade if their privacy rights were more protected and if they had the ability to exercise meaningful control over their data?

Technological and social changes have stimulated many developments over the last decade as the Internet became ingrained in society and social interactions. Substantial technological changes require the law to adapt. When our perceptions change, policymakers amend the law accordingly to address evolved expectations. In this Article, the perceptions and law that we are concerned about are those associated with privacy, especially privacy in the context of services provided over the cloud.

“You have zero privacy anyway. Get over it.”
—Scott McNealy, Chairman and former CEO of Sun Microsystems, 1999
This is not the first time that conceptions of privacy have been shaped by technology. *The Right to Privacy*, published in 1890, was the seminal work of Samuel Warren and Louis Brandeis that substantially influenced privacy law in the United States in the twentieth century.¹ The publication of this piece was spurred by the authors’ concerns about intrusions into personal privacy by the press, especially considering the technological improvements that had enabled the production of small, affordable cameras.²

Portable cameras were just the beginning of technology that prompted major changes in privacy law and theory. Around the middle of the twentieth century, computers were becoming more pervasive and powerful, enabling the creation of databases that could hold and process huge amounts of information. The idea of informational privacy developed in greater detail around this time, as people realized that personal privacy could be threatened not just by appropriation of one’s name and likeness, but by access to and use of other information about a person.³

While these informational privacy concerns were becoming more visible, the future of connecting computers in a global telecommunications network was just a glimmer in the eyes of some of the more innovative researchers. Today, in exchange for our personal information, we have access to free e-mail and free data storage, and we can use free services to keep in touch with former classmates and colleagues around the country and around the world. Thanks to Facebook, attendees of modern high school


reunions who live on opposite sides of the country can focus on catching up on events of the last week instead of the last ten years.

As with any improvement in technology, however, there are also tradeoffs. Free services online are often funded by advertising revenue, and these ads are made more effective by utilizing the user’s personal information to target ads to their interests. To set up accounts for services online, consumers must typically click the ubiquitous box indicating that they have read and agreed to the website’s terms of service (TOS) and privacy policy. These agreements often contain broad provisions for what the provider is permitted to do with the consumer’s information, while giving the consumer few, if any, options for redress. In the majority of cases in which services are marketed to individual users, there is zero negotiability in these terms, and almost no one reads these terms anyway.

In this Article, we urge the creation of baseline regulations that would guarantee a minimum level of protection of consumer privacy while preserving market vitality. One of the essential elements for this baseline regime would be the protection of the consumer’s right to control his data. People are often denied meaningful control over their personal information and the other information that they store with these services. Companies often do not address beforehand how a consumer can exercise control over their information in the event that the service is terminated, and many companies reserve a nonrevocable license to use the consumer’s intellectual property that is stored with its service. We view this right of data control as consisting of two parts: (1) data mobility, which we summarize as a right to move one’s data and terminate a relationship with a particular service provider, under which providers would be required to provide data to departing customers in a generally accepted file format such that customers do not become “locked in”; and (2) a broader right of data withdrawal that would permit a consumer to withdraw his information from the records of any entity, including a third party, through a notice-and-takedown process.

In Part II, we explain the idea of cloud computing and introduce a number of issues related to it. In Part III, we turn to an examination of privacy fundamentals, first examining different theoretical approaches to privacy before turning to a
discussion of privacy law in the United States and an examination of statutes and case law. In Part IV, we describe issues relating to companies and customer data, including concerns about TOS agreements, privacy policies, and data security. In Part V, we turn to the results of our empirical analysis of the TOS agreements and privacy policies of a sample of cloud service industry leaders. Finally, in Part VI, we offer our recommendations based on our empirical work, as well as our research into privacy issues and the cloud.

II. Cloud Computing Fundamentals

In examining the legal implications of privacy and cloud computing, it is important to understand some of the background. In this Part, we will examine some of the technical background of the current technologies before discussing cloud computing and its advantages and disadvantages in more detail. We will also introduce some legal issues that arise in the cloud context and briefly review various calls for action that have sounded with respect to the cloud, such as calls for amending legislation, proposing legislation, or calling for standards or increased transparency.

A. Background Technology

Before the World Wide Web (Web) became so prevalent, there were two paradigms of computer use. The first was mainframe computing, and under this paradigm, users worked at “dumb terminals” that were connected to a large mainframe system, which in turn processed the users’ requests. As microprocessors became available, the personal computing paradigm took over, and the files and data were under the users’ physical control.

The personal computing paradigm has weaknesses, however, including the low degree of scalability of individual systems, the need for technological expertise to assemble and maintain computer systems, and a low level of redundancy such that data loss through equipment failure is a significant danger. Under the traditional model of information technology (IT) management, based on this paradigm, a lot of space and human capital is required to maintain and secure the systems of a large enterprise. Moving our technological worlds to the cloud is another paradigm shift that some view as the future of computing. Mark Weiser predicted in 1991 that the third wave of computing, after mainframe computing and personal computing, would be ubiquitous computing, where computers become so small, inexpensive, and ubiquitous that they virtually disappear. Today, technologies continue to improve, but the truly ubiquitous nature of modern computing is not because of the computer’s size

6. See Robison, supra note 4, at 1200–01 (explaining the inefficiencies that occur when everyone has his or her own computer).

7. Mark H. Wittow & Daniel J. Buller, Cloud Computing: Emerging Legal Issues for Access to Data, Anywhere, Anytime, 14 NO. 1 J. INTERNET L. 1, 5 (2010). Wittow and Buller note that these limitations are mitigated by the use of things like centralized disk storage, the use of more advanced servers with smaller hardware footprints, and system virtualization. Id. Virtualization is one of the major technologies behind some applications of “cloud computing,” where spaces on hard drives are turned into “virtual machines” that segment the processing of different requests. VMWare, Virtualization Basics, http://www.vmware.com/virtualization/virtualization-basics/how-virtualization-works.html (last visited Feb. 3, 2013) (on file with the Washington and Lee Law Review).

8. See Ilana R. Kattan, Cloudy Privacy Protections: Why the Stored Communications Act Fails to Protect the Privacy of Communications Stored in the Cloud, 13 VAND. J. ENT. & TECH. L. 617, 621 (2011) (noting the transitions between paradigms); Soghoian, supra note 5, at 364 (noting that cloud computing has been deemed by many commentators to be the future of computing). Some suggest that the decentralized cloud computing model has the potential to make such services comparable to utilities, with data centers being the equivalent of power plants in the electrical utility context. Kevin Werbach, The Network Utility, 60 DUKE L.J. 1761, 1817 (2011). If cloud providers are utilities, the argument for regulation of services on the cloud becomes stronger. Id. at 1818.

or power. The Internet and high-speed connections allow people to be in touch not only with each other, but with service providers that can essentially rent out processing power and storage space over the Internet. Moving some functions to the cloud can allow users access to high-end services and technology without having to trade quality for mobility. This future of computing, however, may challenge the default assumption that a user will be able to control her own data.

1. The Internet

The history of the Internet is often traced back to the late '60s and ARPANET. Even before ARPANET, however, some recognized the possible future value of computers being connected using communication lines. Regardless of how the ideas emerged, there is no doubt that the Internet is a pervasive element of today's society.

To say that the Internet has become a staple of modern life is an understatement. The Internet has had a substantial effect on the world and how people interact. Cyberspace is a major social...
outlet that is often intertwined with the physical realm.\textsuperscript{15} People keep in touch through a variety of electronic messaging technologies, including e-mail, text messaging, other instant messaging over the Internet, and social networking websites.\textsuperscript{16} Research by the Kaiser Family Foundation suggests that the average youth between the ages of eight and eighteen spends every permissible waking moment using electronic devices, many of which are connected to the Internet, like smart phones and computers.\textsuperscript{17} A study by the Nielsen Company found that across all ages, the average American Internet user is online over fifty-five hours per month.\textsuperscript{18}

The Internet works because computers on the network use identical protocols that enable interconnection so that data can be delivered across the network.\textsuperscript{19} One of the well-known protocols is the Simple Mail Transfer Protocol (SMTP), which enabled e-mail exchanges in the 1980s in the days before the World Wide Web.\textsuperscript{20} In the mid-1980s, the transfer of e-mail was fairly fragmented, with communications being transmitted from server to server, stored at various locations temporarily during the trip before being downloaded by the recipient.\textsuperscript{21} Today, webmail still uses the SMTP protocol, as well as the Internet Message Access everything different.").

\begin{itemize}
\item \textsuperscript{15} Id. at 639.
\item \textsuperscript{16} See John Soma, Melodi Mosley Gates & Michael Smith, \textit{Bit-Wise but Privacy Foolish: Smarter E-Messaging Technologies Call for a Return to Core Privacy Principles}, 20 ALB. L.J. SCI. & TECH. 487, 497–502 (2010) (explaining the five technologies, including telephone systems, e-mail, text messaging, instant messaging, and social networking); Strandburg, supra note 14, at 655–56 (explaining how social media promise to change social interactions by supplementing physical interaction or replacing it).
\item \textsuperscript{17} Andrea Cascia, \textit{Don’t Lose Your Head in the Cloud: Cloud Computing and Directed Marketing Raise Student Privacy Issues in K–12 Schools}, 261 WEST’S EDUC. L. REP. 883, 894 (2011).
\item \textsuperscript{18} Paul Lanois, \textit{Caught in the Clouds: The Web 2.0, Cloud Computing, and Privacy?}, 9 NW. J. TECH. & INTELL. PROP. 29, 29 (2010). About half of that time is spent on social networking, e-mails, games, and instant messaging. Id.
\item \textsuperscript{19} See Werbach, supra note 8, at 1769 (explaining the functionality and concept of the Internet).
\item \textsuperscript{20} OKIN, supra note 12, at 212.
\item \textsuperscript{21} See Robison, supra note 4, at 1205–06 (explaining the functionality of electronic communication services).
\end{itemize}
Protocol (IMAP). IMAP allows e-mails to be accessed from anywhere with an Internet connection, with e-mails being perpetually stored on the provider’s servers. The ability to access information from anywhere is important for mobility and mobile computing.

2. Mobile Computing

Computers have shrunk in size over the last fifty years, from room-size computers to thirty-pound desktops to five-pound laptops to smart phones weighing just a few ounces. The early 1980s saw the invention of the first laptop and the first cellular phone, and the first personal digital assistant (PDA) was released in 1993. This increase in mobility has been helpful for both personal and professional tasks. The Blackberry became a popular office tool after its release in 1999, functioning as both a cell phone and a PDA that permitted remote access to office e-mail. Today’s smart phones go beyond the original Blackberry, giving users access to e-mail, the Web, appointment calendars, and even software that allows the users to review word processing files and full color PDFs in the palms of their hands. It is estimated that by 2013, about half of the mobile phone market will be smart phones, and many if not all of these are likely to have access to 3G or 4G data networks that do not require a separate wireless connection.


24. Id. at 28.


The desire for technologies that can go anywhere makes cloud computing more appealing. Even with improvements in personal computing technology, increased mobility generally requires a tradeoff with the hardware abilities of the device. This is where the value of the cloud becomes clearer: there are fewer tradeoffs from having smaller and cheaper end-user devices because these devices can tap into the power of network-based services.

However, mobile devices are vulnerable to the same sorts of security threats as full-sized computers, including spyware and viruses, and data transmitted using these devices may not be secure. For this reason, and because of the significant security concerns that arise in the cloud computing context, we turn to this topic next.

3. Security

The security of any information in the cloud is often unclear. There was an uproar when Scott McNealy of Sun Microsystems dismissed online privacy concerns by proclaiming in 1999, “You have zero privacy anyway. Get over it.” But regardless of whether consumers will be persuaded by assertions about the nature or extent of privacy, active efforts by third parties to infringe on privacy should properly raise red flags.

One threat to devices accessing the cloud is spyware. One can define spyware as software that installs itself, runs, and uses its host computer, all without the owner’s permission. Similar software has been called “adware,” an example of which was the software produced by Gator, which was ad-supported and sent...


27. See Werbach, supra note 8, at 1816 (explaining how cloud computing is changing the way people think about computers and computer networks).

28. See Rhodes & Kunis, supra note 23, at 32–33 (noting the existence of malware that targets mobile devices, worms with the ability to monitor and record cell phone conversations, and the exploitation by hackers of information transmitted using wi-fi hotspots).


30. Schwartz, Property, supra note 10, at 2065.
information about the user and his computer back to the company.31 In 2004, sources indicated that Gator software was installed on about thirty-five million computers located in the United States.32

4. Related Regulations

The degree to which the Internet is or should be regulated is the subject of much debate.33 Werbach traces the origins of the broadband regulation debate back to the 1960s, when the FCC launched the Computer Inquiries to determine when and how data processing services would become sufficiently intertwined with communications that they would be covered by the Communications Act.34 In the first of the Computer Inquiries, Computer I, the FCC concluded that there was "no public interest requirement for regulation by government of such activities" because of the competitive nature of the market for data processing services.35 However, the FCC did recognize that the communications circuits that carried these services might need to be regulated.36

31. Id. at 2066.

32. Id. at 2065.


36. Computer I Final Decision, 28 F.C.C.2d at 269 ("[W]ithout appropriate regulatory safeguards, the provision of data processing services by common carriers could adversely affect the statutory obligation of such carriers to provide adequate communication services under reasonable terms and conditions and impair effective competition in the sale of data processing..."
At the turn of the century, questions about regulating these communications circuits came to the fore. The Telecommunications Act of 1996[^37] established a category of services called “information services,” which the Act defines as “the offering of a capability for generating, acquiring, storing, transforming, processing, retrieving, utilizing, or making available information via telecommunications.”[^38] Information services are not regulated as a common carrier under Title II of the Telecommunications Act. In 2002, the FCC designated cable Internet as an “information service” instead of a “telecommunications service,” a designation that was upheld by the Supreme Court[^39] and later expanded to include DSL service.[^40] The *National Cable & Telecommunications Ass’n v. Brand X* case was regarded by some as marking a decision to not regulate the Internet, given the lesser degree to which information services were regulated compared to telecommunications services.[^41]

### B. What Is Cloud Computing?

Up to this point, we have referenced “the cloud” in the context of cloud computing as a new computing paradigm. In this subpart, we will go into more detail about cloud computing and what it is.

1. Defining Cloud Computing

The term “cloud computing” has become popular and trendy, but there are many concepts behind this idea. On a general level, “cloud” is used as a metaphor for the “ethereal Internet” and the virtual platform that it provides. Some view cloud computing abstractly as the result of the convergence of computing and communications, or more practically as a “scalable network of servers,” as “IT as a service,” or as the convenience of being able to access a shared pool of computing resources over a network like the World Wide Web.
Denny maintains that there is not a uniform definition of cloud computing. On the other hand, many commentators also authoritatively cite the definition of cloud computing put forth by the National Institute for Standards and Technology (NIST), which currently defines it as “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.” For our purposes, we accept the NIST’s definition because it is broad enough to encompass the variety of uses for cloud computing.

2. Growth of Cloud Computing

Cloud computing is a growing segment of technology services, thanks in part to the availability of high speed Internet service. A study by the Pew Internet and American Life Project concluded that about 69% of Internet users in the United States already use webmail, other software programs located solely

---

47. Denny, supra note 46, at 237.


49. Robison, supra note 4, at 1201.
online, or online data storage.\textsuperscript{50} A survey of technology insiders and critics in 2010 reflected a view by the majority that cloud computing technologies will be heavily used in work environments by 2020, with most expecting the PC model to decrease in importance.\textsuperscript{51} Some suggest that as cloud computing grows and more activities transition onto the Internet, there will be a greater focus on interoperability between cloud platforms and applications.\textsuperscript{52}

As a result of more people using cloud services, the revenue in this industry is expected to grow substantially. The cloud services industry saw revenue of $58.6 billion in 2009, and some analysts are anticipating that the industry’s revenue will increase between $40 billion and $160 billion over the next few years.\textsuperscript{53} Because of these large growth forecasts, many companies are pushing to be at the forefront of this movement.\textsuperscript{54}

\begin{footnotesize}
\begin{itemize}
  \item 50. See John B. Horrigan, \textit{Cloud Computing Gains in Currency}, PEW RES. CTR. (Sept. 12, 2008), http://pewresearch.org/pubs/948/cloud-computing-gains-in-currency (last visited Feb. 3, 2013) (on file with the Washington and Lee Law Review). A majority of those responding in the Pew study also indicated that they were very concerned about the use of their personal data by cloud providers. See \textit{id.}; see also Martin, supra note 44, at 298 (discussing the Pew Research Center study); Wittow & Buller, supra note 7, at 5 (same).
  \item 51. Kattan, supra note 8, at 620. Some have noted that cloud computing has the potential to partially replace the desktop computer. See Stylianou, supra note 44, at 604; see also Werbach, supra note 8, at 1813–14 (discussing how the rise of smart, connected mobile devices will increase incorporation of cloud computing).
  \item 52. See Stylianou, supra note 44, at 597 (stating that some platforms and applications will allow interoperability, which will allow users to transfer content easily).
  \item 53. See Lanois, supra note 18, at 30 (citing a study anticipating growth to $148.8 billion in revenue by 2014, and a study anticipating over a 20% increase in spending on cloud services by organizational customers); Soghoian, supra note 5, at 361 (citing analyst expectations of industry revenue growth between $40 billion and $160 billion).
  \item 54. See Lanois, supra note 18, at 30 (referring to a “recent bidding war between Hewlett-Packard and Dell to acquire cloud storage firm 3PAR”).
\end{itemize}
\end{footnotesize}
3. Uses of Cloud Computing

There are a lot of uses of cloud computing and a lot of aspects to those uses. One of the earliest forms of cloud computing was server-side e-mail storage. There are many companies offering cloud services. Webmail in particular is very popular, and sometimes an organization may contract with cloud providers for e-mail in order to save money over running its e-mail system in-house. Google provides such services to organizations through its Google Apps service, as well as free services to individuals over the Web. Google’s services to the public include webmail through Gmail and Web-based productivity software through Google Docs.

There are also a number of other uses that are not as immediately visible. Users can take advantage of the cloud to improve the functionality of locally run software, like the Weave add-on for the Firefox Web browser, which allows users to synchronize bookmarks, saved passwords, and cookies across multiple computers by storing this information on Mozilla’s servers. Additionally, Ford is working on a system that would bring features of cloud computing and social networking to new cars, perhaps including things like traffic alerts and real-time fuel consumption monitoring. Cloud computing could also be useful in education to increase student engagement and provide

---

55. See Couillard, supra note 42, at 2218 (explaining that server-side e-mail was one of the first iterations of cloud computing); Robison, supra note 4, at 1203 (referring to server-side e-mail storage as one of the first cloud computing services available to the public).

56. See Lanois, supra note 18, at 30 (listing offerings of companies, including Amazon, Microsoft, IBM, and VMWare).


58. See Soghoian, supra note 5, at 367–68 (describing services offered by Google Apps).


60. See Soghoian, supra note 5, at 397 (explaining the characteristics of Firefox, Mozilla’s browser).

61. Lanois, supra note 18, at 32.
students with additional tools like online forums and storage space in the cloud.  

The cloud is also leading to many innovations in entertainment. Some gaming services are appearing in the cloud, like OnLive and Gaikai, and some posit that the cloud has the potential to let gamers play games with high-end graphics without having high-end computers. Other entertainment uses of the cloud include subscription or ad-supported video streaming services like Netflix and Hulu. There are also social networking websites, like Facebook, that behave in ways consistent with the NIST’s definition of cloud computing.

The providers of cloud services may take a variety of approaches to service provision, differing in areas like cost models, user interfaces, and treatment of user data. Because cloud services are still fairly new, some companies may also seek to ease the transition to the cloud by making their services resemble software that is run locally on a computer. In addition to easing the transition by focusing on the user experience, cloud service providers also may make their services more appealing by offering them for free. There are many cloud services that are already provided for free, and these services can remain profitable by relying on ad support. Companies that do so often

62. See Cascia, supra note 17, at 884 (discussing the benefits of integrating cloud computing in schools). The Department of Education takes the position that cloud computing, data mining, and data aggregation could play valuable roles in increasing student performance and keeping school districts accountable. Id. at 887.

63. Lanois, supra note 18, at 31. OnLive launched in June 2010, but is said to already be worth $1.1 billion. Id.


65. See supra note 48 and accompanying text (discussing the NIST definition of cloud computing).

66. See Soghoian, supra note 5, at 369–70 (explaining single-site browser technology). A cloud service provider, looking to ease the transition between local computing and cloud computing, might also choose to provide support for offline access, such as Google’s Gears browser add-on that allows limited access to Gmail. Id. at 370–71.

67. See Jiang, supra note 48, at 415 (explaining different business models for cloud computing).
use customer information to generate targeted advertisements, which some criticize as effectively monetizing users’ private data.\textsuperscript{68}

Providers may also take very different approaches to data protection and encryption depending on the service, and we argue that the public should be made aware of data protection issues. Remotely stored data that is not intended for public access is likely to be encrypted, password protected, or have unlisted links.\textsuperscript{69} Other data, especially data that is not considered “sensitive,” are typically stored in an unencrypted format.\textsuperscript{70} Because cloud computing technology is still emerging, added features like increased security would cost more for early adopters, and this cost plus the current lack of market demand means that cloud service providers currently do not have much incentive to invest in enhancing security for a lot of the data involved.\textsuperscript{71} One of the things that current customers demand, however, is reliability, so cloud service providers often go to great lengths to have their services available at least 99.9\% of the time.\textsuperscript{72}

4. Types of Cloud Computing Services

Cloud services may be private, public, or some hybrid of the two.\textsuperscript{73} Private clouds may also be referred to as “internal” clouds, and are located solely within that organization and use only that

\textsuperscript{68} Soghoian, supra note 5, at 396.

\textsuperscript{69} Couillard, supra note 42, at 2217. Mozy asserts that it uses encryption technologies when user data is transmitted and stored, which is different from most other companies that say that they use SSL encryption for the exchange of data but do not specify whether data in storage is encrypted. Stylianou, supra note 44, at 603.

\textsuperscript{70} Stylianou, supra note 44, at 605. Because Google does not encrypt stored e-mails, for example, Google’s software can scan e-mail content for key words for the purpose of targeted advertising. Id. at 606.

\textsuperscript{71} Id. at 607.

\textsuperscript{72} Id. at 607.

\textsuperscript{73} Barnhill, supra note 48, at 640.
organization’s infrastructure.74 Public clouds are offered over the Internet and are supported by ads or fees.75

There are three primary models for public cloud services: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).76 Companies that provide servers and storage for remote use are providing IaaS, while companies that provide platforms on remote servers to run applications are providing PaaS.77 A company that makes software applications available over the Internet, including webmail, is providing SaaS.78 Gmail and Facebook are examples of SaaS cloud services.79 SaaS goes much further, however, and includes services like online gaming and online legal research.80

SaaS is arguably the level that consumers are most familiar with. The other types of cloud computing services may be more appealing to developers and computing professionals. PaaS, for example, gives customers (often software developers) the ability to deliver their own software applications over the Web to end users at a lower cost to the developer since they are using someone else’s servers to do so.81 IaaS, on the other hand, involves cloud providers giving customers access to raw computing resources in a manner similar to a utility service.82 Because this Article focuses on individual consumers, the most relevant category of cloud service for our purposes is SaaS.

74. Couillard, supra note 42, at 2216; Martin, supra note 44, at 287.
75. See Martin, supra note 44, at 287 (explaining how cloud computing works).
77. Id.
78. Id. at 639; Denny, supra note 46, at 237.
80. Martin, supra note 44, at 287–88 (“Under the SaaS model, a user interacts with an online service through the Internet, and the online service’s vendor provides the necessary software applications and remote data storage.”).
81. See id. at 289 (explaining the lower costs of PaaS compared to SaaS); Robison, supra note 4, at 1203 (noting the use of PaaS by third-party developers).
82. Robison, supra note 4, at 1204.
C. Advantages and Disadvantages of Cloud Computing

Moving more services onto the cloud has many promises and pitfalls. It is possible that the future success of cloud services will depend on how these advantages and disadvantages balance with each other and, more importantly, with the public’s expectations.83

Advantages of the cloud paradigm include data preservation,84 high levels of expertise on the part of cloud service providers,85 scalability,86 affordability,87 and availability.88 Additionally, some studies have shown that businesses that adopt SaaS enjoy a return-on-investment of almost 600%.89 Cloud providers are benefited because they have control over content, can set access terms, and can also monitor usage statistics.90

83. Stylianou, supra note 44, at 606 (“In effect, the combination of the sensitive nature of information that cloud services usually attract, the lack of adequate security from cloud services, and the intensification of governmental intrusiveness, stands as an impediment to the spread of cloud services.”). Stylianou also suggests that if cloud services implemented stronger security measures, like encrypting stored data, such changes could make cloud services more attractive to business customers. Id. at 609.

84. See Martin, supra note 44, at 294 (describing the benefit of being able to access applications and data from anywhere at any time).

85. Id.; Stylianou, supra note 44, at 603.

86. See Cascia, supra note 17, at 888 (citing the Department of Education’s position that the scalability of cloud-based IT services would help schools cut costs); Jiang, supra note 48, at 413; Martin, supra note 44, at 294 (stating that cloud computing offers rapid and intelligent resource adjustment as well as economies of scale); Wittow & Buller, supra note 7, at 5 (noting that cloud computing allows a system’s capacity and capability to be increased without additional infrastructure or personnel investments). Wittow and Buller cite the example of Animoto, which went from 25,000 users to 250,000 users over the course of just three days and was able to keep pace with this very high rate of growth by acquiring more virtual servers. Id. at 5–6. The scalability advantage works both ways, allowing small companies to easily expand their technological resources, and allowing downsizing companies to easily cut unnecessary IT costs. Rhodes & Kunis, supra note 23, at 31.

87. Barnhill, supra note 48, at 640–41; Martin, supra note 44, at 289; Soghoian, supra note 5, at 366.

88. Jiang, supra note 48, at 413; Soghoian, supra note 5, at 366.

89. Martin, supra note 44, at 289

90. See Jiang, supra note 48, at 413; see also Soghoian, supra note 5, at 364–65 (listing the ability to terminate user access and make sure that users are always running the current software version as two advantages of the cloud
These additional advantages for cloud providers also make cloud services attractive to copyright holders because the control exercised by the cloud provider can provide additional security and protect the copyright holder from infringement.91

There are also many disadvantages to the cloud paradigm, and many of these disadvantages arise in part because of consumers’ loss of control over data. Because consumers are entrusting their data to a third party, they are relying on that third party to adequately secure the information,92 have the services and data available at all times,93 and allow the consumer to move their information between providers freely,94 all in a context in which it is unclear how modern privacy law (including the Fourth Amendment and laws related to confidentiality) may

to the service provider).

91. See Jiang, supra note 48, at 422; Soghoian, supra note 5, at 364–65 (noting the value of the cloud for helping content owners better protect copyrights and trade secrets).

92. See Soghoian, supra note 5, at 374 (“[N]early all [] leading cloud providers offer products that are by default vulnerable to snooping, account hijacking, and data theft by third parties.”). Soghoian suggests that the reason hackers are a threat to users of cloud services is because cloud providers have not yet adopted strong encryption technologies. Id. at 361. Businesses are likely to be very concerned about the potential security issues of the cloud, so they will have to balance the financial benefits of moving to the cloud against the costs of data security like encryption and key management. Couillard, supra note 42, at 2217.

93. See Kattan, supra note 8, at 623 (explaining how cloud computing creates dependency); Martin, supra note 44, at 294 (describing the benefit of being able to access applications and data from anywhere at any time). While cloud services strive for reliability, the technology is still developing and thus is still very susceptible to human error and programming bugs, like the leap day bug that caused Microsoft’s Azure service to be unavailable all day on February 29, 2012. Bill Laing, Summary of Windows Azure Service Disruption on Feb 29, 2012, WINDOWS AZURE TEAM BLOG (Mar. 9, 2012, 6:03 PM PST), http://blogs.msdn.com/b/windowsazure/archive/2012/03/09/summary-of-windows-azure-service-disruption-on-feb-29th-2012.aspx (last visited Feb. 3, 2013) (on file with the Washington and Lee Law Review).

94. Martin, supra note 44, at 297–98; see also Kattan, supra note 8, at 623 (noting that a customer who moves data storage and processing onto the cloud may have difficulty if he later decides to revert to the PC model). Martin notes that this lock-in problem is likely to not apply to IaaS because a customer of an IaaS provider will typically have everything on a virtual machine over which the customer can exercise full control. Martin, supra note 44, at 294.
Another disadvantage is related to the risk of loss. If a provider fails to secure data and a consumer's information is compromised, the risk of loss is likely to fall on the consumer rather than the cloud service provider.

In this Article, we emphasize the need for data control in the cloud, which we define as consisting of the ability to withdraw data (data withdrawal) and move data to a new location (data mobility). We argue that data control is essential for meaningful consumer choice. Consumers will inherently have less control over data stored in the cloud, but being able to choose (and switch to) providers that are more reliable or that offer stronger security measures is important for preserving consumer

---

95. See Lanois, supra note 18, at 44 (citing a publication of the World Privacy Forum). Privacy is likely to be especially important to consumers in the context of electronic health records. See Colin P. McCarthy, Note, Paging Dr. Google: Personal Health Records and Patient Privacy, 51 WM. & MARY L. REV. 2243, 2253 (2010) (discussing the potential problems of personal health records). These concerns are not just limited to health services. Confidentiality is a significant concern to a number of other professions when considering the adoption of cloud services as well. See Cascia, supra note 17, at 884 (noting that outsourcing IT management to third parties may make it more difficult for schools to make sure that the personal information of students remains private); Martin, supra note 44, at 295. Martin mentions the legal field by name as one industry that should be hesitant at this point when considering whether to use cloud services in support of its practices. Id. at 300. It is also unclear how the Fourth Amendment will apply to information held by third party cloud service providers. See id. at 295–96; see also Soghoian, supra note 5, at 361 (noting that cloud computing leaves users vulnerable to invasions of privacy by the government, resulting in “evisceration of traditional Fourth Amendment protections of a person’s private files and documents”). Martin also notes concerns that the federal statute governing electronic messaging may be difficult or unable to apply to modern technology. Martin, supra note 44, at 295–96.

96. See Soghoian, supra note 5, at 378–79 (discussing why cloud computing providers have little incentive to protect users); see also infra Part IV.A (discussing contents of TOS agreements, including explicit limitations on providers’ legal liability).

97. See Kattan, supra note 8, at 623 (noting the customer’s dependence on cloud service providers to protect the customer’s data); Martin, supra note 44, at 289 (noting customers’ lack of control over data and the security practices of the cloud vendor); Stylianou, supra note 44, at 595 (explaining that some private data will be transferred away from the user’s immediate physical control); Wittow & Buller, supra note 7, at 6 (noting the lack of control that users have over data in the cloud and the importance that the user be able to trust the cloud service provider).
autonomy. Currently, there are systemic limitations to meaningful choice. SaaS customers may experience lock-in problems because a cloud provider may store the customer’s information in a format unique to the cloud provider and thus make it difficult for the customer to switch cloud providers later.98 This control over content also leads to some concerns about private censorship. Werbach notes the existence of concerns over cloud services having too much power to censor controversial causes, such as when Amazon Web Services dropped Wikileaks as a customer.99

D. Cloud Computing Legal Issues

For our purposes, there are two important categories of legal issues raised in the context of cloud computing: data use and procedural issues. Data use issues could include the use of both public and private information, thus our use of the term “data use” also includes privacy concerns, examined in more detail below. Procedural issues relating to cloud computing can include E-Discovery and jurisdiction questions. The appropriate degree of regulation is also in controversy, so even if we could identify all of the possible legal issues related to cloud computing, it may prove difficult to effectively regulate the industry.100

One data use issue is the problem of “scraping,” specifically the question of how courts should deal with the unauthorized, automated collection of information by, for example, auction services that list relevant auctions in one search across multiple

---

98. Martin, supra note 44, at 297–98; see also Kattan, supra note 8, at 623 (noting that a customer who moves data storage and processing onto the cloud may have difficulty if they later decide to revert to the PC model). Martin notes that this lock-in problem is likely to not apply to IaaS because a customer of an IaaS provider will typically have everything on a virtual machine over which the customer can exercise full control. Martin, supra note 44, at 294.

99. See Werbach, supra note 8, at 1820 (“From a broader perspective, though, the rise of cloud computing changes a default assumption that data will be within the control of the user.”).

100. See id. at 1766 (referring to network neutrality as the “final hurrah” of the regulatory framework under the Telecommunications Act, as views of the industry have shifted “from regulated monopoly to managed competition within defined industry segments”).
auction websites. Claims relating to scraping have been brought based on the Computer Fraud and Abuse Act (CFAA), the tort of trespass, and a “hot news” theory. An analysis of these options and whether they provide adequate means of redress for companies whose data is mined poses an interesting research question for future research. Our concern about the privacy of individual users also makes us question whether recourse for “scraping” might also apply to protect individuals whose data is mined without their consent, though this is outside the scope of our research.

1. Privacy

Our primary focus in this Article is on the implications of cloud computing and corresponding privacy agreements on personal privacy. There are several legal issues relating to privacy and cloud computing, including the uncertain applications of the Health Information Portability and Accessibility Act (HIPAA), the Stored Communications Act, and the Fourth Amendment, especially the third-party doctrine of Fourth Amendment jurisprudence. If a legal regime is put into place to provide stronger privacy protections, it is unclear

---

101. See Wittow & Buller, supra note 7, at 8–9 (discussing how the scraping issue impacts cloud computing).
103. See id. (discussing how the scraping issue impacts cloud computing).
105. See Electronic Communications Privacy Act of 1986, Pub. L. No. 99-508, 100 Stat. 1848 (codified as amended in scattered sections of 18 U.S.C.); Werbach, supra note 8, at 1819 (noting that a search warrant is required to access e-mail stored on a user’s hard drive, but that under the Electronic Communications Privacy Act, a lower standard would be applied if that same e-mail had been stored on Google’s Gmail servers for more than six months).
106. See Stylianou, supra note 44, at 596–97 (“[I]t is still debatable whether access to online stored data should be considered a search . . . or whether by communicating data to a remote server the subject is considered to have knowingly exposed the information.”).
whether data collection should be addressed based on the quantity collected or the type collected, and there is also a lot of uncertainty about how to address the transfer of data between countries with different privacy laws.\textsuperscript{107}

Many aspects of the privacy debate rely on an understanding of privacy theories. Several things influence privacy protections online, including social norms, website architecture, and the law.\textsuperscript{108} Some note that there are societal obstacles to strengthening privacy protections online, arguing that the younger generation values the interconnectedness and low cost of cloud services more than they value their personal privacy.\textsuperscript{109} Werbach asserts that the range of concerns about cloud providers’ information practices goes beyond our current concept of “privacy,” and suggests referring to it as “information governance.”\textsuperscript{110} In lieu of creating a new category, Solove suggests revising the concept of “privacy” to encompass these concerns.\textsuperscript{111} It is likely that there will be an increase in public policy activity in this area in the near future,\textsuperscript{112} underscoring the importance and timeliness of this topic. A significant problem that arises when dealing with technologically sophisticated policy issues, however, is that some judges and other policy makers may be ill-

\textsuperscript{107}. See id. at 595–96 (discussing whether access to cloud data is a search).


\textsuperscript{109}. See Robison, supra note 4, at 1237–38 (“[Y]ounger users are more likely to embrace the Internet’s interconnectedness and convenience by participating in social networking, sharing digital content, and using cloud services.”). But see Chris Hoofnagle, Jennifer King, Su Li, & Joseph Turow, How Different Are Young Adults from Older Adults When It Comes to Information Privacy Attitudes & Policies? 20 (Working Paper Series, 2010), available at http://ssrn.com/abstract=1589864 (noting that their study results failed to show the expected significant differences between the behavior of young adults and older adults online with regard to privacy).

\textsuperscript{110}. Werbach, supra note 8, at 1833.


\textsuperscript{112}. See Werbach, supra note 8, at 1835 (“Public policy activity in this area seems bound to increase.”).
informed about the underlying technology, leading these policy makers to hesitate when faced with current issues.\textsuperscript{113}

There may also be legal harms arising from data gathering practices. Richards discusses the “database problem,” in which there are very large databases that make it efficient and valuable for businesses to use consumer information, but the legal rights of the consumers in these databases are unresolved.\textsuperscript{114} Stylianou acknowledges that cloud computing does result in more private information being collected and this could be harmful, but concludes that most of this increase in information collection happens voluntarily, and that the compromises in privacy appear to be no greater than necessary for the delivery of cloud services.\textsuperscript{115} Some were critical of the settlement in Authors Guild v. Google\textsuperscript{116} for its lack of restrictions concerning data gathering, arguing that privacy issues should be addressed in the settlement to protect people from having their reading choices readily available to third parties.\textsuperscript{117}

2. Jurisdiction

Jurisdiction issues concerning a court’s ability to hear a claim will arise in the context of the cloud for two reasons: (1) the

\textsuperscript{113} For example, in the oral arguments of City of Ontario v. Quon, Justices Roberts and Scalia noted their confusion as to how wireless communications are transmitted, with both indicating that they were not aware that these messages were inherently processed by a third party. See Transcript of Oral Arguments at 48–50, City of Ontario v. Quon, 130 S. Ct. 2619 (2010) (No. 08-1332), http://www.supremecourt.gov/oral_arguments/argument_transcripts/08-1332.pdf (exemplifying the confusion of Justices Roberts and Scalia as to how wireless communications are transmitted).


\textsuperscript{115} See Stylianou, supra note 44, at 594–96 (discussing voluntary information collection).


\textsuperscript{117} See Denny, supra note 46, at 238–39 (“Much of the recent debate surrounding cloud computing and privacy stems from a settlement in Authors Guild v. Google Inc.”); Wittow & Buller, supra note 7, at 7 (“Privacy concerns also have been raised in the context of the pending Authors Guild v. Google [Inc.] book search settlement, which creates a cloud-based database of searchable books.”).
lack of borders in cyberspace; and (2) the vast differences between privacy laws in different locations. If a conflict arises with respect to a cloud service, where could that conflict be resolved? If there is a conflict between a customer and cloud provider within the United States, the customer might be bound by arbitration language in a TOS agreement, or by a choice of law or venue clause.

But what about more geographically vague situations? Some discussions about jurisdiction assume that the applicable law will be determined by the physical location of the data, but this information is often unknown to the customer. Sometimes, a defendant may claim that he has insufficient contacts with the forum state for a particular court to exercise jurisdiction. Because of these jurisdictional problems, it is important that the TOS agreements for cloud services specify where data will be stored and which laws will apply. Otherwise, the uncertainties related to jurisdiction in the cloud may chill some online activity by discouraging people from engaging in electronic commerce.

Approaches to informational privacy can vary between nations, and the United States as a whole has a privacy law

118. See Stylianou, supra note 44, at 596 (discussing the transfer of data between countries).


120. See Lanois, supra note 18, at 44 (“[D]ata that might be secure in one country may not be in another, and in many cases, users of cloud services do not know where their information is being held.”); Stylianou, supra note 44, at 602 (“Because different national laws accord different levels of protection to personal and private information, it is important that users know where their data is stored.”).

121. See Pinguelo & Muller, supra note 46, at 1 (“It is apparent that the use of a cloud can potentially increase the number of ‘contacts’ a party is found to have for personal jurisdiction purposes, and thus raise its exposure to lawsuits in multiple forums.”).

122. See Denny, supra note 46, at 239 (“According to the Privacy Authors, if readers were worried that information about their reading habits could be disseminated to the government, divorcing spouses, or other interested third parties, these readers would be less likely to view books on controversial topics.”).

regime that is much less protective of personal privacy than that of the European Union. 124 Can a court in the European Union exercise jurisdiction over a U.S. company that violates the personal privacy of EU citizens? Generally, the answer will be yes, based on principles of jurisdiction.

In the international context, jurisdiction can be described as the right of one country to regulate actions that are not solely conducted within that nation’s borders.125 Three categories of international jurisdiction are legislative jurisdiction, under which a nation’s laws can apply to cases with a foreign element; adjudicative jurisdiction, when the nation’s courts have the power to try cases involving a foreign element; and enforcement jurisdiction, when the nation has the power to act in another nation’s territory to enforce its own laws.126

Exercise of adjudicative jurisdiction may be justified when the acts were committed or completed within the nation’s territory, when the perpetrator or victim was a citizen of that nation, when the act has effects within that nation (a justification that is commonly criticized for its open-endedness), or when the act jeopardizes the nation’s sovereignty.127 Because adjudicative jurisdiction can be found when the victim of a wrong is a citizen of the adjudicating nation, this means that service providers in

---

124. See Stylianou, supra note 44, at 597 (noting that the use of the Safe Harbor agreement allows U.S companies to process the data of European citizens). This agreement is in lieu of a privacy law overhaul to make the U.S. approach to privacy match the approach of the EU. Id.

125. See Kuner, Part 1, supra note 119, at 178–79 (defining international jurisdiction as “the State’s right under international law to regulate conduct in matters not exclusively of domestic concern.” (citation omitted)).

126. See id. at 184 (discussing categories of jurisdiction). Generally, direct enforcement of one nation’s laws in another nation is not permitted, though a nation may apply its domestic law to conduct that occurs elsewhere, provided recognized legal grounds exist for doing so. Id. at 185. Enforcement jurisdiction, however, is rarely found. See Christopher Kuner, Data Protection Law and International Jurisdiction on the Internet (Part 2), 18 INT’L J. L. & INFO. TECH. 227, 232 (2010) [hereinafter Kuner, Part 2] (“[A] State may not carry out an investigation in another State, if the purpose is to enforce its own administrative, criminal, or fiscal law. These restrictions apply even if the persons or entities in the second State consent to the first State’s enforcement actions.”).

127. See Kuner, Part 1, supra note 119, at 188–90 (examining adjudicative jurisdiction in detail).
the United States must act carefully to comply with the privacy laws of other jurisdictions when a customer is a foreign citizen.

**E. Calls for Action in the Cloud**

The current legal regime applicable to cloud computing has drawn a lot of criticism from organizations that want the law to consider current technologies.\(^{128}\) Legislative reform will likely be necessary to address the new environment created by cloud computing, but such reform will need to take into account many different concerns.\(^{129}\) For example, reforms will need to take data protection into consideration because customers are likely to want data stored in the cloud to be protected the same as it would be on the customer’s own tangible storage devices.\(^{130}\)

The Electronic Communications Privacy Act (ECPA)\(^{131}\) is examined in detail below in Part III.B.2. Several institutions have urged lawmakers to amend the ECPA. Microsoft proposed the Cloud Computing Advancement Act (CCAA)\(^{132}\) in 2010, and the Center for Democracy and Technology has also recommended

\(^{128}\) See Kattan, *supra* note 8, at 645 (suggesting revision of the Stored Communications Act and referencing the position of the nonprofit Digital Due Process that the ECPA should be modernized and clarified); Martin, *supra* note 44, at 286 (noting recommendations made by Microsoft and the Center for Democracy and Technology). Digital Due Process is an organization that is focused on modernizing the approaches of law enforcement to electronic data, and they encourage the reformation of the ECPA to take into account recent and emerging technologies. Lanois, *supra* note 18, at 45.

\(^{129}\) See Werbach, *supra* note 8, at 1826 (“The solution to the contemporary challenges of cloud computing likely requires some legislative reform in addition to FCC action.”).

\(^{130}\) See Couillard, *supra* note 42, at 2205–06 (“Despite the shift in Internet usage, users expect their information to be treated the same on this virtual cloud as it would be if it were stored on their own computer, phone, or iPod.”).


legislative action to address cloud computing issues. The CCAA would strengthen the privacy protections of the ECPA, unifying the concepts of “electronic communications service” and “remote computing service,” and would also enhance the Computer Fraud and Abuse Act (CFAA) by presuming a loss of $500 for each count of unauthorized access. The CDT proposal, on the other hand, focuses more on civil liberties, urging Congress to amend the ECPA to require probable cause before a seizure of online information can be executed without notice.

1. Transparency and Control

Other calls for revisions of the system have focused on the need for transparency. To say that practices of cloud providers should be transparent about information use means that customers should be well-informed of what companies are doing with the customers’ personal data. In examining Internet issues, the FCC maintains that transparency is important for consumer protection in the telecommunications context. Martin suggests that when addressing cloud computing concerns, it will be important to ensure that the practices of cloud providers are understood and that customers have the ability to exercise control over their data. Transparency could have additional advantages by encouraging cloud platforms to be more interoperable, allowing for greater data portability. If the

133. See Martin, supra note 44, at 286 (discussing recently proposed legislation, standards, and governing principles).
136. Id. at 310.
137. See, e.g., Werbach, supra note 8, at 1767 (“To achieve its public interest mandates, the FCC must consider . . . [and examine] transparency.”).
138. See id. at 1837 (discussing the FCC’s adoption of a transparency mandate in its Open Internet Order).
139. See Martin, supra note 44, at 286 (“Any solution needs to incorporate guarantees that data owners would be able to gain control of their data in a usable form should their service providers become inoperable.”).
140. See Werbach, supra note 8, at 1839 (noting the ancillary benefits of transparency).
industry takes an approach to personal data that focuses on the ability of users to control their data, transparency may prove beneficial and alleviate some of the information asymmetry between cloud providers and their customers.\textsuperscript{141}

Industry leaders are conscious of transparency concerns. A consortium of industry leaders put forth the Open Cloud Manifesto, advocating the use of standardization and collaboration to develop an “open cloud.”\textsuperscript{142} The Open Cloud Manifesto focuses on transparency and interoperability between cloud providers, with one of the goals being to minimize the lock-in issue.\textsuperscript{143} If implemented, this manifesto might mitigate some of the data control issues that we are concerned about in this Article.

It could also facilitate transparency for users to be proactive about seeking information. The European Network and Information Security Agency (ENISA) suggests that users ask cloud providers about things like the provider’s personnel security procedures, use of subcontractors, operational security procedures, disaster recovery protocol, and miscellaneous legal issues like data location, jurisdiction issues, and how the customer can recover data upon termination of the service.\textsuperscript{144} We posit that users who are given the right to control their information are likely to be more involved in the process of controlling their own data.

\textbf{F. Cloud Services in Different Industries}

There are a number of professions in which practitioners are required to handle client or patient data with care, making data protection in these sensitive industries very important. One of these industries is the legal field, in which attorneys and their staff

\begin{itemize}
\item \textsuperscript{141} Schwartz & Solove, \textit{supra} note 3, at 1882.
\item \textsuperscript{142} Martin, \textit{supra} note 44, at 286.
\item \textsuperscript{143} See \textit{id.} at 310 (discussing the Open Cloud Manifesto in detail).
\item \textsuperscript{144} See \textit{id.} at 311 (examining the European Network and Information Security Agency report, which recommends a series of user procedures that can be employed for self-protection).
\end{itemize}
are required to take great steps to protect client confidentiality.\textsuperscript{145} Still, some state bar associations may recognize the convenience of cloud services and may be inclined to approve of attorney practices in which client information is stored using public cloud services.\textsuperscript{146} Martin, however, suggests that the ABA should establish ethical guidelines relating to topics like document storage, e-mail, and confidentiality in the cloud.\textsuperscript{147}

In the health care industry, there has been a shift toward using electronic medical records (EMR) as an alternative to paper records.\textsuperscript{148} A more recent push is toward maintaining personal health records (PHR) online through services like Epic, Microsoft’s HealthVault, and Google Health, in which the patient will have control over her records.\textsuperscript{149} However, PHR providers do not fall within one of the statutory categories of “covered entities” under HIPAA, so the storage and transmission of personal health information is not currently regulated by HIPAA or any of the related rules.\textsuperscript{150}

In addition to control, security of health information is also of paramount concern. McCarthy notes that the Health Information Technology for Economic and Clinical Health Act\textsuperscript{151} requires users to be notified if there is a breach threatening PHR data,

\begin{footnotesize}
\begin{enumerate}
\item[145.] Model Rules of Prof’l Conduct R. 1.6, 5.3 (1983) (discussing confidentiality and the duties of nonlawyer staff, respectively).
\item[146.] See Martin, supra note 44, at 300–01 (citing a New York bar opinion about using e-mail services that scan e-mail content to generate targeted advertising).
\item[147.] See id. at 313 (“[T]he ABA should move quickly to establish ethical guidelines for lawyers who use cloud computing services . . . [including] document storage, e-mail, collaboration, due diligence for confidentiality, and breach notification related to cloud services.”).
\item[148.] McCarthy, supra note 95, at 2250–51. EMRs, however, are generally limited to that specific provider, with no sharing of information. See id. (“Each health care provider maintains its own EMRs—physician’s offices maintain their EMRs, hospitals maintain their EMRs, and so on.”).
\item[149.] See id. at 2245, 2251–54 (“Until now, patients could request a copy of her [sic] medical records from their health care providers but have not had the opportunity to control them in the way that PHRs offer.”).
\item[150.] Id. at 2258.
\end{enumerate}
\end{footnotesize}
and that the HHS has also promulgated a rule that requires PHR vendors to comply with notification requirements if a breach occurs.\textsuperscript{152} The increased vulnerability of data in the cloud necessitates strong protections for PHR, like encryption, password protection, and authentication requirements.\textsuperscript{153} One of our recommendations for regulating cloud providers focuses on establishing baseline standards for data protection, which could help address some of these issues.

\section*{III. Privacy Fundamentals}

\subsection*{A. Privacy Theories}

“Privacy” is an example of a word that can mean many different things.\textsuperscript{154} It can be a handmade sign on the door of a teenager’s room prohibiting entry by parents and little brothers. It can be the right to make one’s own decisions without undue burden imposed by the government. On the Web, some people might consider social networking posts “private” if they are only viewable by the poster’s four hundred closest friends,\textsuperscript{155} while others do not consider anything that they do on the Web “private” unless all data is heavily encrypted and all of their traffic is routed through an anonymizer.\textsuperscript{156}

\textsuperscript{152} See McCarthy, supra note 95, at 2263–64 (discussing new federal law governing PHR privacy and security).
\textsuperscript{153} See id. at 2267 (“PHRs should be required to employ best practices in data encryption, password protection, and authentication in order to safeguard PHI stored on their servers.”).
\textsuperscript{154} See Anita L. Allen, Privacy-As-Data Control: Conceptual, Practical, and Moral Limits of the Paradigm, 32 Conn. L. Rev. 861, 864 (2000) [hereinafter Allen, Data Control] (noting the wide variation in how “privacy” is defined, even among people who seemingly are talking about the same privacy paradigm of privacy being data control).
\textsuperscript{155} Some argue, however, that such postings are still functionally private because of the boundaries that exist by making a posting viewable only by certain people. See Richards & Solove, supra note 1, at 1920–21 (citing Lior Jacob Strahilevitz, A Social Networks Theory of Privacy, 72 U. Chi. L. Rev. 919 (2005)). This view arguably does not consider the potential of screenshots of “friends only” postings being reposted elsewhere.
\textsuperscript{156} These three categories of privacy have been referred to as physical and proprietary privacy, decisional privacy, and informational privacy. Allen, Data
Some view privacy as a negative freedom, providing a freedom from something instead of a claim to something else.\textsuperscript{157} Perhaps the most prevalent view of privacy over the years has been the secrecy paradigm of privacy, where privacy is limited to things that are secret.\textsuperscript{158} There is also an “invasion conception” of privacy, where privacy violations are viewed as invasions of an interest.\textsuperscript{159} Some view privacy as referring to inaccessibility, when a person or information about her is inaccessible to others.\textsuperscript{160} Some also address what sort of harm is necessary to find a privacy violation. Solove asserts that there can be an infringement of privacy “even if no secrets are revealed and even if nobody is watching us,” connecting the concepts of privacy and human dignity.\textsuperscript{161}

The importance of privacy is sometimes stated in grandiose terms, tying the concept of privacy to democratic ideals like

\textit{Control, supra} note 154, at 865–66.


\textsuperscript{158} See Solove, \textit{Taxonomy, supra} note 111, at 497–98 (“Under the secrecy paradigm... if the information is not previously hidden, then no privacy interest is implicated by the collection or dissemination of the information. In many areas of law, this narrow view of privacy has limited the recognition of privacy violations.”). This paradigm can be seen in the approach courts have taken to the Fourth Amendment, as well as in the tort of intrusion upon seclusion. \textit{Id.} Solove takes the view that the secrecy paradigm approach to information privacy law is outmoded. \textit{SOLOVE, DIGITAL PERSON, supra} note 2, at 143.

\textsuperscript{159} See \textit{SOLOVE, DIGITAL PERSON, supra} note 2, at 8 (defining the invasion conception of privacy). Solove says that the Warren and Brandeis theory of privacy falls within this conception of privacy, with a focus on the existence of discrete wrongs to individuals. \textit{See id.} at 93–94 (discussing the two models for the protection of privacy). Solove also criticizes the invasion conception of privacy by arguing that it overlooks the structural nature of certain privacy problems that affect not just an individual, but also society as a whole. \textit{See id.} at 97.

\textsuperscript{160} See Allen, \textit{Data Control, supra} note 154, at 867 (“[O]ther than in contexts in which ‘privacy’ holds its decisional and proprietary meanings, privacy refers to a degree of inaccessibility of a person or information about her to others’ five senses and surveillance devices.”); Allen, \textit{Coercing Privacy, supra} note 157, at 724 (“Privacy obtains where persons and personal information are, to a degree, inaccessible to others.”).

\textsuperscript{161} \textit{SOLOVE, DIGITAL PERSON, supra} note 2, at 44, 55.
independent thought and the right to take political actions.\textsuperscript{162} Alan Westin, an early information privacy scholar, defined privacy as “the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others.”\textsuperscript{163} The law has taken a number of approaches to address different concerns associated with privacy. The right to privacy has been recognized in the United States for over a century, though coherent definitions have generally been lacking.\textsuperscript{164} Solove views privacy as a concept that encompasses many different kinds of distinct but interrelated issues.\textsuperscript{165}

The concept of privacy also overlaps with constitutional protections under the Fourth Amendment, where the focus is on a “reasonable expectation of privacy.”\textsuperscript{166} This legal concept is connected to several philosophical questions: what is privacy, where does it exist, and is it reasonable to expect a particular action to be private? If the government conducts surveillance somewhere that there is an expectation of privacy, a warrant is

\textsuperscript{162} See Allen, \textit{Coercing Privacy}, supra note 157, at 734 (“Liberal theorists claim that we need privacy to be persons, independent thinkers, free political actors, and citizens of a tolerant democracy.”); Solove, \textit{Taxonomy}, supra note 111, at 489 (citing Julie Cohen and Paul Schwartz for the argument that “privacy is a constitutive element of a civil society”).

\textsuperscript{163} \textsc{Alan F. Westin}, \textit{Privacy and Freedom} 7 (1967).

\textsuperscript{164} See Richards, \textit{supra} note 114, at 1155 (discussing the sometimes uneasy coexistence of privacy and speech); Paul M. Schwartz & Karl-Nikolaus Peifer, \textit{Prosser’s Privacy and the German Right of Personality: Are Four Privacy Torts Better than One Unitary Concept?}, 98 CAL. L. REV. 1925, 1963 (2010) (“[I]n their comprehensive work, Privacy, Property and Personality, [the authors] argue that the right of privacy in the United States ‘remains somewhat conceptually uncertain and poorly defined.’” (quoting \textsc{Huw Beverley-Smith \textit{et al.}}, \textit{Privacy, Property and Personality} 207 (2005))); Solove, \textit{Taxonomy}, \textit{supra} note 111, at 562 (“But our understanding of privacy remains in a fog, and the law remains fragment and inconsistent.”).

\textsuperscript{165} Richards & Solove, \textit{supra} note 1, at 1914–15; Solove, \textit{Taxonomy}, \textit{supra} note 111, at 562.

\textsuperscript{166} See Rebecca N. Cordero, \textit{No Expectation of Privacy: Should School Officials be Able to Search Students’ Lockers Without Any Suspicion of Wrong Doing?}, 31 U. BALT. L. REV. 305, 308 (2002) (“In his concurrence, Justice Harlan coined the term a ‘reasonable expectation of privacy’ to describe an area subject to the protection of the Fourth Amendment.”).
necessary to protect against unreasonable intrusion.\textsuperscript{167} Generally, public surveillance is not viewed as an intrusion because behaviors are being exposed to the public, but there may be exceptions when such surveillance is overzealous.\textsuperscript{168} As one court said, “The mere fact that a person can be seen by someone does not automatically mean that he or she can legally be forced to be subject to being seen by everyone.”\textsuperscript{169}

The Fourth Amendment protection against unreasonable searches and seizures, the protections afforded to electronic communications under the ECPA, and privacy torts are three large legal categories for the concept of privacy.\textsuperscript{170} As we examine in later sections, the application of the Fourth Amendment and the ECPA to the Information Age is far from clear. Additionally, there is also a sense that privacy tort law is ineffective at addressing these issues.\textsuperscript{171} The traditional model for privacy protection simply does not address the sorts of privacy problems that have arisen recently.\textsuperscript{172}

The desire for privacy is arguably an innate human trait, and privacy theorists thus often make philosophical or literary allusions when explaining the importance of privacy. One of the most vivid images for the modern information privacy problems is Jeremy Bentham’s design for a prison that he called the

\begin{footnotesize}

\textsuperscript{168} See Solove, Taxonomy, supra note 111, at 498 (“In some cases, however, courts have recognized a harm in public surveillance.”).


\textsuperscript{170} Other relevant elements of constitutional law include the freedom of association and the freedom of anonymous speech under the First Amendment. See SOLOVE, DIGITAL PERSON, supra note 2, at 64–65 (discussing the right to privacy).

\textsuperscript{171} See Richards & Solove, supra note 1, at 1889 (“Today, the chorus of opinion is that the tort law of privacy has been ineffective, particularly in remedying the burgeoning collection, use, and dissemination of personal information in the Information Age.”).

\textsuperscript{172} See id. at 1918 (“Tort law has not emerged as the leading protector of privacy.”). Solove argues that many of the privacy problems we confront today are systemic in nature, stemming from information flows, with multiple actors being responsible for these problems. Daniel J. Solove, Identity Theft, Privacy, and the Architecture of Vulnerability, 54 HASTINGS L.J. 1227, 1232 (2003) [hereinafter Solove, Architecture].
\end{footnotesize}
Panopticon. In the Panopticon, prison cells are distributed around a central observation tower, and someone placed in the tower can monitor all of the prison cells without the prisoners knowing when they are being observed, and this fear of observation leads to the prisoners behaving better. In the context of the Internet, Schwartz has argued that there is a danger both of a government Panopticon and private Panopticons operated by private entities that collect and use information while resisting attempts at transparency.

Privacy concerns gained more public visibility in the early 1980s, perhaps due to the era’s relationship with George Orwell’s dystopian novel *Nineteen Eighty Four*. Similar to the Panopticon, the telescreens of *Nineteen Eighty Four* allowed the government to monitor citizens without their knowledge that they were being observed. Perhaps thanks in part to this work of fiction—and the fact that it is required reading for many high school seniors—U.S. citizens are keenly aware when government action has the potential to intrude on privacy and lead to an authoritarian state.

---

173. See Michel Foucault, *Discipline and Punish: The Birth of the Prison* 201 (Alan Sheridan trans., 1977) (listing the essential elements of the Panopticon’s effectiveness being visibility and unverifiability, visibility referring to that of the tower, and unverifiability referring to the prisoners’ inability to know whether they are being observed).

174. Id. at 201; Solove, *Architecture*, supra note 172, at 1240. Solove also notes Foucault’s argument that the Panopticon represents power relations in society. Id. at 1240.


176. See Schwartz & Solove, supra note 3, at 1825–26 (“Part of this attention was driven, in turn, by the arrival of George Orwell’s titular year, 1984.”).


1. Warren and Brandeis

Theoretical discussions of privacy law often begin in 1890, when Samuel Warren and Louis Brandeis published their article about privacy as a right of personality. Warren and Brandeis were especially concerned about the use of private information by the media and the implications of technological developments like new and cheaper photography technologies. Warren and Brandeis also argued that a person’s intellectual property was not a matter of private property, but rather was related to the person’s “inviolate personality.”

Warren and Brandeis were supportive of the idea of applying the common law to protect a right to privacy, which they famously summarized as a “right to be let alone.” In their article, they supported enforcing the right of privacy using tort damages to provide individuals with compensation for the “mental suffering” caused by the privacy invasions. This view

complete understanding of the issues, I turn to . . . Franz Kafka’s depiction of bureaucracy in The Trial.”). In The Trial, the protagonist is under arrest and does not understand why because the bureaucratic government has a large amount of information about the protagonist that it refuses to share with him. See id. at 8–9 (discussing Kafka’s novel The Trial).

179. See Samuel D. Warren & Louis D. Brandeis, The Right to Privacy, 4 HARV. L. REV. 193, 195 (1890); see also Richards & Solove, supra note 1, at 1888 (writing that Warren and Brandeis popularized privacy in American law with their famous article in 1890); Schwartz & Solove, supra note 3, at 1819 (recognizing the Warren and Brandeis article as a famous example of privacy’s early jurisprudence).

180. See Richards, supra note 114, at 1198 (“[M]odern thinking about the right of privacy is often traced to Warren and Brandeis’s privacy article, in which their concern was not primarily data privacy, but rather media use of private information.”); Schwartz & Solove, supra note 3, at 1819 (“The paradigmatic privacy invasion for Warren and Brandeis concerned the press intruding on the privacy of individuals by printing gossip about them.”). This position leads to a balancing of the interest in privacy against interests under the Freedom of the Press Clause of the First Amendment. Richards & Solove, supra note 1, at 1892; Solove, Architecture, supra note 172, at 1229.

182. Richards & Solove, supra note 1, at 1891.
183. Solove, Architecture, supra note 172, at 1229–30. Warren and Brandeis expressed a preference for money damages over injunctions, which they noted may be appropriate in narrow circumstances, and asserted that narrower circumstances would be required for criminal penalties to be appropriate. Id. at
of privacy harms focuses on dignitary harms, like harm to reputation, based on the concept that privacy violations are a type of invasion to the victim’s dignity.\(^\text{184}\)

Case law on privacy was heavily influenced by the Warren and Brandeis article, especially when the dispute involved the use of photographs of ordinary people to promote a company’s product.\(^\text{185}\) However, by the time the Warren and Brandeis article was fifty years old, privacy was still a very minor doctrine in tort law. Only twelve states recognized the right of privacy by common law, and only two recognized it by statute.\(^\text{186}\)

2. Prosser

For modern privacy scholars, the next major development in the privacy law of the United States was the 1960 publication of William Prosser’s article, *Privacy*.\(^\text{187}\) In this article, Prosser argued that there were four categories within privacy tort law: appropriation privacy, intrusion privacy, unauthorized public

\(^{1230}\)

184. See Solove, *Taxonomy*, supra note 111, at 487; Solove, *Digital Person*, supra note 2, at 93–94 (referring to the privacy theory of Warren and Brandeis as being based on an “invasion conception” of privacy, which turns on the existence of discrete wrongs to individuals).

185. See Richards & Solove, *supra* note 1, at 1892–93 (“Warren and Brandeis’s approach to privacy was in one sense profoundly conservative, as it was part of a broader legal strategy employed by late-nineteenth-century elites to protect their reputations from the masses in the face of disruptive social and technological change.”).

186. See *id.* at 1895 (discussing the number of states that recognize certain common law privacy rights).

Disclosure of private facts, and false light. In constructing these four categories, Prosser analyzed hundreds of privacy cases.

Prosser’s approach was thus fairly comprehensive, but the categories he created were also narrow and rigid. The rigidity was perhaps based on Prosser’s concern that privacy torts might swallow up established doctrines like defamation law and intentional infliction of emotional distress. In some ways, Prosser seemed to be skeptical of privacy laws because of their potential to interfere with the flow of information, and he would have been concerned with the balance between newsworthiness and conflicting privacy interests. Unfortunately, this rigidity also makes it difficult to apply these torts to modern information privacy problems.

188. Richards, supra note 114, at 1198–99; see also Solove, Digital Person, supra note 2, at 58; Schwartz & Peifer, supra note 164, at 1941; Schwartz & Solove, supra note 3, at 1820.

189. Richards & Solove, supra note 1, at 1889.

190. See id. at 1890 (“His skepticism about privacy, as well as his view that tort privacy lacked conceptual coherence, led him to categorize the law into a set of four narrow and rigid categories.”).

191. See id. at 1890, 1900 (describing Prosser’s concern that privacy law’s “haphazard development threatened to swallow up established doctrines, such as defamation law, as well as new doctrines, such as intentional infliction of emotional distress, that he felt had more promise”). The intrusion into seclusion tort includes, as its main element, intentional infliction of emotional distress, but does not require a showing of extreme outrage, serious mental harm, or a showing that injuries were nontrivial. Id. at 1890. Prosser expressed concern that the false light and disclosure of private facts torts involved an examination of reputation, overlapping with defamation law. Id. Prosser was also concerned about privacy torts being overbroad and interfering with freedom of speech and of the press. Id.

192. See Schwartz & Peifer, supra note 164, at 1956–57 (“Due to Prosser’s strong belief in liberal flows of information, moreover, his article reflects a strong undercurrent of skepticism about the legal protection of privacy.”).

193. See Richards & Solove, supra note 1, at 1904 (“[W]hile Prosser gave tort privacy a legitimacy it had previously lacked, he also fossilized it and eliminated its capacity to change and develop.”). But see Schwartz & Peifer, supra note 164, at 1929 (disagreeing with some of Richards and Solove’s criticism of Prosser, asserting that if it had not been for Prosser, privacy would likely be much less protected in the United States); id. at 1983 (“Rather than creating an ossified privacy concept, Prosser’s contribution generated useful doctrinal categories where there previously had been unclassified cases and a lingering air of skepticism towards the tort.”).
Breach of confidentiality has been described as a tort that addresses privacy violations in specific contexts.\textsuperscript{194} Though the concepts are related, Prosser did not include breach of confidentiality in his categories of privacy torts.\textsuperscript{195} This may be because Prosser drew a line between privacy and what would be addressed under agency or contract law. Compared to Warren and Brandeis, Prosser was also less focused on the idea of personality as a justification underlying privacy protections.\textsuperscript{196}

\textit{a. Prosser’s Privacy Torts and Information Privacy}

Prosser’s torts differ significantly from each other. In Prosser’s article, he points out that intrusion and disclosure both require an invasion into something secret, which is not required of false light or appropriation.\textsuperscript{197} Disclosure and false light have publicity as an essential element, while intrusion and appropriation do not (though appropriation usually involves publicity).\textsuperscript{198} Additionally, only false light requires falsity, and only appropriation requires that the defendant have gained some advantage from the use.\textsuperscript{199}

The tort of appropriation has evolved somewhat from Prosser’s time. When Prosser originally wrote \textit{Privacy}, he noted

\begin{itemize}
  \item 194. See \textit{Solove, Digital Person}, supra note 2, at 77 (“The common law tort of breach of confidentiality . . . enables people to sue for damages when a party breaches a contractual obligation (often implied rather than express) to maintain confidentiality.”).
  \item 195. See \textit{Richards & Solove, supra note 1}, at 1909 (“A second notable omission from Prosser’s taxonomy was the tort of breach of confidence.”). In Prosser’s treatise, Prosser addressed this concern by stating, “The right of privacy, as such, is to be distinguished from liability found upon the breach of some confidential or fiduciary relation . . . .” \textit{Id.} at 1910 (quoting \textit{William L. Prosser, Handbook of the Law of Torts} 1062 (1st ed. 1941)).
  \item 196. One of Prosser’s contemporaries was Edward Bloustein, who differed from Prosser in the degree to which the former argued for the idea that privacy law protects an “inviolable personality.” \textit{Schwartz & Peifer, supra note 164}, at 1945.
  \item 197. See \textit{Prosser, supra note 187}, at 407 (discussing and defining common facets of privacy).
  \item 198. See \textit{id.} (examining false light in relation to privacy generally).
  \item 199. See \textit{id.} (discussing false light in detail).
\end{itemize}
that under the common law, all four of the recognized categories of privacy claims were specific to the individual and were not assignable, though three states at the time did recognize under statute that a publication-based claim could be brought after a person’s death.200 On the other hand, the modern right of appropriation allows a likeness to be treated as descendible property.201

These torts have questionable utility in the modern context of information privacy. The privacy tort of invasion typically requires the invasion to be of an offensive nature, but a lot of information collection appears largely innocuous.202 In Shibley v. Time, Inc.,203 the litigation concerned the magazine’s sale of their subscriber information to advertisers, but this sale was found to not meet the injury requirements for Ohio’s common law invasion of privacy tort.204 Courts have also rejected the theory that the tort of appropriation could apply to the data collection problem.205 Thus, it is likely that addressing personal privacy issues will require either the revision of privacy torts or the introduction of new alternatives.

3. Modern Informational Privacy Theory

While there are many types of privacy, the type that we are most concerned with is informational privacy and the right of a person to keep information about herself from being used by

200. See id. at 408 (discussing the qualities of invasion that would constitute a tort).
201. See Schwartz & Peifer, supra note 164, at 1965 (“The overwhelming majority of states in the United States have also recognized a postmortem dimension to the publicity right.”). Schwartz and Peifer point to the example of Elvis, whose publicity rights have been sold and resold multiple times since his death. Id. at 1966.
202. See Richards & Solove, supra note 1, at 1919 (stating that many “privacy torts—public disclosure, intrusion, and false light” require a showing that the privacy invasion be highly offensive).
204. See Richards & Solove, supra note 1, at 1919.
205. Dwyer v. Am. Express Co., 652 N.E.2d 1351, 1357 (Ill. App. Ct. 1995) (ruling that there was not an appropriation claim when American Express sold customer names to merchants); Richards & Solove, supra note 1, at 1919.
others. Our understanding of informational privacy in the modern context also needs to take social norms into consideration. In the United States, the younger generation seems to value privacy much less than the older generations, though some empirical research casts doubt on the idea that there is a meaningful difference between how different age groups view and prioritize privacy. Solove has become a leader in modern informational privacy theory, with some scholars asserting that Solove is Prosser's modern heir. Not entirely dissimilar from Prosser's approach, Solove divides privacy problems into four categories: information collection, information processing, information dissemination, and invasion.

Informational privacy implicates the potentially conflicting interests of content owners and content users. If protections of personal privacy are too strong, businesses that use customer information to target people who would be interested in a new product may be prevented from doing so. There are some First Amendment concerns as well. Chiefly among them is to what extent do I have the right to use the law and the courts to prevent you from speaking about me? There is also a question of accountability to prevent legitimate privacy regulation from applying in ways that are not optimal for society.

Schwartz argues that privacy is a “constitutive value” that is valuable not for its own virtues or effects, but because some degree of protection for personal privacy is a necessary condition

---

206. See Allen, Coercing Privacy, supra note 157, at 736–37 (“Generational differences in the taste for privacy may be significant in the United States, as younger Americans appear to be learning to live reasonably well and happily without privacy.”).

207. Hoofnagle, King, Li & Turow, supra note 109, at 20.

208. See Schwartz & Peifer, supra note 164, at 1940 (“In this sense, Daniel Solove proves the modern heir of the Berkeley Dean.”).

209. Solove, Taxonomy, supra note 111, at 489.


211. See Allen, Data Control, supra note 154, at 861 (citing Schwartz for two important normative questions facing contemporary privacy theorists: how to protect privacy while preserving accountability, and the appropriate role of the state in regulating personal privacy).
for a society that values individual identities and deliberative democracy.\textsuperscript{212} Taking the view that privacy is of vital importance, Allen argues that privacy can be lost through voluntarily giving it up, and that this raises similar moral and policy implications as someone who voluntarily sells himself into slavery.\textsuperscript{213} In the same way that people are forced to be free by being prohibited from selling themselves into slavery, Allen argues that people should be forced to be private to better allow them to “reap the full dignitarian and political consequences of privacy.”\textsuperscript{214}

\textit{a. Concepts of Privacy}

Different approaches to privacy find the value of privacy in different things. Under a communitarian view of privacy, privacy is valuable because it protects a social good by allowing citizens to more effectively participate in a deliberative democracy.\textsuperscript{215} The liberal approach to privacy focuses on the individual and on personal autonomy.\textsuperscript{216} A liberal concept of privacy could be

\textsuperscript{212} See Schwartz, State, supra note 175, at 834 (“Informational privacy, whether on or off the Internet, should not be considered a right of control. Instead, it should be conceptualized as a constitutive value.”). A constitutive value is one that derives its value not from the causal effects of the value’s existence or from the existence of the value for its own sake, but for its role in “a larger complex that is itself valued.” Gerald Dworkin, The Theory and Practice of Autonomy 80 (1988). The value of privacy is occasionally the subject of discussion, with some scholars asserting that privacy is not generally worth protecting except for people who have something to hide. Jeffrey H. Reiman, Critical Moral Liberalism: Theory and Practice 171 (1997).

\textsuperscript{213} See Allen, Data Control, supra note 154, at 869 (discussing the moral and political implications associated with privacy loss). Allen has also examined whether people could be forced to be private in the same way they can be forced to be free. See Allen, Coercing Privacy, supra note 157, at 728 (“We are forced to be free. Liberal governments cannot permit us to sell ourselves into slavery. Are we forced to be private?”).

\textsuperscript{214} Allen, Coercing Privacy, supra note 157, at 752.

\textsuperscript{215} See Schwartz, State, supra note 175, at 836 (“In searching for ways to construct this strong democracy, these thinkers emphasize common participatory activities, reciprocal respect among political equals, and the development of consensus about political issues.”).

\textsuperscript{216} See Allen, Coercing Privacy, supra note 157, at 739 (“Liberal moral philosophers maintain that respecting the many forms of privacy is paramount
further broken down into categories including physical privacy, informational privacy, proprietary privacy, and decisional privacy.\textsuperscript{217}

Some view control over data as central to privacy.\textsuperscript{218} Solove points to the lack of control over data as one of the systemic problems that enables identity theft.\textsuperscript{219} One potential privacy right related to controlling data is a right to prevent access to the data.\textsuperscript{220} After personal information is collected, the person to whom the information refers typically has no control over future use of the information.\textsuperscript{221} Even if courts or policy makers decree that a person has a “right” to control his personal data, he still may lack the ability to meaningfully control his information.\textsuperscript{222}

Our recommendations address similar issues under the broad label of “data control,” including the need for data mobility and a right of data withdrawal.

It is unlikely that an individual will ever be able to exercise absolute control over his data. The government can readily access to respect for human dignity, personhood, moral autonomy, workable community life, and tolerant democratic political and legal institutions.\textsuperscript{217}

\textsuperscript{217} Id. at 723–25. Allen notes that the concept of “private choice” is stronger than the general liberal concept of “privacy.” Id. at 727–28. Allen defines informational privacy in this way: “Informational privacy obtains where information actually exists in a state of inaccessibility, whether it is locked in a file drawer, computer, or in someone’s mind. Anonymity, confidentiality, reserve, and secrecy—not merely having the choice to bring these about—are forms of privacy.” Allen, \textit{Data Control}, supra note 154, at 869.

\textsuperscript{218} See Allen, \textit{Data Control}, supra note 154, at 863 (defining privacy as data control, and the right to privacy as the right to control, and asserting that the central aim of privacy regulation should be to promote individuals’ right to control their personal data).

\textsuperscript{219} See Solove, \textit{Architecture}, supra note 172, at 1258 (“Therefore, the problem runs deeper than identity theft. It is the fact that we have so little participation in our personal data combined with the fact that it flows so insecurely and carelessly without sufficient control.”).

\textsuperscript{220} See Birnhack & Elkin-Koren, supra note 108, at 344 (“There are two principal understandings of the right to privacy in personal data: privacy as a right to control data (‘privacy as control’) and privacy as a right to prevent access (‘privacy as access’).”).

\textsuperscript{221} See Solove, \textit{Architecture}, supra note 172, at 1234 (“[P]ersonal information is not only outside our control but also is subjected to a bureaucratic process that is itself not adequately controlled.” (citation omitted)).

\textsuperscript{222} Id.
individuals’ financial information, and medical privacy is subject to the sharing of medical information between medical professionals and health insurance companies. It is thus probably fair to say that to the extent that privacy involves the control of data, this control is qualified in certain circumstances. In the online context, Schwartz goes so far as to say that control over personal information on the Internet is an illusion. When consumers are presented with take-it-or-leave-it TOS agreements on websites, there is typically no negotiation of terms, and thus no ability to exercise meaningful control. Schwartz argues that informational privacy is not just a matter of having a right to control, but instead is a matter of line drawing to shape behaviors and thus either encouraging or discouraging the use of certain categories of expression and action.

b. The First Amendment Critique

Some argue that a right of “data privacy” would conflict with the First Amendment by interfering with the dissemination of truthful information. Volokh is the most prominent proponent of the First Amendment critique, arguing that data privacy regulation amounts to “a right to have the government stop you

223. See Allen, Data Control, supra note 154, at 872 (discussing moral accountability, control, and privacy). However, though the federal government may have access to detailed financial records, this does not necessarily mean that the government can then disclose the information. In Wine Hobby USA, Inc. v. IRS, the court declined to order the government to disclose registered home wine producers under FOIA, concluding that such a disclosure would violate the registered parties’ privacy. Wine Hobby USA, Inc. v. United States IRS, 502 F.2d 133, 135 (3d Cir. 1974); Allen, Data Control, supra note 154, at 873–74.

224. See Schwartz, State, supra note 175, at 832 (arguing that simply declaring a property right in personal information will not resolve any of the major issues relating to information privacy); see also Allen, Data Control, supra note 154, at 869 (discussing the limits burdening the privacy control paradigm).

225. Solove, Architecture, supra note 172, at 1235.

226. See Schwartz, State, supra note 175, at 858 (“As a result, information privacy should not create data fortresses, but shifting multidimensional data preserves that insulate personal data from different kinds of observation by different parties.”).

227. Richards, supra note 114, at 1150–51.
from speaking about me." Cate argues that there should be full First Amendment protection for electronic information flows, and that truthful data should be allowed to flow unimpeded to prevent violations of the First Amendment. Richards counters these arguments, arguing that data privacy properly concerns economic rights, and that bringing the First Amendment into the debate wrongly makes it into a civil rights issue. Richards also points out that the First Amendment critique is weak because it does not consider the many types of "speech" that are outside the First Amendment's protection, like fraud, solicitation, antitrust law, threats, and libel.

There is some case law support for the First Amendment critique. In U.S. West, Inc. v. FCC, the Tenth Circuit analyzed the constitutionality of the FCC's interpretation of a statutory confidentiality provision, which required customers to opt in before a carrier would be permitted to share the customers' confidential information. In that case, the Tenth Circuit stated that privacy "imposes real costs on society." The court concluded that the opt-in regime for the sharing of confidential

228. See Volokh, supra note 210, at 1050–51; see also Richards, supra note 114, at 1161 (discussing the First Amendment and privacy regulation (citation omitted)).
229. FRED H. CATE, PRIVACY IN THE INFORMATION AGE 1–4 (1997); Richards, supra note 114, at 1161.
230. See Richards, supra note 114, at 1151 ("This Article takes issue with the conventional wisdom that regulating databases regulates speech, that the First Amendment is thus in conflict with the right of data privacy, and that the Constitution thereby imposes an insuperable barrier to basic efforts to tackle the database problem.").
231. See id. at 1171–73 (discussing the fact that much "speech" is outside the scope of the First Amendment and providing an alternative approach).
233. See id. at 1230.
234. See id. at 1235 (stating that privacy does not inherently constitute a substantial state interest for First Amendment commercial speech analysis purposes, and further justification is required). As part of its analysis, the court concluded that the FCC's regulation was not narrowly tailored because it did not adequately consider a less restrictive alternative, specifically an opt-out regime. Id. at 1238–39.
information promulgated by the FCC violated the First Amendment commercial speech rights of the carriers. 235

Ultimately, the First Amendment critique has many flaws, but it illustrates the sort of theoretical balancing that Prosser, Warren, and Brandeis were concerned about. Whether informational privacy is truly more of an economic or First Amendment issue is outside the scope of our research. We mention the First Amendment critique here only to emphasize that the balancing of interests is a pervasive theme in discussions of privacy theory.

c. Privacy as a Commodity

Some theorists have suggested understanding privacy as a property right. 236 Property can be described as an interest in an object in which the owner can enforce that interest against all others. 237 To law students, the concept of property is often described as a bundle of rights. A lot of property is freely alienable; that is, it can be sold, traded, and gifted as the owner sees fit. 238 Some property, like human tissue, can be donated but not sold. 239 With regard to privacy, views differ about the extent to which privacy should be alienable at all. 240 Some scholars advocate propertizing personal information, while others

235. See id. at 1230, 1240.

236. See SOLOVE, DIGITAL PERSON, supra note 2, at 77 (noting that Alan Westin took this view). But see Schwartz, State, supra note 175, at 832 (arguing that simply declaring a property right in personal information will not resolve any of the major issues relating to information privacy).

237. Schwartz, Property, supra note 10, at 2058.

238. One scholarly definition of inalienabilities posits that inalienabilities amount to “any restriction[s] on the transferability, ownership, or use of an entitlement.” Id. at 2095 (citing Susan Rose-Ackerman, Inalienability and the Theory of Property Rights, 85 COLUM. L. REV. 931, 931 (1985)).

239. See Moore v. Regents of the Univ. of Cal., 793 P.2d 479, 498 (Cal. 1990) (Arabian, J., concurring) (viewing the question of property rights in human tissue as a moral issue).

240. For example, Allen argues that privacy should not be viewed as optional, and that people should be restrained from trading away their privacy because of privacy's importance in a society that values personal identity. Allen, Coercing Privacy, supra note 157, at 729.
advocate an outright ban on data trade.\textsuperscript{241} Schwartz suggests a category that he calls “information property,” which is itself a bundle of interests made up of five areas: “inalienabilities, defaults, rights of exit, damages, and institutions.”\textsuperscript{242}

Allen argues that the expectations of people with regard to privacy have been decreasing, with people being willing to prioritize informational privacy lower than they prioritize other goods.\textsuperscript{243} There is also some disconnect between what people say they want in terms of privacy, and then what people actually do, often being quick to accept something in exchange for their personal information.\textsuperscript{244} Currently, no value is consistently assigned to personal information, and Schwartz suggests that this lack of a value contributes to the lack of appreciation that people have for their private data.\textsuperscript{245}

\begin{itemize}
\item \textsuperscript{241} See Schwartz, \textit{Property}, \textit{supra} note 10, at 2057 (recognizing that “a strong conception of personal data as a commodity is emerging in the United States” but that some legal scholars have been “suspicious of treating personal data as a form of property”).
\item \textsuperscript{242} \textit{Id.} at 2060. Schwartz suggests implementing a system that has use-transfer restrictions and an opt-in default. See \textit{id.} (“This Article’s model of propertized personal data involves the development of a hybrid inalienability consisting of a use-transfer restriction plus an opt-in default.”).
\item \textsuperscript{243} See Allen, \textit{Coercing Privacy}, \textit{supra} note 157, at 729–30 (arguing that consumer behavior and popular culture show that people prefer less privacy when using technology than other goods). Allen also notes that because of the deprioritizing of privacy, people may be more willing to trade privacy for things like entertainment, personal profit, medical care, and access to a certain community. See Allen, \textit{Data Control}, \textit{supra} note 154, at 871 (noting that people may disclose private information for various benefits and in doing so either send strong messages to policymakers that people do not value privacy in their technological encounters or stimulate them to paternalistically “coerce” privacy). An antipaternalist approach to privacy would say that privacy is a good if people desire it, but that it should not be forced upon them. On the other hand, the paternalist approach would impose privacy on people who might not want it. There are a number of laws that mandate privacy, like laws requiring people to wear clothes in public and building codes regulating the placement and design of residential housing. See \textit{id.} (noting that the idea of privacy coercion is not foreign in American law).
\item \textsuperscript{244} See \textit{Slove, Digital Person}, \textit{supra} note 2, at 80–81 (noting that, despite people's reflexive desire to protect their private data, most people take minimal precautions and would relinquish their data for money).
\item \textsuperscript{245} See Schwartz, \textit{Property}, \textit{supra} note 10, at 2076 (noting a higher appreciation for one's personal data may accompany higher market value for it).
\end{itemize}
In their book, Hagel and Singer proposed the use of “infomediaries,” a label used to describe companies that would serve as intermediaries between consumers and companies that collect their data. While discussion of this idea has not been pervasive over the last decade, the ideas underlying it form the basis for some start-up companies. Though the impact of such efforts has not yet been seen, infomediaries might be an effective private market solution to the problems related to the decline of privacy on the Internet, provided it does not prove counterproductive to view privacy as a commodity. In a similar vein, Ayres and Funk have suggested implementing a system for telemarketers in which telemarketing is shifted to an opt-in paradigm in which customers can opt in to be contacted and are also compensated for receiving telemarketing calls. Solove argues that compensation for information would not solve the problem, however, arguing that the real problem is a lack of control over data, lack of meaningful participation in the process, and lack of transparency about future data use.

B. Privacy Law

Because intrusions into privacy on the Web are so prevalent, some argue that the government should regulate the Internet to promote consumer privacy, but others worry that this could harm

246. See John Hagel III & Marc Singer, Net Worth: Shaping Markets When Customers Make the Rules 28 (1999) (suggesting a role of intermediaries in helping consumers obtain the most value in exchange for their personal information and also protecting that information from being abused).

247. See Joshua Brustein, Start-Ups Seek to Help Users Put a Price on Their Personal Data, N.Y. Times, Feb. 12, 2012, at B3 (noting the existence of start-ups that would allow people to control and maybe even profit from their “digital trails”).

248. See Schwartz, Property, supra note 10, at 2079 (noting that telemarketing is presently inefficient because it reaches an “excessively broad audience” and suggesting that opt-in programs would “add incentives to target likely customers”).

249. See Solove, Digital Person, supra note 2, at 89–90 (noting that, if people only have the right to sell their data, the result is an “all or nothing” exchange in which the consumer is not left with a viable choice between the two alternatives).
the online advertising industry and other interests.\textsuperscript{250} In this subpart, we will first review some of the issues that emerge when discussing privacy regulation, before turning to existing bodies of law to evaluate the extent to which data control issues, like data mobility and data withdrawal, may fall within current law.

1. Steps Toward Regulation of Privacy

When a social issue has to be addressed, there are two primary options: address the problem through the market and self-regulation, or have the government regulate it. The dichotomy of self-regulation versus government regulation also arises in the privacy context. Some say that the government should not regulate privacy, suggesting that it would be too paternalistic to assume that the government knows best, though others argue that self-regulation is not a viable option because of the lack of mechanisms in the market to enable the exercise of informed, meaningful choices by individuals.\textsuperscript{251}

There are a number of arguments in favor of self-regulation of privacy issues online. Birnhack and Elkin-Koren concluded from their data that law was not important in the shaping of website behavior and privacy practices, suggesting the market forces may be more effective than law at protecting privacy.\textsuperscript{252} Others say that rules to protect privacy could have negative

\textsuperscript{250} See Lanois, \textit{supra} note 18, at 34 (noting that Internet tracking and selling of personal data has pushed the government to promote greater consumer privacy and others to seek relief through the courts, resulting in sizable awards).

\textsuperscript{251} See SOLOVE, \textit{DIGITAL PERSON}, \textit{supra} note 2, at 90–91 (recognizing that although proponents of market-based solutions to privacy concerns criticize the government for paternalism, the market fails to provide adequate mechanisms for the protection of privacy).

\textsuperscript{252} Birnhack & Elkin-Koren, \textit{supra} note 108, at 378 (arguing that the “law does not appear to play an important role” in Israeli Internet privacy practices). However, the authors do not think that law is completely irrelevant, arguing that there is a circular relationship between privacy regulations and what amounts to a “reasonable expectation of privacy.” \textit{See id.} at 379 (“In the United States, data protection law plays another role. Given... the ‘reasonable expectations’ test within U.S. privacy law, concrete regulations help shape these expectations... The fact that the law requires certain measures has a large effect on data subjects’ expectations and... the reasonability of expectations.”).
effects like decreasing the information available and increasing transaction costs.\textsuperscript{253} If the private resolutions to privacy conflicts are preferred, Solove suggests that fiduciary relationships could be recognized under the law when a company collects and uses personal information.\textsuperscript{254}

On the other hand, government intervention can be very helpful in advancing change. Consider, for example, the importance of civil rights legislation in ending institutional segregation. If remedying discrimination had been left to the market to self-regulate, improvements may have been much slower. With respect to online privacy, we disagree with the conclusions of Birnhack and Elkin-Koren concerning the value of government oversight, and assert that the enforcement part of the law is of the utmost importance and was not something that these researchers examined in adequate detail.\textsuperscript{255} Additionally, while it is true that government regulation might have negative effects, this is just as true of trusting market self-regulation. In virtually any context, when faced with multiple options, there will be potential downsides to every option. Thus, the most important thing is to balance the positive and negative.

Fair Information Practices (FIPs) are often referenced as a guide for privacy regulations. When examining the core principles of privacy regimes of different governments, some patterns emerge, including an emphasis on notice, confidentiality, and data security.\textsuperscript{256} These principles also underlie the idea of FIPs, which address how to handle and use personal information,\textsuperscript{257} and often focus on responsibility and participation in the collection and use of data.\textsuperscript{258} The Federal Trade Commission

\textsuperscript{253} See Richards, supra note 114, at 1159–60 (noting that this is the view of some law and economics scholars).

\textsuperscript{254} See Solove, Digital Person, supra note 2, at 103 (recognizing that the disparities in knowledge between consumers and market participants of how consumer data is used may support a court’s finding of a fiduciary relationship).

\textsuperscript{255} See supra note 252 and accompanying text (discussing Birnhack and Elkin-Koren’s views on privacy law and the cloud).

\textsuperscript{256} See Birnhack & Elkin-Koren, supra note 108, at 350 (noting that foreign governments, like those of members of the European Union, exhibit data protection standards similar to those contained in U.S. law).

\textsuperscript{257} See Solove, Architecture, supra note 172, at 1266 (describing FIPs).

\textsuperscript{258} See id. at 1268 (describing the two focuses of FIPs: responsibility and
INFORMATION PRIVACY AND DATA CONTROL

(FTC) views FIPs as being based on five core principles: (1) notice and consumer awareness; (2) consumer choice and consent; (3) access and participation in the process; (4) data integrity and security; and (5) enforcement and redress.\textsuperscript{259} Schwartz and Solove suggest using FIPs to varying degrees, dependent on whether the personal information is identified or identifiable.\textsuperscript{260}

Whether new regulation is needed at all ultimately depends on whether the current regulatory scheme is too flawed to offer meaningful guidance. For this reason, we turn now to an examination of current U.S. privacy law.

2. Federal Privacy Statutes and State Laws

Federal privacy law focusing on consumer protection is typically narrow, often focusing on the type of records in issue or a particular industry.\textsuperscript{261} Early congressional action on privacy includes the Fair Credit Reporting Act (FCRA)\textsuperscript{262} in 1970 and the Family Educational Rights and Privacy Act (FERPA)\textsuperscript{263} in 1974.\textsuperscript{264} Other federal statutes addressing specific privacy issues
include the Children’s Online Privacy Protection Act (COPPA), the Health Information Portability and Accessibility Act (HIPAA), the Electronic Communications Privacy Act (ECPA), and the Gramm–Leach–Bliley Act (GLBA). Several federal statutes focus on the presence of personally identifiable information (PII), while others focus on transparency and access to information, on protecting consumers from inappropriate use of their personal data, or on imposing duties of confidentiality. Federal statutes often include requirements for administrative agencies to promulgate regulations. HIPAA requires HHS to enact regulations to support the Act. Under


269. See Schwartz & Solove, supra note 3, at 1827 (explaining the differing focuses of electronic privacy legislation). Statutes concerned with PII include the Cable Communications Policy Act, the Video Privacy Protection Act, and the Gramm–Leach–Bliley Act. See id. at 1824, 1829, 1830 (detailing the characteristics of CCPA, VPPA, and GLBA).

270. See Solove, Taxonomy, supra note 111, at 525 (listing the Privacy Act, CCPA, FCRA, and COPPA as examples).

271. See Richards, supra note 114, at 1167 (listing statutes that protect consumers’ PII from inappropriate uses); see also Kuner, Part 1, supra note 119, at 176 (“Data protection law gives rights to individuals in how data identifying them or pertaining to them are processed, and subjects such processing to a defined set of safeguards.”).

272. See Richards, supra note 114, at 1196 (listing statutes that impose confidentiality on handlers of PII).

273. See, e.g., Martin, supra note 44, at 297 (providing an example of a regulation promulgated under and supporting HIPAA (citation omitted)).
the GLBA, agencies regulating financial institutions are required to promulgate rules setting requirements for safeguarding customers’ personal information.  

States also adopt their own privacy laws to protect consumers. For example, there is a statute in Massachusetts that requires detailed data security procedures, and forty-five states have statutes requiring customer notification in the event of a security breach. Minnesota has a merchant liability statute, under which a merchant can be held liable if there was a security breach and customer credit card information was insufficiently protected. California’s Song–Beverly Act protects PII by prohibiting merchants from requiring customers to give personal information like their address and phone number “as a condition to accepting the [customer’s] credit card.”

There are several federal statutes aimed at protecting children as a vulnerable population, including COPPA, FERPA, and the Protection of Pupil Rights Amendment. COPPA

---

274. See Solove, Architecture, supra note 172, at 1274 (explaining the GLBA).

275. See Denny, supra note 46, at 240 (explaining a Massachusetts Internet privacy law); see also Rhodes & Kunis, supra note 23, at 50–51 (describing the Massachusetts law as being controversial because of the high bar that it sets as a minimum threshold for security).

276. See Rhodes & Kunis, supra note 23, at 49–50 (explaining that most states have followed California’s passage of laws requiring businesses to notify customers in the event of a security breach); Schwartz & Solove, supra note 3, at 1884–85 (noting that forty-four states have enacted laws requiring that businesses notify customers when they experience a security breach).

277. See Rhodes & Kunis, supra note 23, at 50 (noting the existence of a Minnesota law imposing liability for negligent handling of consumer financial data).


279. Schwartz & Solove, supra note 3, at 1831. The California Supreme Court also held that asking for a zip code would be sufficient to violate the Song–Beverly Act if the zip code was being requested as a condition of accepting a credit card. See Pineda v. Williams-Sonoma Stores, Inc., 246 P.3d 612 (Cal. 2011) (“In light of the [Song–Beverly Act]’s plain language, protective purpose, and legislative history, we conclude a ZIP code constitutes ‘personal identification information’ as that phrase is used in section 1747.08. Thus, requesting and recording a cardholder’s ZIP code, without more, violates the Credit Card Act.”); see also Schwartz & Solove, supra note 3, at 1834 (explaining the outcome of Williams-Sonoma Stores).

280. See Cascia, supra note 17, at 891 (identifying federal laws that protect
imposes limitations on the types of information that a website may collect from children younger than thirteen, and privacy policies must address the website’s information collection practices with regard to children. COPPA explicitly spells out what elements are required for notice to be valid. There are also state laws aimed at protecting children’s online privacy, like a law in Maine that requires parental consent before someone collects, transfers, or sells a minor’s personal or health-related information for product promotion purposes.

Some privacy-related laws do not focus on consumer protection, but on procedural elements of government investigations. The Privacy Act of 1974 regulates how federal agencies can collect and use personal records, the USA PATRIOT Act of 2001 grants a right to the U.S. government to demand data in the interest of protecting homeland security, and the ECPA sets out conditions under which the government can obtain a variety of electronic communications. Beyond statutes, the Fourth Amendment protects against unreasonable

---

281. See id. at 892 (discussing COPPA). COPPA is why the terms of service or privacy policies in our sample typically contained language about not collecting data from or marketing to children under thirteen. FTC, Frequently Asked Questions about the Children’s Online Privacy Protection Rule, http://www.ftc.gov/privacy/coppafaqs.shtm (last visited Feb. 3, 2013) (answering common questions from website providers about how to keep within COPPA regulations) (on file with the Washington and Lee Law Review).

282. See Schwartz, State, supra note 175, at 855 (describing ways that COPPA changed the previous practices of using data collected from children).

283. See Cascia, supra note 17, at 899 (discussing laws that states have enacted to support COPPA).


285. See SOLOVE, DIGITAL PERSON, supra note 2, at 68 (describing the Privacy Act of 1974). The Driver’s Privacy Protection Act of 1994 is similar and prohibits states from selling personal information from motor vehicle records to marketers. See id. at 69 (describing the Driver’s Privacy Protection Act of 1994).


287. See 20 U.S.C. § 1232g(j)(1) (2006); Lanois, supra note 18, at 45 (stating that the USA Patriot Act is a “hurdle to the international adoption of cloud computing” and explaining the Act’s expansion of federal power to collect data).

searches and seizures. However, Fourth Amendment protection is also likely to be weaker in the cloud than it would be if the same information were stored solely on a personal computer in the suspect’s home.

a. Electronic Communications Privacy Act

The ECPA was passed partly in response to the findings of the Office of Technology Assessment that the protections of e-mails were “weak, ambiguous, or nonexistent.” The ECPA and a major update to the Computer Fraud and Abuse Act (CFAA) were both passed in 1986, though in subsequent decades, the criminal provisions of the CFAA have been expanded much more than the electronic privacy protection provisions of the ECPA.

The ECPA consists of three federal statutes: the Stored Communications Act (SCA), the Pen Register statute, and the Wiretap Act. Its protections supplement those of the Fourth

289. See U.S. Const. amend. IV (“The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated . . . .”).

290. See Soghoian, supra note 5, at 386–87 (noting that law enforcement agencies have “essentially deputized” technology companies to monitor end-users’ use of their applications operating on the cloud).


292. See Reid Skibell, Cybercrimes & Misdemeanors: A Reevaluation of the Computer Fraud and Abuse Act, 18 BERKELEY TECH. L.J. 909, 912 (2003) (noting that the CFAA was amended eight times between 1986 and 2003); see also Martin, supra note 44, at 308 (listing provisions of the Act that criminalize Internet-based conduct); Robison, supra note 4, at 1196 (describing the ECPA and one component thereof called the Stored Communications Act (SCA)).


294. Id. §§ 3121–27.

295. See id. §§ 2510–22; see also Andrew William Bagley, Don’t Be Evil: The Fourth Amendment in the Age of Google, National Security, and Digital Papers and Effects, 21 ALB. L.J. SCI. & TECH. 153, 167 (2011) (identifying the three parts of the statute and highlighting court decisions that have interpreted the statute); Casey Perry, U.S. v. Warshak: Will Fourth Amendment Protection Be Delivered to Your Inbox?, 12 N.C.J.L. & TECH. 345, 349 (2011) (identifying the three parts of the statute and explaining the contents of the SCA). A “pen register” is a device that records phone numbers dialed, though the language of the statute also applies to other technological means. See “Pen Registers” and “Trap and Trace Devices,” ELECTRONIC FRONTIER FOUND. SURVEILLANCE SELF-DEFENSE
Amendment.296 The application of each depends on what type of information is sought and where it is in the transmission process.297 The Wiretap Act covers interception of wire, oral, and electronic communications.298 Under the Wiretap Act, obtaining e-mail contents in real time requires a Title III order to be issued with Department of Justice (DOJ) approval and a grant by a federal judge, and the order must be renewed every thirty days.299 Under the Pen Register statute, obtaining real time subscriber data requires an ex parte pen register order.300 Stored electronic information and the requirements for obtaining each type are addressed under the SCA,301 which we analyze in more detail in the section below.
(1) Stored Communications Act

The status of the SCA is problematic because much of the language is very unclear or outdated and interpretations of the statute by courts have varied significantly. The two most important sections for our purposes are: (1) § 2702, which addresses the circumstances under which a provider can voluntarily disclose customer information to others; and (2) § 2703, which addresses how the government can compel a provider to produce stored information. This sounds simple enough, but there are so many exceptions, subcategories, and additional requirements that the statute quickly becomes unwieldy. For example, one commonly referenced exception to the prohibition on disclosure allows for disclosure when the subscriber or customer (depending on the type of service) consents to disclosure, but the question then arises as to what actions can amount to consent. For example, is it consent under...
the SCA to accept the terms of a very broad privacy policy without reading these terms? Another exception in § 2702 that raises new questions in the cloud computing context is the exception for disclosure to persons who provide the service.306 The definition of remote computing service in the Wiretap Act supplements this exception, permitting service providers to monitor activities on their networks in real time.307 This exception allows private employers to internally share information about the online activities of employees when the employer provides these services in-house,308 but will employers lose the right to monitor their employees' online activities if they outsource IT to a cloud service provider? There is also some possible overlap between compelled and voluntary disclosures, such as when the government merely tells the provider about an ongoing investigation, and then the provider gives the government relevant information without a formal request being made for the information.309 In such circumstances, which set of exceptions or requirements should apply?

To determine the propriety of a disclosure under the SCA, the government must first determine whether the sought information is stored as part of an electronic communications

306. See 18 U.S.C. § 2702(b)(4) (providing the service provider exception).

307. See id. § 2711(2) (defining RCS and excluding from limitations on interception devices such as those “being used by a provider of wire or electronic communication service in the ordinary course of its business”); Kerr, supra note 297, at 1226–27 (explaining the advantages of a narrow definition of RCS for nonpublic service providers that want to monitor their networks); see also Soma, Gates, & Smith, supra note 16, at 516 (noting problems with the ECS/RCS dichotomy when applied to service providers that outsource their communications services).

308. See Soma, Gates, & Smith, supra note 16, at 516, 521 (“The key to legal [e-messaging] monitoring by closed community service providers, such as employers, is providing notice and [obtaining] consent.” (internal quotations and citation omitted)).

309. See Kerr, supra note 297, at 1224–25 (calling this overlap a “gray zone”). Kerr offers up these examples: If an ISP finds files and contacts law enforcement pursuant to one of the exceptions in § 2702, but the ISP requests a subpoena before turning over the files so that it has a paper trail, is that voluntary or compelled? On the other hand, if the police contact an ISP and ask the ISP if they would like to help in the investigation of a child molester, and the ISP says yes and turns over its files, is that voluntary or compelled? See id. (providing examples of the gray zone).
service (ECS) or a remote computing service (RCS). An ECS is defined as “any service which provides to users thereof the ability to send or receive wire or electronic communications,”310 while an RCS is defined as a “provision to the public of computer storage or processing services by means of an electronic communications system.”311 In the provisions prohibiting voluntary disclosure, ECS providers are prohibited from knowingly disclosing communication contents that the provider holds in “electronic storage,”312 and RCS providers are prohibited from knowingly disclosing communication contents that the provider maintains for the sole purpose of providing the subscriber or customer with “storage or computer processing services.”313 Some cases have thus turned on a party’s ability to establish the difference between when communications are in “electronic storage” and when communications are just in “storage.”314

The process required to obtain information also varies with the type of information sought, with notice required prior to the disclosure of some information types, some of which require a warrant, while others require a special court order under § 2703(d), and still others require only a subpoena.315 These three

---

311. Id. § 2711(2).
312. Id. § 2510(17) (defining “Electronic storage” as: “(A) any temporary, intermediate storage of a wire or electronic communication incidental to the electronic transmission thereof; and (B) any storage of such communication by an electronic communication service for purposes of backup protection of such communication”). Intermediate storage is thus covered by the ECS rules, while long-term storage is covered by the RCS rules. Kerr, supra note 297, at 1216 (noting that intermediate storage is covered under the ECS rules and long-term storage is covered by the RCS rules).
313. See 18 U.S.C. § 2702(a) (2006) (mandating restrictions on providers of RCS). Insofar as RCS providers are prohibited from disclosing contents held for “storage or computer processing” purposes, these protections go away if the provider is authorized to access the communication contents for any purpose other than “storage or computer processing.” See id. § 2702(a)(2)(B) (delineating the scope and context of the prohibition).
314. See, e.g., Theofel v. Farey-Jones, 359 F.3d 1066, 1075 (9th Cir. 2004) (reversing dismissal of plaintiff’s claim that an Internet service provider disclosed e-mails in violation of the SCA on the grounds that the e-mail messages were in “electronic storage” and therefore afforded SCA protections); see also Kerr, supra note 297, at 1229 (discussing the Theofel case).
315. See 18 U.S.C. § 2703 (requiring, for example, a warrant for disclosure of
methods of compelling information are listed in descending order of the strength of the showing required to obtain them. To obtain a § 2703(d) court order, the governmental entity must show “specific and articulable facts” establishing “reasonable grounds” to believe that the information sought is “relevant and material to an ongoing criminal investigation.” This standard is less than the “probable cause” standard for obtaining a warrant, but greater than the “reasonable relevance” standard for obtaining a subpoena. The type of information sought also determines whether a § 2703(d) order or subpoena must be accompanied by prior notice to the target. For example, the statute explicitly states that only a subpoena, and no prior notice to the customer, is required to compel an ECS or RCS provider to disclose noncontent, basic subscriber information, including the customer’s name, address, phone records (including session times and durations), length and type of service, phone number, and how the customer pays for the service. Most of this same noncontent subscriber information, it should be noted, can be freely disclosed.
to nongovernmental entities pursuant to an explicit exception in the voluntary disclosure provisions.\textsuperscript{320}

Considering the many avenues for uncertainty within the SCA, it should come as no surprise that disclosures under the SCA are often a source of contention. A disclosure in violation of the SCA may give rise to a civil cause of action.\textsuperscript{321} However, good faith reliance on a seemingly lawful document compelling disclosure acts as a complete defense to a civil action against a provider who is compelled to disclose communications.\textsuperscript{322}

\textit{(2) Applying the SCA to the Cloud}

Orin Kerr has written a very detailed and well-received article analyzing and explaining the SCA.\textsuperscript{323} The SCA is a complex statute that Congress wrote based on how early computer networks operated.\textsuperscript{324} The category of RCS provider was intended to address the business model in which companies

\textsuperscript{320} See 18 U.S.C. § 2702(c) (permitting disclosure to the National Center for Missing and Exploited Children). The SCA thus leaves a hole for the disclosure to private parties of personally identifiable information, so the privacy policies of these providers would thus be more applicable to the protection of PII than the SCA. Because there is no explicit exception for subpoenas in civil litigation, courts interpret this omission as meaning that private litigants cannot obtain data other than noncontent information from ECS and RCS providers. See Robison, \textit{supra} note 4, at 1208–09 (citing a number of cases that have so decided). However, because of the importance of civil discovery, courts may promote alternative methods of obtaining information held by an ECS or RCS provider, such as a Rule 34 motion to compel the party to produce data held by these providers. See, e.g., Flagg v. City of Detroit, 252 F.R.D. 346, 349–55 (E.D. Mich. 2008) (concluding that the SCA does not preclude civil discovery for electronic stored communications that are maintained by a nonparty service provider because the other party has control over that information and thus can be compelled to produce it under Rule 34).

\textsuperscript{321} See 18 U.S.C. § 2707(a) (creating a private cause of action for knowing or intentional violations of the law).

\textsuperscript{322} See \textit{id.} § 2707(e)(1) (extending the good faith exception to “a court warrant or order, a grand jury subpoena, a legislative authorization, or a statutory authorization”).

\textsuperscript{323} See Kerr, \textit{supra} note 297, at 1213–33 (analyzing and explaining the SCA).

\textsuperscript{324} See Robison, \textit{supra} note 4, at 1205 (explaining the legislative history of the SCA).
outsourced a lot of storage and processing functions due to the high cost of doing this in-house.\textsuperscript{325} It is thus likely that the RCS category would easily apply to commercial cloud services that provide options for outsourcing IT, but in most other contexts, there is substantial overlap between RCS and ECS. In part because of the rigid language of the statute and the substantial changes that have come about in the electronic communications field, it is unclear how the privacy protections of the SCA apply to other communications in the cloud.\textsuperscript{326} Kerr suggests eliminating the categories of ECS and RCS to address some of the confusion.\textsuperscript{327}

Currently, the degree of privacy in an e-mail likely depends on whether it is stored on a hard drive or in the cloud.\textsuperscript{328} E-mails downloaded from a service provider are easily covered by the requirement in the SCA that requires a warrant to obtain unopened e-mails fewer than 180 days old, but it is unclear whether a webmail provider would be considered an “electronic communication service” or a “remote computing service.”\textsuperscript{329} Some industry actors seem to oppose the 180-day rule because people now leave information on webmail services for long periods of time.\textsuperscript{330} Generally, many support the proposal to revise the SCA to better address cloud computing.\textsuperscript{331}

Some question whether the SCA would protect free cloud services at all because advertising-supported business models

\begin{itemize}
  \item \textsuperscript{325} See \textit{id.} at 1206–07 (explaining the legislative history of the SCA).
  \item \textsuperscript{326} See Kattan, \textit{supra} note 8, at 619 (recognizing the uncertainty of the SCA’s protections of cloud-based electronic communications).
  \item \textsuperscript{327} Kerr, \textit{supra} note 297, at 1209 (calling the categories “confusing” and suggesting their removal); see also Kattan, \textit{supra} note 8, at 653 (echoing this recommendation and suggesting that Congress consider whether it even makes sense to continue to distinguish between ECS and RCS).
  \item \textsuperscript{328} See Lanois, \textit{supra} note 18, at 45 (discussing cloud privacy issues).
  \item \textsuperscript{329} See Bagley, \textit{supra} note 295, at 167–68 (noting the SCA’s ambiguous relationship to webmail services).
  \item \textsuperscript{330} See Kattan, \textit{supra} note 8, at 642–43 (noting Microsoft’s objections to the 180-day rule).
  \item \textsuperscript{331} See \textit{id.} at 645 (noting that Digital Due Process, a consortium of privacy advocates, are lobbying for amendment of the ECPA); see also \textit{supra} Part II.E (discussing the need for legislative reform to create adequate privacy protections for the cloud).
\end{itemize}
often give the providers access to communication contents for targeted advertising purposes.\textsuperscript{332} This may prevent these services from being considered RCS providers because the provider is authorized to access communication contents for purposes other than rendering storage and computer processing services.\textsuperscript{333} TOS agreements and privacy policies thus have potentially significant effects on the extent to which the SCA protects the customer’s privacy because these terms may give the provider explicit authority to take actions that would disqualify the provider from being considered a provider of RCS.\textsuperscript{334}

3. Case Law

While there is not an explicit clause in the U.S. Constitution that states the existence of a general right to privacy, courts have held that such a right exists and is protected by the Constitution. Much discussion of Supreme Court privacy jurisprudence focuses on decisional privacy; that is, the right of individuals to make decisions free of government intervention.\textsuperscript{335} In \textit{Whalen v. Roe}, the Supreme Court recognized “the individual interest in avoiding disclosure of personal matters,”\textsuperscript{336} which has influenced many lower courts in recognizing a constitutional right to information privacy.\textsuperscript{337}

\textsuperscript{332} See Kattan, \textit{supra} note 8, at 638–40 (arguing that Gmail might not be considered an RCS provider because the privacy policy allows Google to access user communications for advertising purposes); Robison, \textit{supra} note 4, at 1196 (noting that this quid pro quo violates the SCA’s provisions).

\textsuperscript{333} See Robison, \textit{supra} note 4, at 1213 (noting that cloud services may not qualify as RCS because these services allow advertisers access to customer data).

\textsuperscript{334} See id. at 1215, 1220–21 (identifying three varieties of terms-of-service agreements and explaining how courts have interpreted terms-of-service agreements).

\textsuperscript{335} Some of the best known examples of decisional privacy cases in the Supreme Court over the last fifty years concern contraception and abortion. See, \textit{e.g.}, \textit{Roe v. Wade}, 410 U.S. 113, 153 (1973) (citing a right to privacy under the Constitution in prohibiting states from outright banning abortion); \textit{Griswold v. Connecticut}, 381 U.S. 479, 485 (1965) (ruling unconstitutional a Connecticut law that prohibited the use of contraception).


\textsuperscript{337} See Solove, \textit{Taxonomy, supra} note 111, at 558 (discussing the privacy
U.S. courts have often examined the distinctions between the public and private spheres. The public-disclosure-of-private-facts tort, for example, has been found to not apply to the republication of public postings on the web, even when the original posting is deleted a few days after it was first posted.\textsuperscript{338} In contrast to public postings, courts may be more protective of seemingly nonsensitive private data, like search queries.\textsuperscript{339}

Courts have also examined privacy in the context of the Fourth Amendment and statutes that relate to privacy, like the ECPA. Also relevant to our research, some courts have examined the implications of TOS agreements and privacy policies. It is to these topics we now turn.

\textbf{a. Fourth Amendment}

The Fourth Amendment declares that people have a right “to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures.”\textsuperscript{340} The Fourth Amendment also applies to seizure of digital evidence, though seizing digital evidence stored in the cloud is likely to be much easier than seizing identical data that is stored solely on a suspect’s personal computer within the suspect’s home.\textsuperscript{341}

\begin{itemize}
\item \textsuperscript{338} See Moreno v. Hanford Sentinel, Inc., 91 Cal. Rptr. 3d 858, 862–63 (Cal. Ct. App. 2009) (concluding that initial publication was not sufficiently obscure or transient). The Moreno court concluded this in spite of the fact that the public backlash did not rise to unacceptable levels until after the college student’s MySpace posting about the town was republished in the local newspaper. See id. at 861–62.
\item \textsuperscript{339} See, e.g., Gonzales v. Google, 234 F.R.D. 674, 687–88 (N.D. Cal. 2006) (declining to compel disclosure of 5,000 search queries and noting the potential privacy concerns of such disclosures).
\item \textsuperscript{340} U.S. Const. amend. IV. Typically, this protection requires a search warrant to be issued, though in some circumstances it is acceptable for the warrant to be executed by parties other than law enforcement. See United States v. Bach, 310 F.3d 1063, 1068 (8th Cir. 2002) (permitting an ISP’s technicians to execute a search warrant outside the presence of law enforcement).
\item \textsuperscript{341} See Soghoian, supra note 5, at 386–87 (noting that digital search and seizure is far easier because of the development of the cloud).
\end{itemize}
Fourth Amendment cases often focus on the need for a warrant, the issuance of which requires a finding of probable cause by a judicial officer. When searches are executed without a warrant, Fourth Amendment jurisprudence requires courts to decide if the target of the search had a subjective expectation of privacy that society recognized as reasonable. Much turns on the existence of this “reasonable expectation of privacy” (REOP). Courts recognize a REOP in papers and effects sent in the mail, and some courts have held that e-mail is analogous to postal mail and thus a sender has a REOP in e-mail. A REOP might not exist, for instance, in a library’s shared computers that are available for public use, but may exist in a personal Yahoo! webmail account. On the other hand, a public employee may have a reduced REOP in equipment provided by

342. See United States v. Leon, 468 U.S. 897, 923 (1984) (ruling that a warrant is deficient for failure to show probable cause if the facts articulated on the face of the warrant are “so lacking in indicia of probable cause as to render official belief in its existence entirely unreasonable”); supra note 340 and accompanying text (discussing the Fourth Amendment).


344. See United States v. Warshak, 631 F.3d 266, 274 (6th Cir. 2010) (extending Fourth Amendment protections to e-mail). But see Rehberg v. Paulk, 611 F.3d 828, 847 (11th Cir. 2010) (ruling in favor of immunity because there is not a “clearly established” constitutional right to privacy in e-mail content “voluntarily transmitted over the global Internet and stored at a third-party ISP”).

345. See Wilson v. Moreau, 442 F. Supp. 2d 81, 104, 108 (D.R.I. 2006) (citing cases that declared a lack of reasonable expectation of privacy when using public library computers); see also United States v. D’Andrea, 648 F.3d 1, 5–14 (1st Cir. 2011) (analyzing a government search of a password protected website used to store private images). On the other hand, courts might not find a REOP in information disclosed in a chatroom with other individuals. See United States v. Charbonneau, 979 F. Supp. 1177, 1185 (S.D. Ohio 1997).

346. See United States v. Lifshitz, 369 F.3d 173, 193 (2d Cir. 2004) (ruling that a probation condition allowing for frequent or random monitoring of probationer’s computer use may be overbroad and violate a Fourth Amendment interest); Trulock v. Freeh, 275 F.3d 391, 403 (4th Cir. 2001) (concluding that a third party’s authority to consent to search of shared spaces did not extend to the defendant’s password protected files).
his employer. On this point, the Supreme Court recently examined the issue of a public employer-issued pager and the extent to which the Fourth Amendment protected text messages sent over this pager, concluding that a search that is justified by noninvestigatory work-related purposes is reasonable. In City of Ontario v. Quon, the Court assumed, but did not conclusively determine, that there was otherwise a REOP in text messages.

If a warrantless search is conducted where a REOP exists, a court will examine if one of the exceptions to the warrant requirement applies, and if one does not, the evidence derived from the violation may be suppressed at trial. One of the exceptions is when the party with a REOP is the object of the search, or a third party with adequate authority, gives consent to the search. The scope of the permission is also important. Suppression may be an option when investigators seize more than permitted under the warrant, which is an especially big danger with e-discovery. However, suppression might not be an available remedy in the case of evidence that was within the

347. Compare United States v. Angevine, 281 F.3d 1130, 1134 (10th Cir. 2002) (noting that a public employer’s computer use policy may reduce an employee’s REOP in the employee’s office computer), with Maes v. Folberg, 504 F. Supp. 2d 339, 347 (N.D. Ill. 2007) (finding a REOP in a public employee’s work laptop computer).

348. See City of Ontario v. Quon, 130 S. Ct. 2619, 2632–33 (2010) (holding that the city’s review of a police officer’s text messages did not violate the officer’s Fourth Amendment rights).

349. See id. at 2619.

350. See id. at 2630 (“Even if Quon had a reasonable expectation of privacy in his text messages, petitioners did not necessarily violate the Fourth Amendment by obtaining and viewing the transcripts.”).

351. Compare United States v. Andrus, 483 F.3d 711, 721 (10th Cir. 2007) (holding that a cotenant had apparent authority to consent to a search of the computer in defendant’s room when police used forensic software to bypass possible password protection), with Trulock, 275 F.3d at 402–03 (holding that third-party consent does not extend to the defendant’s password protected files).

352. See United States v. Comprehensive Drug Testing, Inc., 621 F.3d 1162, 1177 (9th Cir. 2010) (emphasizing the importance of procedures to segregate data covered by the warrant from data that is not covered, including the need to have disinterested computer technicians go through the information to separate covered data).
scope of the warrant, even if the seizure of evidence was outside the warrant.\footnote{353}

There is also an important distinction between “content” and “noncontent” information, with the latter category including things like addresses on the outside of an envelope and phone numbers dialed on a phone. In the cloud context, a person is likely to not have a REOP in subscriber information provided to an ISP.\footnote{354} Other noncontent information, like e-mail addresses in the “To” field of an e-mail or IP addresses of visited websites, are likely to not be protected by the Fourth Amendment.\footnote{355}

The third-party doctrine of Fourth Amendment jurisprudence is the focus of much discussion in the online privacy context, because the doctrine prevents a REOP from being found in papers and effects turned over to a third party.\footnote{356}

\footnote{353}{See United States v. Hill, 459 F.3d 966, 973–76 (9th Cir. 2006) (declining to suppress evidence found on storage media seized pursuant to a warrant that only addressed the seizure of a computer). If digital evidence within the scope of the warrant has been deleted by the computer owner, it is not outside of the warrant’s scope for police to restore deleted files when executing the warrant. See United States v. Upham, 168 F.3d 532, 537 (1st Cir. 1999) (“The seizure of unlawful images is within the plain language of the warrant; their recovery, after attempted destruction, is no different than decoding a coded message lawfully seized or pasting together scraps of a torn-up ransom note.”).}

\footnote{354}{See United States v. Perrine, 518 F.3d 1196, 1204 (10th Cir. 2008) (finding that the defendant had no REOP for subscriber information given to ISP).}

\footnote{355}{See United States v. Forrester, 512 F.3d 500, 510 (9th Cir. 2007) (holding that it is not a Fourth Amendment search to use computer surveillance techniques to obtain only information concerning the to/from fields of e-mails and website addresses visited); Viacom Int’l, Inc. v. Youtube, Inc., 253 F.R.D. 256, 261–62 (S.D.N.Y. 2008) (stating that privacy concerns of disclosing noncontent information like data logs and IP addresses were only “speculative”). The Perrine court also suggested that the defendant’s use of peer-to-peer software also decreased the defendant’s expectation of privacy, especially as to the files shared over the P2P service. See Perrine, 518 F.3d at 1205 (ruling that Perrine’s use of peer-to-peer software, which allowed other Internet users to access files on his computer, partially reduced the expectation of privacy that he would otherwise enjoy in his computer, therefore leaving Perrine with no Fourth Amendment interest in subscriber information he gave to Yahoo! and Cox).}

Bailments of concealed items, however, might not be limited by the third-party doctrine because the bailee entrusted with concealed items does not necessarily have the authority to view the items or to consent to the search of the items by others. Thus, much will turn on the authority that the third party has with respect to the entrusted items. If a private carrier’s terms retain the right to inspect a package for any reason, acceptance of these terms by a customer of the private carrier may also result in a loss of a REOP in packages sent using this private carrier.

Information privacy scholars often argue that this third-party doctrine will prevent Fourth Amendment protections from applying in the cloud because users must inherently reveal their information to third parties in order for it to be transmitted or processed. On the other hand, recent case law casts doubts on this view. In the Sixth Circuit case *United States v. Warshak*, the court distinguished e-mail interception from other third-party doctrine cases by holding that the e-mail provider was an intermediary in the communication, not a recipient. However, the *Warshak* court also noted that if an agreement with a service provider gave the service provider the authority to “audit, inspect, and monitor” the e-mails of its subscribers, that might cause the subscriber to lose a REOP in those e-mails.

---

357. See *United States v. James*, 353 F.3d 606, 613–15 (8th Cir. 2003) (overturning a child pornography conviction because the police obtained the evidence from a person entrusted with computer disks with specific instructions to not use the disks).

358. See *United States v. Young*, 350 F.3d 1302, 1307–09 (11th Cir. 2003) (finding no REOP when a private carrier retained the right to inspect packages for any reason and when defendant also violated other terms set out by the carrier). The court in *Young* also held in the alternative that reserving a right to inspect gave the carrier the ability to later consent to a search of the package by law enforcement. See *id.* at 1308 (“Just as the 'right to inspect' notice defeated Young's privacy interest, we believe it also served to defeat Young's Fourth Amendment challenge because it authorized Federal Express, as a bailee of the packages, to consent to a search.”).

359. See, e.g., Bagley, *supra* note 295, at 173–74 (arguing that Fourth Amendment protections should be extended to data that is revealed involuntarily and incidentally to using a service).

360. See *United States v. Warshak*, 631 F.3d 266 (6th Cir. 2010).

361. See *id.* at 288 (distinguishing *Miller*).
For Fourth Amendment protection to apply to a search, it must be executed by a state actor because the Fourth Amendment only protects against intrusions by the state.\textsuperscript{362} Searches conducted by a private party thus do not inherently raise Fourth Amendment concerns unless the private party is behaving as a state actor,\textsuperscript{363} an analysis that often turns on state entanglement with the private party’s business.\textsuperscript{364} Under the Supreme Court’s holding in \textit{United States v. Jacobsen},\textsuperscript{365} if the papers or effects are secured such that it is clear that the nonstate-actor searcher does not have a right to look at the contents, but the searcher executes a private search anyway, a REOP might not protect the owner of the contents if law enforcement then duplicates the private search.\textsuperscript{366} The \textit{United States v. D’Andrea}\textsuperscript{367} case suggests a narrow interpretation of \textit{Jacobsen}, however, with the former court implying that a private search of password protected online storage would not validate a subsequent warrantless search unless the content owner was careless with his password security.\textsuperscript{368}

\textsuperscript{362} See U.S. Const. amend. 4 (providing protections against unreasonable and warrantless searches); United States v. Jacobsen, 466 U.S. 109, 113 (1984) (stating that the Fourth Amendment is construed as “proscribing only governmental action; it is wholly inapplicable ‘to a search or seizure, even an unreasonable one, effected by a private individual not acting as an agent of the Government or with the participation or knowledge of any governmental official’” (citation omitted)).

\textsuperscript{363} See United States v. Richardson, 607 F.3d 357, 365–66 (4th Cir. 2010) (holding that an ISP was not turned into a state actor by a statute that required ISPs to report to law enforcement if the ISP found child pornography on its network).

\textsuperscript{364} See Burton v. Wilmington Parking Auth., 365 U.S. 715, 716–17 (1961) (finding state action sufficient to implicate the Equal Protection Clause of the Fourteenth Amendment when a restaurant in a city-operated parking garage denied service to a customer because of his race).

\textsuperscript{365} See Jacobsen, 466 U.S. at 109.

\textsuperscript{366} See id. at 126 (holding that if a third party violates the person’s expectation of privacy, the government may use that information to the same extent, but cannot exceed the scope of the private search). But see United States v. D’Andrea, 648 F.3d 1, 8 (1st Cir. 2011) (ruling that the private search doctrine would not apply when the property owner did not give the private searcher permission or means to search the property, unless the property owner was very careless about security of the property).

\textsuperscript{367} See D’Andrea, 648 F.3d at 1.

\textsuperscript{368} See id. at 8 (noting that the private search doctrine might apply if the
b. Stored Communications Act

As discussed above, the ECPA includes the Wiretap Act, the SCA, and the Pen Register statute. 369 The SCA is a complicated statute that courts often interpret differently. The Ninth Circuit, for example, adopted a broad interpretation of the phrase “for backup purposes” within the definition of “electronic storage.” The Theofel v. Farey Jones370 court held that a provider was holding e-mails in “electronic storage,” and thus the ECS terms applied rather than the RCS terms that would apply if the e-mails were merely in “storage.” 371 Kerr criticizes the Ninth Circuit’s interpretation of “electronic storage” under the SCA, 372 specifically its broad interpretation of “backup purposes” that may permit more frequent findings that a service is acting as an ECS provider. 373

It is unclear how the SCA will apply to webmail. The Theofel court itself suggested that storage would not be for “backup...
purposes" if the information were not stored anywhere else. In United States v. Weaver, the court held that Hotmail was an RCS provider, so only a trial subpoena would be required to compel Hotmail to produce previously accessed e-mails under 181 days old. It is also unclear the extent to which the SCA would prevent a party in civil litigation from seeking records maintained by a cloud service provider during discovery. Even if a provider does violate the SCA, the remedy for such a violation is not suppression, as it would be with a Fourth Amendment violation because the available remedies are explicitly limited under § 2708.

The intersection of the SCA and the Fourth Amendment may also raise issues of the ultimate constitutionality of the SCA. Because the e-mail acquisition actions of law enforcement in Warshak relied on the SCA, and the Warshak court viewed e-mail as having the same Fourth Amendment protection as postal mail, the Warshak court ultimately concluded that insofar as the SCA permitted law enforcement to obtain e-mails without a warrant,

---

374. See supra note 371 and accompanying text (discussing the Theofel court’s interpretation of the “backup purposes” language contained in the SCA).


376. See id. at 771 (interpreting 18 U.S.C. § 2703(b)(2)’s subpoena requirements and ruling that e-mails under 181 days old must be seized using a subpoena).

377. On its face, the SCA does not include an exception for disclosure for civil discovery purposes, so adverse parties may be precluded from requesting stored information directly from service providers. However, because of the importance of civil discovery, courts are likely to find other avenues for allowing such information to be compelled, such as by compelling the opposing party to produce it directly under Rule 34 of the Federal Rules of Civil Procedure. See Flagg v. City of Detroit, 252 F.R.D. 346, 352–53 (E.D. Mich. 2008) (concluding that the SCA does not preclude civil discovery for electronically stored communications that are maintained by a nonparty service provider because the other party has control over that information and can be compelled to produce it under Rule 34). Thus, because the subscriber requests the information from the nonparty service provider, the information is being disclosed consistent with the statutory exceptions.

the SCA was unconstitutional. Accordingly, if more courts adopt the reasoning of Warshak, the influence of the SCA in the e-mail search context may be significantly reduced.

c. Contracts and Privacy

In the cloud context, one should also consider case law precedent concerning contract law and agreements between consumers and service providers. As the Warshak court noted, excessively permissive TOS agreements may deprive a customer of a REOP in contents stored or transmitted using a service, so the validity of these contracts has implications for privacy law. Additionally, privacy policies and TOS agreements have implications for SCA cases because one of the most important exceptions under the SCA is for information obtained after the subscriber or customer has given valid consent. Thus, a privacy policy that reserves a license to use the customer’s information and content for business or marketing purposes may be read as consent to disclosure under the SCA. To the extent that the consent applies to communications stored for computer storage and processing purposes, consent renders the SCA completely inapplicable if the service otherwise only qualified as an RCS instead of an ECS.

Contract law is typically state law, so standards will vary across cases. Generally, contracts can be invalidated if they are found to be unconscionable. Some courts have invalidated excessively one-sided TOS agreements on unconscionability

---

379. See United States v. Warshak, 631 F.3d 266, 288 (6th Cir. 2010) (ruling that the portion of the SCA that allows the government to obtain e-mails without a warrant is unconstitutional).

380. See id. at 287 (noting that “if the ISP expresses an intention to ‘audit, inspect, and monitor’ its subscriber’s emails, that might be enough to render an expectation of privacy unreasonable” (citation omitted)).

381. See 18 U.S.C. §§ 2702(b)(3), 2703(c)(1)(C), 2703(c)(2) (explaining the consent exceptions to the SCA).

382. However, this may be limited to content for which the customer does not elect stronger privacy protections. Viacom Int’l, Inc. v. Youtube, Inc., 253 F.R.D. 256, 264–65 (S.D.N.Y. 2008) (concluding that acceptance of Youtube’s TOS and privacy policy does not amount to consent to the disclosure of private content).
grounds, often when the customer was seeking to avoid mandatory arbitration provisions.\textsuperscript{383} Courts have also examined terms not related to arbitration, and may also invalidate terms that the court finds to be excessively unfair.\textsuperscript{384}

At least one court has held that privacy policies are purely aspirational, and thus not contracts that are enforceable at law, when they do not afford any rights or remedies to the customer.\textsuperscript{385} However, this position does not take into account recent actions by the FTC to enforce privacy policies against companies on the grounds that violating its own privacy policy amounts to an unfair business practice.\textsuperscript{386} Thus, we do not anticipate that this “purely aspirational” characterization of privacy policies will be adopted.

\textsuperscript{383} See Specht v. Netscape Commc’ns Corp., 306 F.3d 17, 35 (2d Cir. 2002) (“Reasonably conspicuous notice of the existence of contract terms and unambiguous manifestation of assent to those terms by consumers are essential if electronic bargaining is to have integrity and credibility.”); Bragg v. Linden Research, Inc., 487 F. Supp. 2d 593, 611 (E.D. Pa. 2007) (finding procedural and substantive unconscionability sufficient to invalidate an arbitration provision that was part of a TOS that amounted to a very one-sided adhesion contract); People v. Network Assocs., Inc., 758 N.Y.S.2d 466, 470 (N.Y. Sup. Ct. 2003) (holding that some terms of a software license were unenforceable due to their unfairness when the provisions included terms prohibiting customers from publishing reviews of the product or benchmark test results without the company’s permission).

\textsuperscript{384} See Network Associates, 758 N.Y.S.2d at 470 (holding that some terms of a software license were unenforceable due to their unfairness when the provisions included terms prohibiting customers from publishing reviews of the product or benchmark test results without the company’s permission). Courts value the rights of parties in contracts, however, so perhaps they are likely to only invalidate unfair terms that pass a certain threshold. See MDY Indus., LLC v. Blizzard Entm’t, Inc., 629 F.3d 928, 955–56 (9th Cir. 2010) (acknowledging the contract claim that arises when a user violates a game’s prohibition on the use of “bots” to automatically play a game); Bowers v. Baystate Techs., Inc., 320 F.3d 1317 (Fed. Cir. 2003) (enforcing a shrinkwrap agreement that prohibited reverse engineering).

\textsuperscript{385} See Freedman v. Am. Online, 325 F. Supp. 2d 638, 640 (E.D. Va. 2004) (finding that a subscriber agreement “is plainly aspirational only . . . and is not intended to confer any rights and remedies upon the subscriber”).

\textsuperscript{386} See infra note 465 and accompanying text (discussing FTC enforcement actions).
4. European Privacy Law

The United States and the European Union take very different theoretical approaches to privacy. In the United States, the idea of privacy is often related to concepts like secrecy and intrusions. In the United States, privacy is viewed as an aspect of liberty, with the goal of protecting against intrusions by the state. The European approach, however, views privacy as a right of human dignity, with a focus on an individual’s personal autonomy in deciding how his personal data will be used by anyone, including the free market.

European privacy law is very complicated, and a detailed examination is outside of the scope of this Article. However, because we now live in a global information economy, cloud providers must inevitably consider how their services and practices will need to be altered for a European market. For this reason, we will give a fairly brief introduction to this complicated topic.

In the European Union, privacy is considered to be a fundamental right. The first European data protection laws were enacted in the 1970s, followed by the adoption of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data in 1981, and the enactment of the EU Data Protection Directive 95/46 (DPD 95/46) in 1995.

387. See supra Part III.A (discussing fundamental privacy theories).
388. See Birnhack & Elkin-Koren, supra note 108, at 341 (“In the American model, privacy is understood as a liberty, protecting citizens against the State.”).
389. See id. (“[T]he common European understanding is of a right to human dignity—an individual right to determine the end uses of our personal data—in which threats to privacy arise from both the State and the free market.”).
390. Lanois, supra note 18, at 37 (stating that “the European Union has enshrined the status of privacy as a fundamental right”).
393. See Kuner, Part 1, supra note 119, at 176–77 (providing a discussion of the early data protection laws in Europe).
DPD 95/46 focuses on the protection of “personal data,” which it defines as “information relating to an identified or identifiable natural person.” By its terms, it applies EU law to data controllers that use “equipment” within the EU for the processing of personal data, but the term “equipment” has been read broadly to apply to things like cookies and JavaScript. Other provisions that U.S. cloud providers must comply with include DPD 95/46’s requirements for robust authentication and access safeguards.

a. The Safe Harbor Framework

DPD 95/46 also governs the transfer of data, permitting data transfers only to other countries with adequately protective privacy laws. The United States does not have sufficient privacy laws, but the data of European users can nonetheless be transferred to the United States if the company handling the transfer complies with the Safe Harbor agreement between the

394. See Council Directive 95/46, supra note 392, at 38; see also Schwartz & Solove, supra note 3, at 1873–74 (comparing the reduction in privacy law in the U.S. to its expansion in the EU).

395. See Kuner, Part 2, supra note 126, at 228–29 (explaining the controversy over the use of the term “equipment” to encompass things such as “cookies”). Other EU regulations also restrict the use of cookies and similar technologies, requiring the informed consent of a user to be obtained prior to the provider commencing to store and access information that is on the user’s computer. See Lanois, supra note 18, at 40 (discussing the consent requirement under the EU Data Protection Directive and both the 2002 and 2009 ePrivacy Directives). Because cookies involve information that is considered personal data under the Data Protection Directive, which includes IP addresses, a company that uses cookies in the EU must comply with the terms of both the Data Protection Directive and the more recent ePrivacy directive of the EU. See id. at 41 (“[T]he use of cookies or similar devices involving a unique user ID or an identifier will result in the application of both the Data Protection Directive and the ePrivacy Directives.”).

396. See Lanois, supra note 18, at 47 (“[D]ata may only be transferred outside of the EU if that country provides an ‘adequate’ level of protection . . . .”).

397. See Stylianou, supra note 44, at 596 (noting that “[t]he gravest expression of the implications of different levels of privacy protection occurred when the European Union . . . passed the Data Protection Directive, which allows the transfer of data intended to undergo processing to third countries only if they ensure an adequate level of protection”).
United States and the European Union. Because of the limits of DPD 95/46 on transferring personal data, some cloud providers have also established segregated EU clouds.

The Safe Harbor framework provides a method for companies to certify compliance with European privacy standards without necessarily using segregated clouds. Under the Safe Harbor Privacy Principles, organizations must: (1) provide notice about data collection; (2) give individuals a choice to opt out of the disclosure of their personal information to third parties or before their information is put to a secondary use (or to opt in to sharing if the personal information is considered sensitive); (3) extend these standards to onward transfers—that is, ensure that third parties to whom personal information is transferred also adhere to the Safe Harbor Privacy Principles or have comparable controls in place; (4) provide individuals with access to their personal information held by the organization; (5) take reasonable security precautions to protect personal information; (6) take reasonable steps to protect data integrity; and (7) provide adequate measures for enforcement of the principles.

Adhering to the Safe Harbor Privacy Principles provides a mechanism for companies in the United States to preserve the

398. See Lanois, supra note 18, at 48 (discussing how the Safe Harbor program works between the United States and the European Union); Stylianou, supra note 44, at 597 (explaining the Safe Harbor agreement, “according to which American companies could transfer data from Europe as long as they abided by a commonly agreed upon privacy framework set by the United States Department of Commerce and the European Commission”). The European Commission has found that membership in the U.S. Safe Harbor system provides an “adequate level of data protection,” but the Commission suggests that further transfer of the data beyond the Safe Harbor member must comply with EU privacy law as well. See Kuner, Part 2, supra note 126, at 231 (discussing personal data protection concerns regarding “onward transfers of data” from U.S. Safe Harbor members to third parties).

399. See Lanois, supra note 18, at 48 (“The most simple and obvious way to comply with the EU Data Protection Directive is to ensure that personal data does not leave the EU . . . which is why certain cloud vendors offer segregated EU clouds that keep personal data from being transferred outside of the European Union.”).

status quo of a self-regulatory approach to privacy while still being eligible to serve customers in the European Union.\footnote{401}{See James T. Sunosky, Privacy Online: A Primer on the European Union’s Directive and United States’ Safe Harbor Privacy Principles, 9 CURRENTS: INT’L TRADE L.J. 80, 85 (2000) (explaining the benefits of the Safe Harbor Privacy Principles).} We note that these principles also resemble many of the ideas underlying FIPs, and that the Safe Harbor Privacy Principles might provide another model for how companies should handle personal information belonging to customers located in the United States.

\textit{IV. Companies, Customer Data, and Customer-Company Interactions}

\textbf{A. Companies and Customer Data}

In this Article, we examine the interaction between privacy theories, privacy law, and the relationships between consumers and the companies that serve them in the cloud. These relationships are largely defined by TOS agreements and privacy policies, and these agreements typically enumerate what a consumer can expect concerning the use of his personal information. The concerns about how companies handle customer data go beyond these agreements, however, and include issues like data security, identity theft, and behavioral marketing.

\textit{1. Terms of Service Agreements}

TOS agreements set forth terms governing the relationship between a service provider and its customers.\footnote{402}{See Joshua A.T. Fairfield, Contemporary Issues in Cyberlaw: Nexus Crystals: Crystallizing Limits on Contractual Control of Virtual Worlds, 38 WM. MITCHELL L. REV. 43, 44 (2011) (referring to terms of use and EULAs as the “social contract of the new millennium,” setting forth the rights and redresses of citizens).} Generally, cloud-based services targeted at individual users are accompanied by non-negotiable TOS agreements that favor the service provider.
over the end user.\(^{403}\) TOS agreements will generally address things like metering, monitoring, and data backup,\(^{404}\) and often include clauses in which the provider disclaims liability for harm and forbids customers from using the company’s intellectual property without authorization.\(^{405}\) Some also include terms concerning the retention, control, and ownership of a user’s information.\(^{406}\) TOS agreements take a variety of approaches to customer information. Some include terms that allow providers to access customer information for advertising and other purposes relating to the business, while others are less transparent about what the company may do with customer information, and still others make explicit promises in their TOS agreements that the companies will not access customers’ data.\(^{407}\)

The terms of TOS agreements can have a significant impact outside the context of the provider–customer relationship, potentially affecting the consumer’s legal rights. The DOJ has recently argued that violating a website’s TOS agreement amounts to unauthorized access under the CFAA,\(^{408}\) and courts

---

403. See Bagley, supra note 295, at 163 (“Google’s profit model is based on offering free services to consumers in exchange for their consent to non-negotiable terms of service.”); Wittow & Buller, supra note 7, at 7 (“The SLAs of cloud-based applications and services generally are non-negotiable and much more favorable to the provider than to the end user.”).

404. See Martin, supra note 44, at 311 (noting the difficulties of providing good customer service because of the lack of standards to measure a cloud’s performance).

405. See Bagley, supra note 295, at 178 (“[L]anguage in a TOS agreement merely disclaims liability for any damage to a user’s computer data and forbids unauthorized use or redistribution of intellectual property.”).

406. See id. (explaining that TOS clauses “also dictate the terms by which the entity will retain, control, and own a user’s information”). Google’s TOS agreement includes a provision giving the company a license to use the customer’s data in ways that would otherwise violate the customer’s copyright. See Google Policies and Principles, Terms of Service, GOOGLE (Mar. 1, 2012), http://www.google.com/policies/terms/ (last visited Feb. 3, 2013) (providing that Google may use personal data in accordance with their privacy policies) (on file with the Washington and Lee Law Review).

407. See Robison, supra note 4, at 1215–17 (providing an examination of existing cloud providers).

have also examined whether agreeing to an expansive TOS agreement or a broad privacy policy may cause a person to lose a reasonable expectation of privacy. As discussed above in Part III.B.2.a, the terms of TOS agreements may also impact the application of the SCA.

It is very important that consumers read and understand the terms of cloud services' TOS agreements because of the large amounts of sometimes sensitive information stored with these services. Consumers should pay special attention to how the TOS agreements address customer data, including the information that the company claims rights in, and how the consumer can terminate his relationship with the cloud provider. Consumers might be storing information solely in the cloud, making it very important for the TOS agreements to


409. See United States v. Warshak, 631 F.3d 266, 287 (6th Cir. 2010) (stating that such protections likely would not apply to content stored with a provider that includes terms in an agreement reserving the right to "audit, inspect, and monitor" e-mail content); Bagley, supra note 295, at 181 (discussing cases examining the Fourth Amendment in the context of Terms of Service agreements).

410. See supra Part III.B.2.a (discussing the Electronic Communications Privacy Act).

411. See Soma, Gates, & Smith, supra note 16, at 534 (examining the "blurred lines" between work and home life that e-technology has created and suggesting that "users must make a good faith effort to read, understand, and ask questions about service provider privacy and terms of use policies"); see also Stylianou, supra note 44, at 593 (noting that such terms attract greater scrutiny because of the large amount of data stored with these providers).

412. See Wittow & Buller, supra note 7, at 7 (asserting that cloud service TOS agreements should address data migration issues to assure business continuity and to protect the customer's continued access to data after the customer's relationship with the provider is dissolved).
include provisions protecting customers’ ability to retrieve their content if, for example, a service is shut down.413

a. TOS Agreements as Contracts of Adhesion

Under the common law of contracts, forming a contract requires mutual assent.414 When a contract is not subject to negotiation and is offered by the more powerful party on a “take it or leave it” basis, the contract is often referred to as a contract of adhesion.415 Privacy policies and TOS agreements typically meet this definition for an adhesion contract.416 Such contracts are not automatically invalid, but they may be subject to greater scrutiny.

Excessively oppressive TOS terms may be invalidated if the court concludes that the terms are unconscionable.417 Unconscionability analysis often has two prongs, and courts evaluate the circumstances for both procedural and substantive unconscionability.418 Courts might be more willing to find

---


416. Solove, *Architecture, supra* note 172, at 1235 (arguing that the idea that users give informed consent to these terms is a fiction, due to the total lack of negotiation).


418. Id.
unconscionability when there are no market alternatives, but the diverse reality of the cloud market makes it unlikely that a lack of market alternatives will be a persuasive argument.\textsuperscript{419}

The unequal bargaining power between the provider and its customers means that providers often subject customers to terms that are more favorable to the provider.\textsuperscript{420} At least one court has looked favorably on a provider prohibiting the use of “bots” with its service,\textsuperscript{421} and Martin expresses concern that this opens the door for “predatory software vendor[s]” to prohibit customers from using third party software with the vendor’s projects, thereby eliminating beneficial effects of innovation by third parties.\textsuperscript{422}

2. Privacy Policies

Privacy policies and TOS agreements often overlap, though for our purposes we consider privacy policies to be more focused on making the customer aware of the company’s policies regarding their data instead of the customer’s obligations concerning the service. Terms in a provider’s privacy policy might address things like the quantity and nature of collected data and the company’s policies on data retention and customer control over data.\textsuperscript{423} Privacy policies often also address data security issues, like the use of SSL encryption during data transmission.\textsuperscript{424} However, many of these providers insert

\textsuperscript{419} See Bagley, supra note 295, at 179 (discussing the difficulty of demonstrating unconscionability “in the search engine, e-mail, and digital media services market, where there are many companies even though only a few giants dominate”).

\textsuperscript{420} See supra note 403 and accompanying text.

\textsuperscript{421} See MDY Indus., LLC v. Blizzard Entm’t, Inc., 629 F.3d 928, 950 (9th Cir. 2010) (finding that the prohibition of the use of “bots” was permitted under the Digital Millennium Copyright Act).

\textsuperscript{422} Martin, supra note 44, at 312.

\textsuperscript{423} See Stylianou, supra note 44, at 599, 602 (discussing the quantity and nature of collected data as well as data retention policies and data storage location).

\textsuperscript{424} See id. at 603 (providing a discussion of data safety, security, and integrity).
provisions in their privacy policies or TOS agreements that repudiate any liability for data loss, and reserve to the provider the right to discontinue the service at the provider’s sole discretion.425

Privacy policies typically consist of information provided by the service provider about how the provider may gather, use, disclose, and manage the personal information of its customers.426 Privacy policies, like TOS agreements, are often adhesion contracts marked by significant advantages being reserved for the service provider, such as the right to amend its privacy policy unilaterally with little notice to its customers.427 Privacy policies may include broad permissions to allow the provider to access information for its own marketing purposes and to disclose customer information to its business partners for business-related purposes.428 Privacy policies also might not be considered contracts at all, but purely as notices about a company’s policy. However, few consumers actually read a company’s privacy policy, and even fewer understand it.429 Solove criticizes many

425. See id. at 604 (examining Amazon’s, Mozy’s, and Apple’s data protection disclosures).

426. See Cascia, supra note 17, at 888 (“A privacy policy is a legal agreement between the user and the provider that discloses some or all of the ways the provider gathers, uses, discloses and manages a customer’s personal information.”).

427. See SOLOVE, DIGITAL PERSON, supra note 2, at 82–83 (arguing that privacy policies are not a meaningful contract, with no bargaining over terms and containing mostly unreliable, vague promises); Cascia, supra note 17, at 889–90 (discussing Google’s privacy policy and noting that “Google reserves the right to unilaterally amend its privacy policy leaving it essentially meaningless”). Birnhack and Elkin-Koren argue that if such a term is included, user privacy is not being effectively guaranteed by upfront notice and consent. See Birnhack & Elkin-Koren, supra note 108, at 365 (arguing that “if the user agrees upfront to any use of data as detailed by an adjustable privacy policy, the user does not exercise real control over the collection and use of personal data”).

428. See Soma, Gates, & Smith, supra note 16, at 532 (noting that providers often include terms in e-messaging policies and usage agreements permitting the provider to access the systems for “routine monitoring purposes” and to comply with lawful requests by the government or litigants); infra Part V (providing an empirical analysis of agreements and policies in the cloud).

429. See Schwartz & Solove, supra note 3, at 1856 (“[S]tudies have shown that few consumers read privacy policies, and that those who do frequently fail to understand them.”).
privacy policies as being “written in obtuse prose,” containing large amounts of extraneous information.430

Cloud services collect a lot of data, both through the customer’s voluntary disclosure of data and through the provider’s automatic collection of information through its operations or advertising policy.431 Many privacy policies assure limited use of customer information.432 Some, however, are vague, leaving ambiguities and loopholes. Transparency in privacy policies is very important, and consumers should be informed about how their data will be collected and used.433 In this Article, we posit that reserving explicit rights for consumers to control their data will raise consumer awareness of privacy issues. We anticipate that this raised awareness, combined with the increased control that an individual has over the use of his data, will have a positive effect on the market for cloud services.

a. Sharing Information with the Government

Consumers will often encounter inherent limitations in how much control they can exercise over their data because of common policies permitting the sharing of data with government entities. Privacy policies typically contain provisions reserving to the provider the right to disclose customer information pursuant to lawful government requests.434 Companies like Google and AT&T

430. SOLove, DIGITAL PERSON, supra note 2, at 82.

431. See Stylianou, supra note 44, at 599 (examining the quantity and nature of collected data by cloud services). Some companies may also collect information from other sources that pertains to the user indirectly. See id. at 601 (using Microsoft as an example to demonstrate that the practice of indirect data collection is increasing).

432. See id. at 601, 604 (discussing Microsoft’s, IBM’s, and Amazon’s privacy and data protection policies).

433. See Birnhack & Elkin-Koren, supra note 108, at 353 (explaining the importance of user consent to data collection).

434. See Google Privacy Policy, GOOGLE (July 27, 2012), http://www.google.com/policies/privacy/ (noting that Google may share user data for legal reasons) (last visited Feb. 3, 2013) (on file with the Washington and Lee Law Review); see also Soghoian, supra note 5, at 393–94 (citing a public statement by the CEO of Google in which the CEO listed assisting with lawful investigations as being one of the main reasons that Google keeps detailed data of the online
collect large amounts of personal user data from customers.\footnote[435]{See Bagley, supra note 295, at 155–56 ("[T]hird parties such as information service provider, Google, and telecommunication giant, AT&T, amass large amounts of personal user data.").} This sort of information was formerly used for marketing and research purposes, but recently the U.S. government has been building national security databases that contain personal user data provided by cooperating telecommunications companies like AT&T.\footnote[436]{See id. at 156 (noting that “in recent years the United States government has built national security databases with personal user data allegedly obtained from cooperating telecommunication companies” that has resulted in Fourth Amendment litigation); Soghoian, supra note 5, at 385–86 (discussing wiretaps obtained through telecommunication companies and Internet providers working with law enforcement officers). Bagley cites the wiretapping controversy as an example that did not involve warrants or subpoenas, but instead relied on voluntary agreements with private companies. \textit{See Bagley, supra} note 295, at 156–57 (criticizing that the “traditional legal process was evaded” in this situation because “private companies did the data gathering and managed the phone calls” and the companies involved waived their Fourth Amendment rights).} Sometimes, providers may voluntarily provide data to government entities to improve the provider’s own security.\footnote[437]{See Bagley, supra note 295, at 154 (citing the example of Google voluntarily providing data to the NSA).}

Governments have requested personal user information from various companies for a variety of purposes over the years.\footnote[438]{See id. at 161–62 (noting that the government sought user information from airlines after the September 11th attacks and from hotels and car rental agencies in 2003 to thwart terrorist threats against Las Vegas).} This is not limited to the United States. For example, the government of the United Kingdom is considering using data obtained by social networking sites for the purpose of monitoring users to prevent terrorism and crime.\footnote[439]{See id. at 164 (discussing the United Kingdom’s potential plan to use data collected by social networking sites).} Generally, private companies that turn over information to the government are not considered state actors by doing so.\footnote[440]{See id. at 162 ("[P]rivate companies are not restrained as state actors when they voluntarily hand consumer data to the government ... they are treated as a third party in whom a consumer is placing their trust."). \textit{But see id.} at 188 (arguing that there may be entwinement sufficient to find state action if a communication provider assigns employees to work with government agencies activity of its customers).} Cloud providers sometimes also are required...
to comply with certain content laws of other countries, like Skype’s Chinese counterpart that was required to implement a filter to prohibit text messages that included phrases like “Falungong” and “Dalai Lama.”441 There are also some concerns about the U.S. government’s ability to exploit software vulnerabilities or even enable the microphones of cellular phones remotely as part of criminal investigations.442

There are a number of other reasons why government officials might request information. The federal government recently used data associated with customer shopping cards to trace the source of salmonella poisoning.443 The DOJ has also requested search records from companies like Google and Microsoft in the course of its investigation into the effectiveness of child protection legislation.444 However, the court in that case did not compel Google to turn over actual search queries, noting in dicta that there may be an expectation of privacy in such queries.445

In addition to requesting the cooperation of private companies, the government itself has been collecting personal information for many years. Solove notes in his book that there are almost 2,000 databases of personal information maintained and respond to government requests).

441. Soghoian, supra note 5, at 408. Skype denied allegations that its Chinese software contained a backdoor to allow surveillance by the Chinese government, but it came out in 2008 that when text messages using this software were filtered, the offending message and the identities of the sender and recipient were forwarded to a publicly accessible server in China. See id. at 408–09 (providing a discussion of the TOM-Software).

442. See id. at 400–02 (discussing the FBI’s use of “roving bug” software).

443. See Martin, supra note 44, at 299 (examining use of consumer data by the federal government).

444. See Gonzales v. Google, 234 F.R.D. 674, 679 (N.D. Cal. 2006) (examining a subpoena by the U.S. Attorney General to Google to compile and produce information from the search engine’s index and search queries); Bagley, supra note 295, at 165 (discussing litigation involving subpoenas for online user data).

445. See Gonzales, 234 F.R.D. at 684 (denying the motion to order Google to disclose search queries of its users); Bagley, supra note 295, at 165 (noting that, “[i]n the end, Google was compelled only to generate a list of URLs, rather than actual user search queries”).
by the federal government. Personal information collection as part of the census began in 1790, with the questions becoming more personal until the 1890 census, which included questions about things like diseases, disabilities, and finances. The massive databases that are already maintained by the government and over which citizens have no control might appear to threaten any attempts to improve informational privacy. Requiring data control protections in the private sector may seem like a relatively small issue compared to government databases. However, private data held by governments generally do not leave the government’s possession, and thus the circulation of this information is not as problematic as the circulation of information collected in the private sector.

3. Effects of Security Breaches

A major reason that we argue for consumers to be in control of their data is that we think consumers should be empowered to take proactive steps to protect their information. Consumers, in our view, should be free to withdraw their data from a service if they learn of security failings in that service. One of the dangers of insufficient data security for data in the cloud is the risk of identity theft as a result of data breaches. According to the Identity Theft Resource Center, in 2009 there were at least 498 publicly reported data breaches, impacting 222 million total

446. SOLOVE, DIGITAL PERSON, supra note 2, at 15. Richards also notes that the government has huge databases of information about citizens. See Richards, supra note 114, at 1156 (discussing the history of personal data collection by the federal government that began as early as the nineteenth century).

447. See SOLOVE, DIGITAL PERSON, supra note 2, at 13 (providing a historical look at the collection of public data by the federal government). The public outcry in response to the intrusiveness of the questions in the 1890 census eventually led to legislation to ensure the confidentiality of census data. See id. (“When the 1890 census included questions about diseases, disabilities, and finances, it sparked a public outcry, ultimately leading to the passage in the early twentieth century of stricter laws protecting the confidentiality of census data.”).

448. See Lanois, supra note 18, at 44 (discussing the increasing amount of “commercial, personal, and even secret data and other sensitive information . . . flowing around the globe in the cloud”).
records.\footnote{Wittow & Buller, supra note 7, at 9.} A single data breach of a credit card processing company in 2012 may have resulted in 1.5 million credit card accounts being compromised.\footnote{Credit Card Data Breach Contained, Says Global Payments, BBC NEWS (Apr. 3, 2012, 5:59 ET), http://www.bbc.co.uk/news/technology-17596394 (last visited Feb. 3, 2013) (on file with the Washington and Lee Law Review).}

Identity theft is a federal crime and has been referred to as the most rapidly growing white collar crime,\footnote{See Solove, DIGITAL PERSON, supra note 2, at 110 (stating that the FTC estimated that 10 million Americans were victims of identity theft in 2003).} though some criticize the law as not being adequately supported by resources or sufficient criminal sentences.\footnote{See Solove, Architecture, supra note 172, at 1248 (noting the problems with viewing identity theft as an exclusively criminal matter).} Approximately half a million people are victims of identity theft every year.\footnote{Id. at 1244.} Twenty-six percent of consumer complaints submitted to the FTC in 2008 concerned identity theft.\footnote{Wittow & Buller, supra note 7, at 9.}

But identity theft is not the only risk related to data breaches.\footnote{See Solove, Architecture, supra note 172, at 1258 (“With ever more frequency, we are hearing stories about security glitches and other instances of personal data being leaked and abused.”).} Some breaches can involve very personal and embarrassing information, such as when a firm accidentally posted to the Internet the names, addresses, phone numbers, credit card information, and details of the sex lives of ninety psychotherapy patients.\footnote{See Solove, DIGITAL PERSON, supra note 2, at 53 (discussing “irresponsible and careless uses of personal information”).} Sometimes, breaches are due to a serious failing in a company’s procedures. In one instance, Metromail Corporation hired prison inmates to enter personal information into Metromail’s databases, and one inmate started sending sexually explicit letters with information about the recipients’ lives.\footnote{Id. (providing examples of instances of security breaches of personal information in recent years).} In another more troubling instance, the
company Docusearch provided a man with information about a woman named Amy Lynn Boyer, which the man then used in finding and murdering Boyer.\textsuperscript{458}

Cloud providers might not bear the risk of loss due to fraud, but companies have many incentives to secure data and prevent security breaches because large-scale breaches often result in negative publicity. Security breaches can destroy consumer confidence and devastate a company’s bottom line.\textsuperscript{459} However, this decrease in consumer confidence may not effectively incentivize the creation of stronger security protocols if cloud service providers store data in proprietary formats, making it difficult for current customers to leave. Thus, we argue that data control and format transparency could have benefits for security in the cloud by giving providers incentives to keep data secure in order to retain customers.

4. Protecting Consumer Data—Who Watches the Watchers?

Currently, consumers have fairly little control over their data, but there are other entities to help address data security issues. Several private bodies have set standards enabling companies to either seek certification as to the adequacy of their privacy practices, or otherwise measure their own actions against industry standards. These options include SAS 70 certification, which involves audits of firms’ control mechanisms to protect information;\textsuperscript{460} the Payment Card Industry Data Security

\textsuperscript{458} See Remsburg v. Docusearch, Inc., 816 A.2d 1001, 1009 (N.H. 2003) (finding that Docusearch owed a duty of reasonable care when the company disclosed Boyer’s information to Liam Youens); SOLOVE, DIGITAL PERSON, supra note 2, at 54 (providing the facts of the Docusearch case); Richards & Solove, supra note 1, at 1923 (discussing the holding in Docusearch and noting duty of care issues arising from computer databases).

\textsuperscript{459} See Rhodes & Kunis, supra note 23, at 26 (“A security breach affecting a corporation can destroy consumer confidence and be devastating to the bottom line.”). There was recently a breach at Heartland Security, leading to a loss of 130 million credit card numbers. \textit{Id.} at 45. Heartland has suffered major financial damages since the breach, including a $60 million settlement with Visa over the breach. \textit{Id.}

\textsuperscript{460} See SAS 70 Overview, SAS 70, http://sas70.com/sas70_overview.html
Standard, which created IT guidelines for the credit card industry aimed at reducing the risk of a security breach; and the Financial Industry Regulatory Authority (FINRA), which requires members to have policies and procedures addressing customer record safety, protecting against unauthorized access, and protecting against relevant anticipated threats. Companies on the Web may also seek TRUSTe certification for their privacy practices. These organizations are elements of the self-regulatory framework that U.S. businesses currently use with regard to privacy. However, these certification authorities are largely sector-specific, and thus we recommend broader protections that do not rely on sector-specific self-regulatory bodies.

When the sector-specific self-regulatory framework fails, there are sometimes other private solutions available. Customers may, for example, sue companies in the event of a database security breach, though courts disagree about whether a customer has standing based on a mere risk of future identity

461. See PCI SSC Data Security Standards Overview, PCI SEC. STANDARDS COUNCIL, https://www.pcisecuritystandards.org/security_standards/ (last visited Feb. 3, 2013) (providing an overview of the “comprehensive standards and supporting materials to enhance payment card data security”) (on file with the Washington and Lee Law Review); see also Martin, supra note 44, at 297 (“[P]ublic companies that fail to obtain SAS 70 qualification by adhering to certain procedures and controls can easily lose the confidence of investors and customers.”).


theft or if standing requires actual identity theft to have occurred.464 There are also other organizations that focus on online consumer protection issues, including the Electronic Privacy Information Center (EPIC) and Digital Due Process (DDP). These organizations are more policy-oriented and may do things like filing privacy-oriented amicus briefs in relevant litigation.

In terms of government intervention, the FTC has also become involved with personal data security and other privacy issues, using its authority to challenge unfair or deceptive practices.465 The first FTC action that primarily concerned a company’s data security practices was in 2004 against BJ’s Wholesale Club after hundreds of instances of identity theft arose due to BJ’s data security failings.466 An advantage to FTC involvement over private litigation by consumers is the ability of the FTC to bring an action against a company in the absence of identity theft. For example, the FTC fined Choicepoint in 2006 after a breach resulted in 163,000 private financial records being compromised, citing Choicepoint’s privacy policy as containing “inaccurate and misleading assertions about its security procedures.”467 The FTC may also bring an action when a company fails to adequately secure its data, even if there has not


465. See Rhodes & Kunis, supra note 23, at 36 (discussing the FTC’s jurisdiction and enforcement authority); Wittow & Buller, supra note 7, at 9 (noting that, as of the time of the authors’ writing, the FTC had filed twenty-seven enforcement actions concerning the data security practices of companies). The FTC requires companies to institute “reasonable safeguards” to protect information, and what is “reasonable” depends on factors like how sensitive the data is and how costly it would be for the company to avoid potential risks. See Rhodes & Kunis, supra note 23, at 36 (discussing the “reasonableness” standard applied by the FTC beginning in 2006 to bolster the enforcement of data security risks).

466. See Rhodes & Kunis, supra note 23, at 37 (noting the FTC’s conclusion that the security failings amounted to an unfair practice in violation of federal law).

467. Id.
actually been a data breach. The FTC could also potentially bring an action against a business that uses deceptive practices to obtain information.

As an alternative to extending current regulations to new data issues in the cloud, some argue that the FTC and its current authorities could be used to enforce a company’s privacy policy against it. However, after examining a number of privacy policies, we argue that this approach would not be wise given the reality that many companies adopt vague privacy policy language regarding the company’s own obligations. It is also unclear whether the FTC would be the appropriate regulatory body in all instances because the FTC usually regulates e-commerce issues, but providers whose services count as telecommunications or information services would also be governed by FCC regulations. We also assert that relying on government

468. See Schwartz & Solove, supra note 3, at 1856–57 (“[T]he [FTC] has taken actions against companies that fail to provide adequate data security . . . even in the absence of a data breach, though more typically it acts only once a data spill has occurred.”). The FTC also settled an enforcement action against Sears in 2009, based on Sears’s practice of tracking customers without adequately disclosing details of the tracking program to the customers, and another action against EchoMetrix in 2010 concerning parental control software that also provided information to marketers about children’s computer activity. See id. at 1858 (discussing the “more substantive approach to disclosure of company behaviors” taken by the FTC in enforcement actions).

469. See Richards, supra note 114, at 1185 (“The use of fraud or other deceptive practices in obtaining consumer data could also constitute a violation of the Uniform Deceptive Trade Practices Act (UDTPA), and would fall within the powers of the Federal Trade Commission (FTC) to deter and punish unfair trade practices . . . .”).

470. See SOLOVE, DIGITAL PERSON, supra note 2, at 72 (noting that the FTC has recently brought actions for “unfair or deceptive acts or practices” against companies that violate their own privacy policies); McCarthy, supra note 95, at 2260 (discussing the possibility of enforcing PHR vendors’ privacy policies against them). McCarthy argues, however, that HIPAA would be a stronger way to address privacy issues with personal health records. See id. at 2261 (contrasting HIPAA and the FTC by stating that “HIPAA mandates that covered entities take constant concern over privacy and security by continually auditing, monitoring, and augmenting security when necessary”).

471. See infra Part V.C (providing an analysis of and statistical information on privacy policies).

472. See Soma, Gates, & Smith, supra note 16, at 490–91 (suggesting the possibility of a joint rulemaking between the FTC and FCC to address these issues).
agencies to address the failure of companies to give consumers meaningful control over their data would be ineffectual because the most likely approach would be through adjudication, in which individual consumers would not be clearly represented, in an adjudicatory process that by definition would only address problems on an ad hoc basis. On this point, we argue that regulating this behavior in advance would be more beneficial to consumers than case-by-case adjudication.

5. Tracking Technologies and Behavioral Marketing

Another element of privacy policies that is relevant to the issue of data control is the use of technologies to track consumer behavior. Privacy policies typically address the tracking technologies that a website uses for advertising or other purposes. When tracking users, advertisers may use technologies like cookies, flash cookies, and Web beacons. The degree to which companies disclose the use of these tracking technologies varies. The information collected using these tracking technologies can then be used by companies to profile consumers. Consumers typically have the option to decline some tracking technologies, often by adjusting the settings of their Web browsers to decline all cookies. However, we suggest that this option does not represent a meaningful exercise of

473 See Birnhack & Elkin-Koren, supra note 108, at 372 (providing data comparing actual privacy practices to declared privacy practices of numerous websites); Lanois, supra note 18, at 34 (referencing a study that found that the top fifty websites installed, on average, sixty-four pieces of tracking technology when a visitor loaded the site, and usually did not provide a warning that they were doing so).

474 See Richards, supra note 114, at 1157 (discussing the “profiling industry” and noting that the profiles may include “a person’s social security number, shopping preferences, health information . . . financial information, race, weight, clothing size, arrest record, lifestyle preferences, hobbies, religion, reading preferences, homeownership, charitable contributions, mail order purchases and type, and pet ownership”); see also SOLOVE, DIGITAL PERSON, supra note 2, at 50 (noting private companies’ recent use of information to categorize people as either angel customers or demon customers, and the practices of some banks to deny credit card applications from college students majoring in liberal arts).
control because many websites require cookies to be enabled for website functionality.

A cookie is a text file that is downloaded to a user’s computer when she accesses a website, and it acts as an identifier for the computer on which it is stored.475 Cookies by themselves do not contain a user’s personal information under most definitions of the term,476 but a company called DoubleClick provides a service to websites, connecting cookies to personal information to enable more targeted advertising.477 Flash cookies have a similar effect to text cookies, but some flash cookies may be able to reconstruct previously deleted browser cookies and cannot be controlled by the user.478 Recent research revealed that out of the one hundred most popular websites, fifty-four used flash cookies, but only four sites mentioned the use of flash cookies in their privacy policies.479 Web beacons, the third type of tracking technology noted above, permit the advertiser to observe a user’s website activity in real time.480

Behavioral marketing is advertising that is targeted at individuals based on their past behavior patterns.481 The environment of behavioral marketing has developed substantially

475. See SOLOVE, DIGITAL PERSON, supra note 2, at 24 (referring to cookies as “a form of high-tech cattle-branding”); Lanois, supra note 18, at 33 (explaining how cookies work and why they are useful for both advertising and consumers).

476. However, because cookies typically collect a user’s IP address, this is sufficient to find that cookies collect “personal data” for purposes of the EU’s Data Protection Directive. See Lanois, supra note 18, at 41 (“In practice, almost all cookies involve the processing of personal data because even if the user’s real identity remains anonymous, cookies typically involve the collection of the user’s IP address, the processing of unique identifiers, or both which are personal data within the scope of the Data Protection Directive.”).

477. See SOLOVE, DIGITAL PERSON, supra note 2, at 24–25 (explaining how DoubleClick functions).

478. See Lanois, supra note 18, at 35 (discussing a lawsuit that involved the distinction between flash cookies and traditional cookies).

479. Id. at 36.

480. See Schwartz & Solove, supra note 3, at 1851 (“Some technology, particularly the beacon, or ‘Web bug,’ permits real-time observation of a user’s activity on an Internet page, including where one’s mouse moved and the information that one typed, such as search queries or personal information that an individual filled into a form.”).

481. See id. at 1849 (introducing the concept of behavioral marketing).
over the last century, becoming more effective as marketers have gained access to more detailed information.\textsuperscript{482} Behavioral marketing has led to advertisers buying access to individuals who match a particular consumer profile.\textsuperscript{483} There is a market for consumer data that is collected and can be used for targeting advertisements, with information about an individual’s browsing habits selling for a fraction of a cent on the data exchange.\textsuperscript{484}

Because declining all cookies would likely lessen a user’s Web browsing experience, researchers have worked to develop a technology that focuses on collection by third parties, like third-party advertisers that collect data for behavioral advertising. Concerns over such data collection and the possible privacy implications thereof have led to calls for a “Do Not Track” (DNT) standard, similar to a “Do Not Call” registry, that would allow users to opt out of tracking by third parties.\textsuperscript{485} Mozilla’s Firefox already includes DNT capabilities.\textsuperscript{486} Additionally, Microsoft made DNT the default setting for Internet Explorer 10, and Google announced that Google Chrome would have DNT capabilities by the end of 2012.\textsuperscript{487}

\textsuperscript{482} See Solove, Digital Person, supra note 2, at 19 (noting that direct mail has a yield-per-cost ratio double that of television advertisements).

\textsuperscript{483} See Lanois, supra note 18, at 34 (noting that user profiles are bought and sold on exchanges that resemble the stock market); Schwartz & Solove, supra note 3, at 1851 (“Marketers draw on extensive databases . . . . They are able to cross-reference online activity with offline records including home ownership, family income, marital status, zip code, and a host of other information, such as one’s recent purchases as well as favorite restaurants, movies, and TV shows.”).

\textsuperscript{484} See Richards, supra note 114, at 1157–58 (noting that in some places, consumer profiles can be bought for $65 for a thousand names); Schwartz & Solove, supra note 3, at 1852 (explaining that browsing information sells for as little as a tenth of a cent but that it adds up to a billion-dollar industry).


Modern consumers are often uneasy about the pervasiveness of behavioral advertising,\textsuperscript{488} and some research questions the ultimate value of targeted advertising.\textsuperscript{489} However, there is currently not much recourse available to consumers whose data is mined. The privacy torts typically require an invasion to be of an offensive nature, but most of the time, information collection is of largely innocuous information.\textsuperscript{490} For these reasons, one of our proposals relevant to data control focuses on the possibility of withdrawing data that was mined using these technologies. A DNT system, as described above, may also assist with limiting future unauthorized collection, provided that most websites eventually adopt it. At the time of this writing, however, many websites and advertisers have not adopted a DNT-friendly implementation.\textsuperscript{491} Even if the market solutions become more viable, our proposed data withdrawal and data portability rights are designed to inform and empower consumers, enabling more meaningful participation in the vigorous market for cloud services. In our view, such rights would be complementary to, and not supplanted by, an effective opt-out DNT regime.

\textsuperscript{488} See Schwartz \& Solove, supra note 3, at 1854 (suggesting that consumer objections to behavioral advertising should be addressed through policy).

\textsuperscript{489} Aleecia M. McDonald \& Lori Faith Cranor, The Cost of Reading Privacy Policies, 4 J.L. \& POL’Y INFO. SOC’Y 540, 541 (2008) (concluding that targeted advertising “may have negative social utility” after taking into account the opportunity costs required if everyone read and understood privacy policies).

\textsuperscript{490} See Richards \& Solove, supra note 1, at 1919 (citing Shibley v. Time, Inc. for its holding that disclosure of subscriber information did not meet the requirements of causing “mental suffering, shame or humiliation to a person of ordinary sensibilities.” (citing Shibley v. Time, Inc., 341 N.E.2d 337 (Ct. App. Ohio 1975))).

\textsuperscript{491} See Solomon, supra note 487 (“The main problem with DNT, though, is that not all that many websites and advertisers actually abide by it, since it’s more of a guideline than an actual rule.”). In fact, some critics say that DNT simply does not work and, in addition, that advertisers are adopting an interpretation of DNT that is contrary to the intent of those promoting DNT. See Ed Bott, Why Do Not Track is Worse Than a Miserable Failure, ZDNET (Sept. 21, 2012, 12:35 GMT), http://www.zdnet.com/why-do-not-track-is-worse-than-a-miserable-failure-7000004634/ (last visited Feb. 3, 2013) (criticizing DNT and arguing that it does not work) (on file with the Washington and Lee Law Review).
6. Personally Identifiable Information and “Anonymous” Information

The final concept that we will address in the context of privacy policies is the treatment of certain types of information. Arguments about the degree of protection to which information is entitled often turn on the type of information being protected. In the context of information privacy, the focus is often on personally identifiable information (PII), and on information that is considered sensitive. Computer use in the 1960s led to PII becoming more of an issue because companies and government entities were processing a lot of personal data.492 PII is a term that is often used to describe information that is clearly connected to a specific person, though there is no uniform definition of the term.493 Service providers often focus on assuring customers that their PII will be kept safe.

Regulatory intervention is often focused on protecting PII, in part because of the threats posed by identity thieves. Statutes define PII in several different ways. Some define PII as information that is personally identifiable, some define PII as information that is not public, and some define it by providing specific examples of information that is PII.494 With PII, the question is often whether information is identified or identifiable, which respectively refers to whether information immediately connects to an identified person or can be used to lead to an identified person, given more information.495 In the United States, the concept of PII is largely limited to identified data, whereas the European Union takes an expansionist view of PII that treats identified data the same as data that is only identifiable.496 Schwartz and Solove argue that the European

492. See Schwartz & Solove, supra note 3, at 1820 (explaining why the PII became an issue in the 1960s).
493. See id. at 1816 (“Given PII’s importance, it is surprising that information privacy law in the United States lacks a uniform definition of the term.”).
494. See id. at 1828 (identifying the competing definitions of PII).
495. See id. at 1817 (setting forth a “PII 2.0” model that proposes “two categories of PII, ‘identified’ and ‘identifiable’ data,” and treats them differently).
496. See id. (comparing the United States and European models); see also id.
Union’s expansionist approach is more consistent with the technology than a reductionist approach that limits PII protections to identified personal data.497

The idea of categorizing information as PII, however, has become more problematic over the years. The line between identified and identifiable has become increasingly blurred, as has the line between sensitive and nonsensitive. A social security number is generally viewed as very sensitive information, but date of birth may be considered less so. However, computer science has shown that a person’s social security number can be estimated to some degree of accuracy if one knows the person’s date of birth and the city in which they were born.498 If a database contains a very large amount of nonsensitive information, the aggregation of the information can track a person’s whole existence.499 A person’s search queries are an example of seemingly anonymous information that could nonetheless lead to an identifiable person, especially considering common behaviors like searching for local businesses, information on particular medical diagnoses, and vanity searches when an individual will often search for her own name to see what results emerge.500

at 1875 (noting that Canada takes a similar approach to that of the European Union).

497. See id. at 1875 (“The European Union’s expansionist approach to PII is more in tune with technology than is the United States’ reductionist approach.”).

498. See id. at 1846 (citing a recent study by Alessandro Acquisti and Ralph Gross).

499. See Bagley, supra note 295, at 164 (“The synthesis of data from a user’s web search history coupled with email, photos, documents, voicemails, phone logs, and location, creates a profile of an individual that serves as behavior modeling for advertisers. This same data could just as easily be disclosed to law enforcement officials for criminal profiling.”); see also Richards, supra note 114, at 1158 (acknowledging the privacy concern that “uber-databases can be created, composed of nonsensitive information in such enormous quantities that the database constitutes a highly detailed dossier of a person’s entire existence”).

500. See Paul Ohm, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 57 UCLA L. REV. 1701, 1717–18 (2010) (providing an example of AOL search queries being used to identify individuals); Schwartz & Solove, supra note 3, at 1848 (explaining that “if the user has engaged in a highly specific search, or multiple searches, she becomes more
Even if identifying information is removed, that does not necessarily solve the privacy problems. Reidentification science is a new field in computer science research that reattaches anonymized information to identified individuals. Researchers may reidentify a dataset by, for example, comparing two databases, one anonymized and one containing PII and some information fields in common with the anonymized database. The FTC has recently acknowledged that the distinction between PII and de-identified information is often blurred. Because of the ease with which data can be reidentified, Ohm suggests rejecting the concept of PII entirely, though Schwartz and Solove instead suggest a reunderstanding of what information should be considered PII.

501. See Ohm, supra note 500, at 1704 (arguing that the science of reidentification should be of more concern to policymakers than PII).

502. See id. at 1725–26 (explaining the basic principles of how reidentification functions). One potential source of databases containing PII is public records, which may many times be obtained upon a showing that the request is not for an improper purpose. See SOLOVE, DIGITAL PERSON, supra note 2, at 133 (discussing the history of public record disclosure laws). Public records that may be so obtained include a person's vital records and records of a person's interactions with government. See id. at 128–29, 134 (demonstrating the breadth of personal information contained in public records). When a person makes a state or federal Freedom of Information Act (FOIA) request, the state may redact private personal information, but the person to whom the information corresponds cannot object to a disclosure if the state does not redact such information. See id. at 135 ("The federal FOIA doesn't require that a person be given notice that his or her personal information is encompassed within a FOIA request. Even if an individual finds out about the request, she has no right under FOIA to prevent or second-guess an agency's decision . . . .").

503. See Schwartz & Solove, supra note 3, at 1828 (discussing the core conceptual problems with PII recently identified by the FTC). The Seventh Circuit has also noted the problem of reidentification, holding that redacting patient identities in a series of records about recipients of partial birth abortions was not sufficient to avoid violating the patients' privacy rights. See id. at 1844 (recognizing that "de-identified data can readily be re-identified" (citing Nw. Mem'l Hosp. v. Ashcroft, 362 F.3d 923, 929 (7th Cir. 2004))).

504. Compare Ohm, supra note 500, at 1742–45 (arguing that the concept of PII must be replaced to allow for privacy law to move forward), with Schwartz & Solove, supra note 3, at 1817 (arguing that "PII must be re-conceptualized if privacy law is to remain effective in the future").
Considering the technological issues, theorists who urge government regulation should evaluate which approach to PII should be taken. Reidentification science should be examined by policymakers to determine whether the concept of PII should be expanded to include both identified and identifiable information. Schwartz and Solove propose a model in which information is considered identified when the person’s identity is ascertained, identifiable when there is a nonremote possibility of future identification, and nonidentifiable when the risk of identification is remote and the information is not relatable to a person.\footnote{505} We argue that limited government intervention would be beneficial to set a baseline for protection of PII to mitigate threats to identity security, but we do not take a position on the identified–identifiable dichotomy. The existence of such regulations would likely raise consumer awareness of these threats. Once informed, we expect that consumers will express a preference for exercising meaningful control over their PII, whether identified or identifiable.

V. Empirical Analysis of Agreements and Policies in the Cloud

In the interest of empirically establishing a baseline for the current status of “data control” terms in contemporary agreements, we examined the privacy policies and TOS agreements of several different cloud providers. Our sample size is fairly small, consisting of twelve TOS agreements and nineteen privacy policies, but because so many companies use boilerplate language for these agreements, even though they may include different types of provisions, our small sample size is nonetheless very likely to be fairly representative of the industry. In fact, insofar as our sample emphasizes several enterprise-oriented companies with fee-based structures, in addition to consumer-oriented companies that rely on advertising revenue, our findings may provide a more generous estimate of the degree to which the terms of agreements favor the customer.

\footnote{505} See Schwartz & Solove, supra note 3, at 1878 (discussing their proposed model of PII).
A. Methodology

In collecting the privacy policies and TOS agreements, we first used a report by BTC Logic that identified thirteen companies that are viewed as leaders in cloud computing. In addition to the BTC Logic report, we also collected TOS agreements and privacy policies for six additional consumer-oriented cloud services whose information was available through Quantcast’s website. Most of these companies made their privacy policies available on a company website. In addition to privacy policies for all nineteen companies, we collected twelve separately labeled TOS agreements and one set of disclaimers from EMC that did not include other terms commonly found in TOS agreements.

506. See BTC Logic Ranks: Top 10 Cloud Companies, BTC LOGIC (2010), http://www.btclogic.com/documents/BTCLogic_TopTen_Q22010.pdf (providing a ranking and report of the top ten cloud computing companies). Based on their appearances in the list of the top thirteen cloud companies, we included in our sample: Google, Amazon, Microsoft, Cisco, Citrix, EMC, Level 3, Oracle, Red Hat, Sales Force, Symantec, VMWare, and IBM.

507. The additional six companies whose agreements we analyzed were: Carbonite, Dropbox, Flickr, Facebook, GoDaddy, and Apple. For Apple, we specifically looked at the TOS agreement and privacy policy for Apple's new iCloud service. Quantcast is a company that is very active in the Web advertising arena, with a website that provides detailed information about services on the web. Quantcast also provides a list of the top 100 websites in terms of visits. Top Sites, QUANTCAST, http://www.quantcast.com/top-sites (last visited Feb. 3, 2013) (providing a ranked list of websites based on the number of people in the United States who visit each month) (on file with the Washington and Lee Law Review).

508. Some companies declined to disclose sample TOS agreements. We attribute this in part to the different business models of the companies. If a product is anticipated to be widely deployed to a large number of people, such as Amazon’s AWS or Google’s ad-supported services, the TOS will most likely be standardized to address the company's relationships with a large population. When a service provider is contracting with an established enterprise that is paying a large sum for these services, we anticipate that the contracting is likely to be more balanced, with TOS agreements being tailored to the specific customer. Some companies such as IBM focus on very customer-specific services with a target audience of large enterprises, and the terms will change based on the specific needs of the customer. In those situations, the TOS will be more like a standard contract between two parties than the click-wrap agreements that individual consumers are familiar with through installing software on their systems.
Table 1

<table>
<thead>
<tr>
<th>Company</th>
<th>Privacy Policy</th>
<th>Disclaimers/Warranties Separate from TOS</th>
<th>TOS Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Microsoft</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cisco</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Citrix</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMC</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Oracle</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Red Hat</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salesforce</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Symantec</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>VMware</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carbonite</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Dropbox</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Flickr</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Facebook</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>iCloud</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>GoDaddy</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The remaining six companies, all from the BTC Logic sample, did not make a boilerplate TOS agreement available to noncustomers, and these companies often targeted their services at enterprise customers. In these situations, TOS agreements may be closer to a traditional contract. However, consumers with fewer resources, like end users and small businesses, are likely to have no bargaining power. These consumers are the anticipated beneficiaries of the changes we suggest. Currently, small businesses and end users are simply not given reasonable alternatives for controlling their data. Thus, the current market
is inefficient, notwithstanding the availability of more involved contracting when the customer is a wealthy enterprise.

One of the challenges of educating the public about privacy policies and TOS agreements is that many times, people assume that these agreements are all the same. We posit that a detailed, side-by-side comparison of agreement provisions based on provision categories would prove the most helpful in noting both similarities and differences between these provisions. Thus, once we had collected the privacy policies and TOS agreements, we first carefully read the language of the agreements with an eye to creating categories and subcategories to permit us to compare the language of several agreements from a top-down perspective. We then categorized the different provisions and noted in our research whether certain provisions were present or absent in a given company’s available policies.

The purposes of privacy policies and TOS agreements are very different. Privacy policies generally focus on information that can identify the individual, whereas TOS agreements generally focus on matters relevant to potential conflicts between a company and its customers. Both types of agreement reflect a company’s policies with regard to data control. Privacy policies are more directly relevant to data control issues relating to the control of personal information. TOS agreements, on the other hand, often address matters like the ownership of intellectual property and the processes to be followed to obtain stored data upon termination of service.

Analyzing a collection of nineteen privacy policies and twelve TOS agreements applicable to cloud services, a number of patterns emerged. The following subparts will first give a brief overview of our findings with respect to TOS agreements and privacy policies, and then discuss the implications of these patterns.

B. Terms of Service Agreements

TOS agreements typically address potential legal conflicts in advance. These agreements may use choice of law and venue provisions to state where litigation must take place, indemnify the company against third parties, limit damages either by type
or by setting a cap for damage awards, and disclaim warranties to the extent permitted by law. The TOS agreements in our sample addressed major topics like these. Seven of the twelve noted that the company might alter the services offered, but only two of the twelve stated that they would give notice to customers of such changes. This is significant because that means that most of the companies in our sample provide little information upfront about the degree to which customers will be notified of service changes.

For our purposes, one relevant aspect of this relationship is the issue of exercising control over data upon termination. Within our sample, eleven of the twelve TOS agreements set out conditions for account termination. Ten of the twelve set out conditions in which the company is authorized to terminate an account for cause, and two of the twelve set out conditions in which either party to the contract can terminate an account for cause, including a breach of the agreement between the parties. Six of the twelve also include provisions allowing customers to terminate their accounts for any reason.

Only five of the twelve, however, address the issue of data access after an account is terminated, and how and when a customer may access the provider’s servers to back up their files and delete them from the servers. The removal of information from the company’s servers also implicates document retention policy, which is sometimes addressed in a company’s privacy policy. Eleven of our full sample of nineteen companies address document retention to some extent, and refer to the company’s possible limitations under the law concerning permanent deletion of a customer’s data.


510. States following the Uniform Commercial Code will often not recognize a contractual waiver of certain implied warranties unless the waiver is conspicuous within the written contract. See U.C.C. § 2-316 (1977).
Table 2

<table>
<thead>
<tr>
<th>TERMINATION OF ACCOUNT (out of 12)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lists when company may terminate for cause</td>
<td>10</td>
</tr>
<tr>
<td>Lists when either may terminate for cause</td>
<td>2</td>
</tr>
<tr>
<td>Permits customer to terminate without cause</td>
<td>6</td>
</tr>
<tr>
<td>Includes provisions for temporary account suspension</td>
<td>1</td>
</tr>
<tr>
<td>Specifies time period for former customer to access</td>
<td></td>
</tr>
<tr>
<td>and delete information*</td>
<td>4</td>
</tr>
<tr>
<td>TOS does not address</td>
<td>2</td>
</tr>
</tbody>
</table>

* However, 11 of the 19 in our full sample also refer to the possibility that they may be required by law to retain customer information.

In TOS agreements, we were especially concerned about the extent to which the information remains the property of the customer. All twelve companies in our TOS sample included statements asserting the company’s rights in its own intellectual property that it was licensing to its customers, though only six of the twelve included a provision reiterating that the customer’s intellectual property remains his own. Three of the twelve (Amazon, Flickr, and Apple) reserve a license in the customer’s IP to the company limited to the purpose for which the customer submitted the content. Another set of three companies in the sample (Facebook, GoDaddy, and Google) also address the granting of a license to the customer’s IP, but these latter three do not include explicit restrictive language that would limit the scope of the license to the customer’s initial purposes. Eleven of the thirteen companies for which we had TOS agreements or sets of disclaimers also include provisions whereby the company retains full rights in suggestions or ideas submitted by customers, and can therefore use or implement these suggestions as they see fit without giving the submitter of the idea any form of credit or acknowledgment.
Table 3

<table>
<thead>
<tr>
<th>INTELLECTUAL PROPERTY (out of 12)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The company retains full rights in its intellectual property that it is</td>
<td>12</td>
</tr>
<tr>
<td>licensing to the customer</td>
<td></td>
</tr>
<tr>
<td>The customer retains full rights in his intellectual property that is</td>
<td>6</td>
</tr>
<tr>
<td>maintained on the company’s servers</td>
<td></td>
</tr>
<tr>
<td>The company obtains a license to use the customer’s content, not explicitly</td>
<td>3</td>
</tr>
<tr>
<td>limited to purpose for which it was originally submitted</td>
<td></td>
</tr>
<tr>
<td>The company obtains a license to use the customer’s content for marketing</td>
<td>1</td>
</tr>
<tr>
<td>purposes</td>
<td></td>
</tr>
<tr>
<td>The company owns all rights in any e-mails, suggestions, or ideas that the</td>
<td>11</td>
</tr>
<tr>
<td>customer sends to the company</td>
<td></td>
</tr>
</tbody>
</table>

C. Privacy Policies

There are some things that all or almost all of the policies in our sample addressed. All of the companies that we examined gave examples of situations in which they would gather users’ personal information and how they would make use of it, such as obtaining the user’s name, e-mail address, and other contact information in order to register the user’s account and process the user’s requests. Eighteen of the nineteen companies in our sample also purport to give their customers some control over the collection of their personal information, which may include the ability to opt out of data collection or the ability to access and edit personal information already on file with the company. Eighteen of the providers in our sample also addressed compliance with either TRUSTe or Safe Harbor, security concerns, changes to the privacy policy, and included a section about the use of tracking technologies.
Table 4

<table>
<thead>
<tr>
<th>19 out of 19 Privacy Policies included:</th>
<th>18 out of 19 Privacy Policies included:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data gathering by the company</td>
<td>Customer control over information</td>
</tr>
<tr>
<td>How the company uses that data</td>
<td>Security of user information</td>
</tr>
<tr>
<td>When may customer data be disclosed to third parties</td>
<td>Changes to privacy policy</td>
</tr>
<tr>
<td></td>
<td>Complies with TRUSTe or Safe Harbor</td>
</tr>
<tr>
<td></td>
<td>Use of tracking technologies</td>
</tr>
</tbody>
</table>

All of the providers in our sample also included information about sharing customer information with third parties. When the question turned to with whom and when a user’s personal information would be shared, however, the differences between the agreements began to stand out. For example, all nineteen companies included lawful government requests as a condition for disclosing customer information, ten said customer information may be disclosed in order to investigate, prevent, or take action concerning violations of the law or the company’s TOS, and four said that customer information may be disclosed so that the company can defend itself in court or assert its legal rights. Generally, the enumerated situations in which user information can be disclosed are fairly rational, with an eye to protecting customers’ privacy interests. However, two companies noted in their privacy policies that they may disclose customer information to the company’s business partners for the partners’ direct marketing purposes.
We identified eleven total categories of situations for information disclosure, and noted that out of our nineteen privacy policies, the companies listed between two and nine of these when describing information disclosure. Listing more categories of disclosure does not necessarily mean that a company is less protective of privacy, but it does underscore the variety and complexity inherent in analyzing these agreements.
People have been told for years to be careful of how much they disclose about themselves online in public forums, so it is not a surprise that many of the companies that provide chat forums and bulletin boards for their customers also include disclaimers in their privacy policies that information disclosed in these forums is not protected by the company’s privacy policy. Another fairly common-sense provision that fifteen of the nineteen providers include in their privacy policies is a disclaimer that the company does not control the privacy policies of third parties whose websites the customer may access through links on

the company’s website. These provisions usually advise the customers to read the privacy policies of the third-party websites. Of the nineteen, only Symantec and Citrix explicitly stated that their customers’ information may be disclosed to the company’s business partners for direct marketing purposes. Fifteen of the nineteen policies state that the company will not step outside the language of the privacy policy unless the customer gives consent. Disclosure to business partners is generally covered by another button on a form that a customer either checks or unchecks to give permission to the company to share information with these business partners for marketing purposes. Generally, the companies in this sample seem very aware of the negative press associated with selling user information to data farming firms.\footnote{See, e.g., Mitch Lipka, Twitter Is Selling Your Data, REUTERS (Mar. 1, 2012, 11:35 AM), http://www.reuters.com/article/2012/03/01/twitter-data-idUSL2E8DTEK420120301 (last visited Feb. 3, 2013) (“Twitter users are about to become major marketing fodder, as two research companies get set to release information to clients who will pay for the privilege of mining the data.”) (on file with the Washington and Lee Law Review); Jason Morris & Ed Lavandera, Why Big Companies Buy, Sell Your Data, CNN (Aug. 23, 2012 3:42 PM), http://www.cnn.com/2012/08/23/tech/web/big-data-acxiom/index.html (last visited Feb. 3, 2013) (“Acxiom . . . is just one of hundreds of companies who are peering into your personal life, collecting data that is generated from everything you do online . . . .”) (on file with the Washington and Lee Law Review).} Four companies in our sample note that they may disclose aggregated information for statistical purposes. Red Hat is one of the four, but its privacy policy also assures that once this information is aggregated, it is no longer traceable to the original individual.

Another element found in all but one of the privacy policies was a discussion of security measures to protect customer data. Data security is not directly related to data control, but as we noted above in Part IV.A.3, giving customers the power to withdraw and move their data may result in security-driven decisions to change services, thus giving companies incentive to implement stronger security measures to retain customers. Thirteen of the nineteen companies in our sample referred to their use of industry standard or commercially reasonable security measures. Some of these companies also listed specific technological or organizational measures in place, but others
simply included a vague statement about implementing industry standards. Another common element of security is encryption, and twelve of the nineteen stated that they use SSL encryption during the transfer of data. Twelve of the nineteen also stressed the importance of customers being proactive with the security of their own systems, and six emphasized that no electronic storage or transmission would ever be 100% secure. Only one company, Oracle, gave any information about what steps would be taken in the event that a customer’s user information was compromised.

Because information will be governed by different laws when it is located in different countries, companies also must keep jurisdictional issues in mind when describing their privacy practices. A majority of the companies (fifteen of nineteen) include a provision in the privacy policy notifying the customer that their data may be transferred to and processed in other countries. Only Amazon appears to give customers a meaningful choice of where their information is stored and processed, with its privacy policy stating that data will not be moved to other regions without the customer’s consent.

The question of where data is stored is significant, especially when the company has customers within the European Union. Because the European Union has strict privacy requirements,
U.S. companies that wish to transfer personal information from the European Union to the United States must certify compliance with the Safe Harbor program.\textsuperscript{513} Sixteen of the nineteen companies in our sample certify in their privacy policies that they are in compliance with this program. Two of the three companies that do not state compliance with Safe Harbor do, however, certify compliance with TRUSTe standards. Of the nineteen companies, eight certify compliance with both Safe Harbor and TRUSTe. Only one company, Citrix, does not refer to either the Safe Harbor program or adherence to TRUSTe standards in its privacy policy.

To exercise meaningful data control, consumers should be informed about how data is collected. Most of the companies in our sample detailed their use of Web tracking technologies within their privacy policies. Eighteen of the nineteen companies disclosed their use of cookies on their websites, sixteen of the nineteen disclosed that they used IP logs to track user behavior, and thirteen of the nineteen indicated that they used Web Beacons to track user behavior. Sixteen of the nineteen also listed other information that they collected from users, including information on the user’s browser and operating system, and other information that can be obtained using Javascript. Only four of the nineteen companies include flash cookies in the list of technologies utilized. However, as earlier research has noted, the use of flash cookies is sometimes unreported by companies.\textsuperscript{514}

\textsuperscript{513} See Commission Decision of 26 July 2000 pursuant to Directive 95/46/EC, 2000 O.J. (L 215) 7 (EC); see also EXPORT.GOV, supra note 400.

\textsuperscript{514} See supra note 473 and accompanying text (“The degree to which companies disclose the use of these tracking technologies varies.”).
Privacy policies and TOS agreements for services in the cloud may also change periodically to reflect new priorities or activities on the part of the provider. When companies are empowered to change the terms of agreements unilaterally in material ways, this can undermine user efforts to control data at the outset of the contractual relationship. The privacy policies of eighteen of the nineteen companies in our sample addressed changes to the privacy policy. All eighteen indicated that notice of material changes would be posted on the company’s website. Only six of those, however, gave any indication of either how long the notice would be posted or whether it would be posted before the changes went into effect. Eleven of the eighteen also indicated that the company might contact the customers directly to notify them of material changes to the privacy policy. These varying approaches to keeping consumers informed of changes are troubling, and are also reminiscent of our findings in the TOS agreement portion of this study, where we found that of the seven companies that discussed changes to the service, only two indicated that the company would notify current customers of the changes.
D. Analysis and Discussion

In analyzing these documents, we placed the provisions into broad categories. For privacy policies, the categories included provisions listing how and why personal information would be gathered, when and with whom it would be shared, security of information, whether personal information is transferred to other jurisdictions, the collection of non-personally identifying data, provisions addressing third party content and advertisements, and provisions addressing customer control over information. In terms of these larger categories, the companies that we looked at generally have similar priorities. With respect to security, the companies that we examined were more likely to speak in vague terms, with little specific detail (with the exception that most companies referenced SSL encryption). The TOS agreements that we analyzed tended to be more detailed, but this may be because TOS agreements are designed to protect specific rights of the companies. In a TOS agreement, a company will generally address the company’s liability for harms to the customer and what recourse a customer may have. A company is also likely to address intellectual property issues, though their primary
interest is typically on protection of the company’s rights rather than the customer’s rights.

As noted above in Part IV.A.4, consumers are likely to be unable to enforce a company’s privacy policy against it. Under § 5 of the FTC Act, however, the FTC is empowered to take action against companies that engage in unfair or deceptive trade practices, and the FTC has a precedent of using this authority to take action against companies that fail to comply with their own privacy policies. Given the possibility that a company could face legal problems for failing to comply with its own privacy policy, this gives a perverse incentive for a company to commit to as little as possible and inform its customers of as little as possible within its privacy policy. This threat of legal recourse may provide a partial explanation for why the companies that we examined are generally vague with respect to the rights of their customers and the methods used to protect their data.

In general, the privacy policies and TOS agreements that we have examined are much more protective of the company that wrote them than of the customer that agrees to them. This is not much of a surprise, especially considering the literature about asymmetric “click wrap agreements” that customers are often required to agree to in order to install software or use services online.

---

516. See id. § 45.
517. See supra Part IV.A.4 (describing FTC enforcement efforts).
518. See, e.g., Jared S. Livingston, Comment, Invasion Contracts: The Privacy Implications of Terms of Use Agreements in the Online Social Media Setting, 21 ALB. L.J. SCI. & TECH. 591, 625 (2011) (“There are several kinds of asymmetric information in this market: (1) failure to read provided information about the agreement; and (2) failure to appreciate the risk of loss of private information.”); Lucille M. Ponte, Getting a Bad Rap? Unconscionability in Clickwrap Dispute Resolution Clauses and a Proposal for Improving the Quality of These Online Consumer “Products,” 26 OHIO ST. J. ON DISP. RESOL. 119, 119 (2011) (“In undertaking these online transactions, millions of consumers each day simply click on ‘I Agree’ to a site’s standard terms of use, often without reading or understanding the terms and conditions of their purchases.”).
E. Implications

Publicly traded companies respond to public demand. As long as consumers ignore privacy policies and TOS agreements, accept that their control over their own data is subject to the will of the service provider, and remain uninformed about existing privacy threats, the companies have little incentive to write privacy policies and TOS agreements with their customers’ best interests in mind. On the other hand, if a data control policy became a legal mandate, the customer service experience of various cloud providers could see a marked improvement as providers compete to retain customers. Our research of the literature and our own analysis of the terms of TOS agreements and privacy policies lead us to conclude that there is currently a significant failure on the part of the market to ensure that consumers have sufficient control over their data in the cloud. We argue that this market failure is something that must be remedied to ensure the protection of personal privacy in the cloud.

Given the prevalence of the data trade and the value of profiles to marketers, consumers who are willing to trade personal information and privacy for free services are not just using their own data as currency; to some extent, it could even be said that they are becoming a commodity themselves. This commodification is part of a trade-off, and consumers may even be trading some of their legal rights in exchange for services in the cloud. Currently, a company that provides services to individuals and small businesses can demand any number of allowances as to the use of customer data in its privacy policy, and can substantially limit customers’ permissible actions in its TOS agreement. The customer has no bargaining power to challenge these terms. These terms, in turn, can affect the customer’s legal rights, limiting the extent to which the Fourth Amendment and the SCA protect data that the customer entrusts to the company, and potentially even making the customer vulnerable to liability under the DOJ’s current interpretation of the CFAA.

Throughout this Article, we promote the idea that consumers should have the ability to exercise meaningful control over their own data, including the ability to withdraw their personal information and move their data from one provider to another. If
a regulatory intervention like we propose in this Article gives consumers the power to exercise control over their data, but the consumers choose to not exercise these data control rights and still choose to trade their information for various services, such informed decisions may indicate that there is not a systemic flaw in this approach to personal data. However, we argue that the current utter lack of meaningful control prevents us from determining if the current data trade business model can be optimal.

VI. Recommendations—Building a Baseline for Facilitating Transactions in the Cloud

In the United States, privacy is largely protected using narrow laws that apply only to specific categories of information. To the extent that laws of general applicability apply to privacy in the cloud, like the Fourth Amendment and the SCA, customers may inadvertently remove their own privacy protections by agreeing to excessively broad terms in a cloud service’s privacy policy. Even though the FTC has brought actions against companies that violate their own privacy policies, these actions arguably serve only to give the providers incentives to write privacy policies that are as vague about the providers’ obligations as possible.

After examining the privacy policies and TOS agreements in our sample and analyzing a variety of legal issues and privacy theories, we have arrived at a series of recommendations to what we see as the failure of the contractarian paradigm to adequately protect parties that indicate agreement with these terms. We recommend a new legal regime that would emphasize empowering consumers by setting a baseline of protection to ensure that a consumer has control over her own data. The baseline would be designed to protect the most sensitive information without hindering market development.

A. Building the Baseline

One of our foundational arguments is that relatively modest regulatory intervention into the relationship between providers
and consumers could support positive social change with regard to privacy protections. To some extent, legal regulations can provide structure for social interactions, and the strength of the legal control can affect perceptions of social control and personal freedom.  

Regulating privacy would involve the regulation of relationships, perhaps by placing limits on organizational power.  

When implementing a legislative system to address problems, policy makers can either choose to implement rules, which tend to focus on strict requirements, or standards, which tend to be more flexible and open-ended. In regulating technologies, standards may be superior to rules because standards are more adaptable to further technological change. Detailed and inflexible sets of rules can either chill technological development, or in the alternative can quickly become obsolete if the progress of technology continues unimpeded. On the other hand, if the implemented regulations are too open-ended and vague, they can end up being entirely ineffective. A study by Birnhack and Elkin-Koren questions the very idea that regulation of personal data collection and use would be effective at all. While we do not suggest specific language for regulations

519. See Solove, Architecture, supra note 172, at 1240–41 (explaining his use of the term “architecture” to describe the protection or diminishing of privacy in our society).  
520. See id. at 1242 (“Protecting privacy thus depends upon regulating relationships, often by enforcing limits on the power of bureaucratic organizations.”).  
521. See Schwartz & Solove, supra note 3, at 1871–72 (describing how “standards are generally the superior choice for dealing with situations of rapid change because . . . rules can become obsolete”).  
522. See Solove, Architecture, supra note 172, at 1275 (summarizing research that shows how regulations, “if too specific, can quickly become obsolete, discourage innovation, and be costly and inefficient”).  
523. See id. (“However, rules that are too open-ended and vague can end up being toothless. Although security standards must not be overly specific, they must contain meaningful minimum requirements.”).  
524. See Birnhack & Elkin-Koren, supra note 108, at 343 (noting a low level of compliance with information privacy laws across several categories of websites in Israel). The authors noted that popular websites were more likely to comply with the privacy protection laws, perhaps because popular websites were likely to be maintained by organizations with the resources to have legal departments, and perhaps because complying with the law also serves as a
in this Article, we encourage policy makers to construct a regime that strikes a balance between rules and standards to make the new data protection regime specific enough to address discrete problems and open-ended enough to allow it to evolve.

1. Baseline Regulation

We recommend a regime that includes baseline privacy protections that would set a floor for the permissible approaches of companies that handle consumer information. The variation in the approaches taken by companies in our relatively small sample underscores the need for more uniformity.

Many questions exist about the appropriate levels of baseline protections, posing interesting questions for future research. Baseline regulations should first identify minimum requirements in order to protect certain types of sensitive information. Such regulations should explicitly address the protection of personal health information, social security numbers, and financial information like bank account numbers and credit cards. The baseline regulation could also include a provision that places the risk of loss for online fraud on a cloud provider. Opponents of our approach may point to the results of the Birnhack and Elkin-Koren study, an empirical study of Israeli websites that suggests that regulations setting a baseline for privacy agreements are not truly effective due to low compliance rates. But the authors of that study failed to focus on enforcement, and more effective enforcement would likely improve the efficacy of such regulations.

After establishing categories of sensitive information that must receive special protection, the next question concerns what minimum requirements should be included to protect consumer information. Baseline regulations might, for example, include

signal to consumers that the company is more reliable. See id.
525. See Soghoian, supra note 5, at 378–79 (noting that cloud computing providers do not have the same incentive as banks and online merchants to protect customers from online fraud because the banks and online retailers legally bear the risk of loss instead of the consumer).
526. See Birnhack & Elkin-Koren, supra note 108, at 383 (describing how the authors “found that some areas of the law are simply irrelevant in the daily practices of websites”).
requirements for data security. End users are generally ignorant of many data protection issues, so there is not sufficient market demand for firms to pay more attention to security issues like the need to encrypt information. This could be addressed using regulations that require data to be encrypted. We envision two primary options for security baseline regulation: language requiring the use of “best available security technology,” or language requiring the use of “industry standard security technology.” The comparison of these two options is another possible direction for future research.

We further suggest that baseline regulations should also address issues related to data breaches. First, the regulation should include security breach notification requirements in order to give users the information necessary to assess the negative consequences of a cloud vendor’s security failures. Second, there should be viable private causes of action for data breaches to address the current problem of consumers not having standing to sue a company after a breach in the absence of a distinct injury like identity theft. Some scholars have suggested finding companies strictly liable for data leaks, creating a new common law tort based on the use of Fair Information Practices, and

527. See Soghoian, supra note 5, at 380 (stating that many consumers know very little about data encryption and describing how this provides “no incentive to [devote resources] to something for which most customers have not expressed a want”).

528. See id. at 382–83 (proposing that government regulators require cloud service providers to use encryption just as this has already been done in the banking and health industries).

529. Martin suggests a similar approach. See Martin, supra note 44, at 313 (“Congress should create new breach notification requirements that allow users to assess the exposure, damage, and operational costs of any security failures on the part of a cloud vendor.”).


531. See Sarah Ludington, Reining in the Data Traders: A Tort for the Misuse of Personal Information, 66 Md. L. Rev. 140, 140 (2006) (suggesting “a new common law tort . . . to force reform and accountability . . . and to provide remedies for individuals who have suffered harm to their core privacy interests” and stating that this tort “borrows from . . . the Fair Information Practices from the Privacy Act of 1974”).
imposing liability for breach of trust if a company misuses information.\textsuperscript{532} Imposing fiduciary obligations in some circumstances may also provide adequate private causes of action in response to security breaches. It is possible that some of the issues relating to information privacy could be resolved through the common law, such as if privacy tort law were expanded to take intangible harms into account, including the harm from the disclosure of data that is not embarrassing.\textsuperscript{533} We argue that an emphasis on private enforcement options would preserve the viability of the market by limiting excessive legislative oversight of business practices.

\textbf{B. Data Control}

The most important part of our proposal for a new legal regime that sets a floor for the use of data by private companies concerns data control, which we have defined in this Article as encompassing the ideas of data mobility and data withdrawal. In the cloud context, there are two sets of information that we are concerned about: PII, and what we call “course-of-business” data that is stored as part of the customer’s use of the service. Related to the use of PII, we are also concerned about secondary use of such information, including secondary use by third parties. We further argue that data mobility and data withdrawal provisions as described below would attract consumers who are more risk averse and who would not use these services in the absence of these protections, thus leading to a net benefit to the industry.

\textsuperscript{532} See Jessica Litman, \textit{Information Privacy/Information Property}, 52 \textit{STAN. L. REV.} 1283, 1288 (2000) (“[A] rubric based loosely on breach of confidence might persuade courts to recognize at least limited data privacy rights.”).

\textsuperscript{533} See Richards & Solove, \textit{supra} note 1, at 1922–23

Courts can readily understand the harm caused by the disclosure of a naked photograph of a person, but they struggle in locating a harm when non-embarrassing data is disclosed or leaked. A broader understanding of harm is needed in order for the privacy torts to apply to the extensive gathering, dissemination, and use of information by various businesses and organizations.
Thus, these provisions should be mandatory, and the regulations should prevent parties from contracting around these terms.

1. Personally Identifiable Information

Baseline privacy regulations should protect the ability of consumers to control the use of their PII in the cloud. The security of PII should be paramount to prevent fraud and identity theft. This part of our recommendation is by no means revolutionary, however, because privacy policies are centered on protection of PII, and most privacy theorists focus on PII as the class of information that must be afforded the most protection.

We also encourage discussions of PII to consider the commodification of data. If consumers are free to use their PII as a form of currency, disclosing it to obtain desired services, should there be limits on what information consumers can trade? If privacy is viewed as property, and property itself is really a bundle of rights, there may be some types of information where it would be against the best interest of society to permit the free trade thereof. For example, the relationship between doctors and patients is typically viewed as sacrosanct. We thus suggest that personal health information is one category of information that service providers outside this circle would not be able to seek under our proposed legal regime.

The problem of reidentification raises additional issues because it can lead to anonymized, descriptive information about the consumer being reattached to the consumer’s identity. While we would not recommend a regime that stifles innovation and academic creativity, a legal regime to protect PII in the cloud also needs some forward-looking provisions addressing the possibility that reidentification science could lead to threats to personal privacy in the future. These provisions, for example, might prohibit the use of public records for reidentification purposes unless the user certifies compliance with some form of privacy standard.
2. Secondary Use

Secondary use of PII is another very important consideration. Those who argue for limitations on secondary use suggest that the use of data should be limited to the purpose for which it was initially collected, absent further consent being obtained.534 Existing rules prohibiting secondary use include legal ethics rules that prohibit a lawyer from using client information for a purpose unrelated to the interests of the client, and restrictions in the Fair Credit Reporting Act that prohibit an employer who obtains an employee’s credit report from using this information for nonemployment purposes.535

Once PII is properly collected, we suggest imposing further limits on secondary use of the PII. One option is to give the consumer the ability to restrict secondary use of her PII. Privacy policies often give the customer the ability to access and amend PII stored on the collecting company’s system, so requiring these provisions to address secondary use would likely not be excessively burdensome.

However, privacy policies do not give consumers control over PII given to third parties unconnected to the consumer. To address this third-party problem, the baseline regime should guarantee consumers a right of data withdrawal. By permitting data withdrawal when a consumer’s information is being used in a way that goes against the wishes of the consumer, we secure the right of consumers to control their data and feel more secure.536 To solidify this data withdrawal right, we recommend

534. See Richards, supra note 114, at 1190 (defining secondary use prohibitions as “the requirement that data collected for one purpose may be used for that purpose only, absent consent”); Solove, Taxonomy, supra note 111, at 521 (“Secondary use’ is the use of data for purposes unrelated to the purposes for which the data was initially collected without the data subject’s consent.”).

535. See Richards, supra note 114, at 1190–91. However, Solove argues that the restrictions in the Fair Credit Reporting Act do not adequately restrict secondary uses of covered information. See Solove, Digital Person, supra note 2, at 67–68 (describing how effective lobbying by the credit reporting industry led to an exemption for “names, addresses, former addresses, telephone number, SSN, employment information, and birthdate”).

536. Our proposed right of data withdrawal is ideologically similar to the proposed “right to be forgotten” in European privacy law, which is supported by the European Commission, though many worry that a right to be forgotten is
giving consumers the ability to serve a notice-and-takedown order on third parties to require the removal of the consumer’s PII from the third party’s system. We recognize that consumers may have difficulty obtaining information about the secondary use of their PII, but argue that combining a notice-and-takedown regime with controls to enable meaningful informed choices could potentially address some of the problems relating to the secondary use of PII by third parties. Designing controls to enable meaningful informed choices is outside the scope of this Article, but it is an important and related issue that should be the subject of further study.

Under a regime allowing for notice and takedown of PII, a party who wants his PII removed from a specific service could contact the operator of that service to (1) assert his rights in the PII, and (2) request that the PII be taken down. At that time, the operator would have to comply and notify the original submitter of the information about the takedown. The original submitter would then have an opportunity to contest the takedown and assert that the PII was not wrongfully made available. This proposal of a notice-and-takedown approach is patterned after the procedures of the Digital Millennium Copyright Act (DMCA), which permits copyright owners to serve a notice on a website when infringing material has been posted. Our notice-and-takedown proposal would permit consumers to request that entities take down information that was either posted by the consumer and then republished elsewhere, or that was derived from information posted by the consumer. The notice and


538. See 17 U.S.C. § 512(c)(3) (2006) (establishing a “notice and takedown” procedure to handle allegations that content on a host website infringes an owner’s copyright).
takedown approach could apply to secondary use by the original entity entrusted with the information, as well as to third parties.

Another option we suggest is for the baseline regulation to declare that some information, like personal health information, should never be tradable. Thus, someone within the necessary circle encompassing the doctor–patient relationship would not be able to trade health information, even if it is anonymized, to marketers seeking to create profiles based on health needs. If some information is not tradable but others are, this can still leave room for many different business models to survive, as long as a minimum level of privacy and security are provided.

3. Course-of-Business Data

Recommendations about PII are very common in the privacy literature, but the information disclosed to cloud providers goes far beyond PII. One of the elements that we think deserves more discussion is the control of what we call “course-of-business” data, which consumers store with cloud providers as part of the service. Many cloud services permit customers to store photos, writings, and business data in the cloud. The storage of this information is often the customer’s purpose for using this service to begin with, whereas the transfer of PII is typically incidental to the rendering of service. Because the storage of this information is essential to the service, the terms relating to such storage should be explicit as a condition of the contract between the parties.

In Part II, we noted that many private actors have called for improved transparency and control in the cloud. In the cloud context, the question of data control does not only involve targeted advertising, but also the importance of data mobility so that customers would not lose everything if a service provider became inoperable or if the data had to be moved to a new service provider.\textsuperscript{539} However, as our analysis of TOS agreements and privacy policies showed, companies often do not address the handling of such data after the contract has terminated.

\textsuperscript{539} Martin, \textit{supra} note 44, at 286 (“Any solution needs to incorporate guarantees that data owners would be able to gain control of their data in a usable form should their service providers become inoperable.”).
Above we emphasized the data withdrawal aspect of data control. The right of withdrawal may have less application to course-of-business data like writings and photos because such data may be protected by intellectual property law (IP law), and therefore a right of withdrawal may be duplicative of IP law protection. However, for course-of-business data entitled to lesser IP law protection, like databases, the right of withdrawal via notice and takedown should be available.

More importantly, the baseline regulations must require a minimum level of protection to ensure data mobility. This means that data must be converted to an acceptable format before being delivered to a departing customer, such that the customer is not locked in to a particular service provider, and could easily move their data from one provider to another. Data mobility focuses on the access and consumer choice aspects of data control and would facilitate market transactions by enabling customers to move their data freely between competing services. What happens if the customer decides for any reason that she wants to use the cloud services of a competing provider? Is a user’s course-of-business data stored in a proprietary format such that the user encounters a “lock-in” problem if she decides she wants to change providers? Currently, privacy policies and TOS agreements often may not address these issues at all. As part of the legal regime that we propose, format transparency would be required, and providers would also be required to include terms addressing end-of-relationship handling of course-of-business data. Under our proposed regime, a company could still store the data in a proprietary format, but would be required to convert the data to a generally accepted format upon account termination to enable the data to be easily moved to a competing service.

Data mobility is important because it allows consumers to more fully participate in the features and services that cloud providers offer. The importance of data mobility in the cloud can be emphasized by analogizing to mobile phone numbers. In November 2003, an FCC regulation became effective that required cell phone carriers to allow numbers to be ported from one carrier to another.540 There was a great deal of resistance on

---

540. See Telephone Number Portability, First Report and Order and Further
the part of service providers that claimed that this rule would be too costly to carriers and might not be beneficial to consumers. The FCC, however, concluded in 2006 that the number portability requirement did not significantly increase “wireless churn,” and did in fact have a positive impact on service quality due to the need that it created for carriers to devote extra effort to customer retention. We expect that a data mobility requirement may be met with the same initial resistance as the wireless number portability requirement, but that like wireless number portability, data mobility requirements will have a net positive effect on both the industry and on consumers. By allowing consumers to “port” their phone numbers into another provider’s system, cellular subscribers are better equipped to participate in the market because such porting greatly reduces costs that might otherwise be associated with switching mobile service providers. Similarly, data mobility in the cloud would facilitate consumer participation and reduce transaction costs for consumers when moving from one provider to another.

Protection of course-of-business information could also be achieved through some application of the principles surrounding Notice of Proposed Rule, 11 FCC Rcd. 8352 (1996). The first compliance date was set for June 30, 1999, but after two requests for forbearance, the agency pushed the deadline for compliance to November 24, 2003. See Cellular Telecommunications & Internet Ass’n v. F.C.C., 330 F.3d 502, 503–04 (D.C. Cir. 2003).

541. See Caron Carlson & Carmen Nobel, Carriers Resist Porting Numbers, EWEEK, Apr. 21, 2003, at 20 (describing how “[w]ireless carriers [were] looking for relief from a requirement that would . . . allow cell phone customers to keep their numbers when they change phone companies”).


[T]he advent of porting . . . did not lead to a significant increase in wireless churn, but did appear to have had a positive impact on service quality by inducing carriers to engage in aggressive customer retention efforts . . . . Significantly improved retention efforts (better deals on upgrade handsets, incentives for signing longer contracts, better customer service, and higher network spending) following the implementation of local number portability . . . have led to lower churn rates . . . (citations and internal quotations omitted).

543. See Carlson & Nobel, supra note 541 (noting that the extra cost associated with changing a cellular phone number was sometimes viewed as the most important reason to stay with the same provider).
the law of confidentiality. Confidentiality as a concept is related to privacy, but primarily arises in the context of contracts between private parties. When fiduciary relationships exist, the law often recognizes obligations to keep information confidential. Solove has pointed out the potential application of fiduciary relationships in the privacy context. If elements of fiduciary relationships were integrated into the customer–cloud provider relationship, this would impose on the providers an obligation to keep not only the customer’s PII secure, but other data stored on the provider’s servers as well. Consumers could obtain stronger protections by opting into a fiduciary relationship with the service provider for a price. The consumer would thus get a guarantee that if the service provider acts badly, the consumer has a right of action against them. This differs from Solove’s proposal because we are more focused on consumer choice than on making fiduciary relationships into a default rule.

VII. Conclusion

Privacy issues online are not going to disappear overnight. Changes to the law are necessary to facilitate optimal market development that takes into consideration the autonomy of consumers in controlling their personal information. Foucault’s view of Bentham’s Panopticon as a metaphor for power relations in society is even more apt today than when Foucault was originally writing. The market forces peering into private lives may not be doing so with malicious intentions, but the corresponding decrease in consumer control of their personal information is nonetheless harmful.

In this Article, we have examined issues relating to cloud computing through the lens of privacy theories and privacy law. In analyzing a sample of terms of service agreements and privacy policies, we have concluded that these documents have potentially serious implications for the rights of consumers who agree to them without reading the terms.

544 See Solove, Digital Person, supra note 2, at 103 (making the “radical proposal” that the law should recognize a fiduciary relationship when a company collects and uses personal information).
Ultimately, we recommend the implementation of baseline regulations to guarantee some minimum level of protection for consumers in the cloud. These regulations should emphasize the importance of preserving consumer control of their data, and these control mechanisms should focus on data mobility and a right of data withdrawal. Data mobility will require cloud providers to make consumer data available in a generally acceptable format such that consumers can freely move their own data from one provider to another in the interest of maintaining a healthy, competitive marketplace. For data withdrawal, we propose a notice-and-takedown approach patterned after similar provisions in the DMCA, which would permit a consumer to request that entities take down his personal information.

Our proposal raises a number of interesting new research questions. One of the most interesting problems is the effect that our proposals would ultimately have. Once the efficiency of the market is protected and consumers are in control of their data, would there actually be any statistically significant changes in consumer behavior? At the end of the day, if consumers are empowered to control their data, but behavior is largely unaltered, this may indicate that the current state of the market is actually optimal for society. However, the uncertainty that currently exists with regard to data ownership is harmful to consumer autonomy and makes it impossible to conclusively determine the optimality of the current regime. Thus, reduction of this uncertainty is essential to protecting the interests of consumers, and sufficient reduction will likely require some degree of regulatory intervention. We posit, however, that the degree of this regulatory intervention could be very modest, with narrow goals focusing on minimum protections and consumer choice, thus balancing the need to protect consumers with the need to preserve market vitality.
Docket Dividends: Growth in Shareholder Litigation Leads to Refinements in Chancery Procedures

Donald F. Parsons, Jr.*
Jason S. Tyler**

Table of Contents

I. Introduction ................................................................. 474

II. The Importance of Shareholder Representative Litigation to the Substantive Law of Corporations, or a Brief History of Why the Court of Chancery Hears Shareholder Cases .............................................. 479

III. Recent Issues in Chancery Procedure .............................. 487
   A. Corporate Benefit on a Contingency Basis:
      Calculating Attorneys' Fees ....................................... 488
   C. Hercules and the Hydra: The Multi-Jurisdictional Litigation Problem ...................................................... 506

* Vice Chancellor, Delaware Court of Chancery.
** Associate, Sullivan & Cromwell LLP.

This Article is based on a speech delivered by Vice Chancellor Parsons at the Washington and Lee University School of Law on March 26, 2012. We would like to thank the following individuals for their helpful comments and discussions: William B. Chandler, III, Jill Fisch, Lawrence Hamermesh, Hon. Jack Jacobs, Marcel Kahan, Hon. J. Travis Laster, Christine Shepard, Francis Stapleton, and Hon. Leo E. Strine, Jr. We also would like to thank Daniel Kirshenbaum, Matthew Murphy, and Nicole Sciotto for their research assistance. Lastly, we thank the Washington and Lee Law Review and all of its faculty advisors and student staff for their invaluable efforts in readying this Article for publication. The views expressed in this Article do not necessarily represent the views of the Delaware Judiciary, Sullivan & Cromwell LLP, or any of the individuals whom we consulted in our research. All errors are our own.
I. Introduction

Every year or so, the caseload of the Delaware Court of Chancery tends to gravitate toward one or more timely topics in the law of corporations. In the mid-2000s—after a booming economy but a triad of corporate scandals at Enron, Worldcom, and Tyco—the topic was backdating of stock options. That issue culminated in the Court’s decision in Ryan v. Gifford. From that case, and a handful of others involving similar issues, a fairly cohesive body of law developed around 2007 regarding potential improprieties in the issuance of stock options. After the subprime mortgage crisis, the issue of the relationship between directors’ duty of oversight and risk management came to the fore. In In re American International Group, Inc. Consolidated Derivative Litigation (AIG), the Court of Chancery held that corporate insiders could be personally liable for failing to implement and monitor effective internal reporting controls against fraudulent

---

1. Ryan v. Gifford, 918 A.2d 341, 358 (Del. Ch. 2007) (holding that the intentional violation of a shareholder-approved stock option plan is disloyal to the corporation and constitutes bad faith conduct).

2. See, e.g., MBKS Co. v. Reddy, 924 A.2d 965, 967 (Del. Ch. 2007) (regarding grant of common stock for inadequate consideration), aff’d sub nom. Reddy v. MBKS Co., 945 A.2d 1080 (Del. 2008); In re Tyson Foods, Inc., 919 A.2d 563, 592–93 (Del. Ch. 2007) (regarding “spring-loaded” options). Another important line of cases clarified how the duty of good faith functions as part of the fiduciary duties owed by corporate directors. See, e.g., Stone v. Ritter, 911 A.2d 362, 369–70 (Del. 2006). After much speculation about whether the duty of good faith was a third fiduciary duty, the Delaware Supreme Court defined the duty of good faith as “a subsidiary element, i.e., a condition, of the fundamental duty of loyalty” and not as a discrete fiduciary duty. Id. at 370 (internal quotation marks omitted).

financial transactions. Shortly after the AIG decision, the Court again addressed directors’ duties of oversight, but this time as they related to business risks as opposed to compliance risks. The Court of Chancery held in In re Citigroup Inc. Shareholder Derivative Litigation that “the mere fact that a company takes on business risk and suffers losses—even catastrophic losses—does not evidence misconduct, and without more, is not a basis for personal director liability.” And just two years ago, as the economy showed tepid signs of improvement and a handful of corporate acquirers went bargain hunting, the poison pill made a curtain call. In four successive cases, Selectica, Inc. v. Versata Enterprises, Inc., Yucaipa American Alliance Fund II, L.P. v. Riggio, eBay Domestic Holdings., Inc. v. Newmark, and, most prominently, Air Products & Chemicals, Inc. v. Airgas, Inc., the Court of Chancery and the Delaware Supreme Court examined the use of the poison pill in a wide variety of factual scenarios. Collectively, these cases provided new clarity on when and how directors can make use of this defensive measure.

4. Id. at 776–78, 799.
6. Id. at 130. As the court further expounded, [t]here are significant differences between failing to oversee employee fraudulent or criminal conduct [as allegedly occurred in AIG] and failing to recognize the extent of a Company’s business risk. . . . While it may be tempting to say that directors have the same duties to monitor and oversee business risk, . . . [t]o impose oversight liability on directors for failure to monitor ‘excessive’ risk would involve courts in conducting hindsight evaluations of decisions at the heart of the business judgment of directors. Oversight duties under Delaware law are not designed to subject directors, even expert directors, to personal liability for failure to predict the future and to properly evaluate business risk.

This phenomenon of multiple cases posing different facets of timely questions of corporate law is not the result of intelligent design or stellar forces, but the natural consequence of the Court of Chancery’s role as the United States’ premier business court. Delaware is home to just shy of one million business entities—more than its population—and more than 60% of the companies comprising the Fortune 500.11 The internal affairs doctrine holds that internal disputes between a company’s managers and shareholders are governed by the law of the state of incorporation.12 And, if Delaware law applies to these disputes, it makes sense for Delaware jurists—at the trial level, the Chancellor and four Vice Chancellors of the Court of Chancery and, at the appellate level, the five Justices of the Delaware Supreme Court13—to be the ones to apply it.14 Consequently, the


12. See CTS Corp. v. Dynamics Corp. of Am., 481 U.S. 69, 89 (1987) (“No principle of corporation law and practice is more firmly established than a State’s authority to regulate domestic corporations . . . .”); In re Topps Co. S’holders Litig., 924 A.2d 951, 953–54 (Del. Ch. 2007) (“[W]hen a corporation forms under the laws of a particular state, the rights of its stockholders are determined by that state’s law . . . .”); RESTATEMENT (SECOND) OF CONFLICTS OF LAWS §§ 302–10, 313 (1971) (discussing a state’s interest in regulating corporations established under its laws).


14. There are a number of principled justifications, from a societal perspective, for preferring the judges of the state of incorporation to adjudicate disputes concerning a corporation’s internal affairs. Perhaps most obviously, doing so “promotes the consistent application of relevant doctrine . . . . That, in turn, promotes the growth of precedent to guide future transactions. Fostering that consistency and growth is the primary public interest implicated in corporate and contract cases, because consistency and predictability contribute to wealth creation.” Leo E. Strine, Jr., Lawrence A. Hamermesh & Matthew C. Jennejohn, Putting Stockholders First, Not the First-Filed Complaint, 4 (Harvard John M. Olin Ctr. for Law, Econ. & Bus. Discussion Paper No. 740, 2013), available at http://ssrn.com/abstract=2200499. Thus, although “judiciaries throughout the United States are staffed with competent judges capable of producing serviceable opinions applying other jurisdiction[s’] laws[,] . . . doctrinal complexity means optimal applications of a jurisdiction’s
Delaware courts often adjudicate cases involving many of the country’s major companies and issues stemming from the latest developments affecting the national economy. It is not surprising, therefore, that several independent cases dealing with similar issues related to the financial headlines, especially those questions having potentially systemic ramifications, arise at any particular time.

Unlike most of the issues du jour that have arisen in recent years, the issues that arose in 2011 and early 2012 did not invoke a discrete, substantive question of corporation law. Rather, the issues recently at the forefront relate to the growth in shareholder representative litigation itself. Five years ago, in 2007, less than 40% of public company merger and acquisition (M&A) deals were challenged. By 2010, over 80% of those transactions gave rise to litigation, with each challenged transaction drawing an average of nearly five independently filed complaints. The preliminary data for 2011 shows that such
shareholder suits are becoming even more ubiquitous: in 2011, over 90% of M&A deals over $100 million were challenged. This rise in shareholder representative litigation does not pertain necessarily to any particular substantive area of corporation law. Nevertheless, it has broad implications. The plethora of cases has brought into sharper focus the civil procedure of shareholder representative actions. The results may affect how lead plaintiffs, their counsel, defense counsel, and courts in multiple jurisdictions can best handle such litigation involving Delaware companies and ensure that it effectively serves the purposes for which representative forms of shareholder litigation were created.

This Article addresses emerging issues in the civil procedure of shareholder representative litigation along two parallel tracks. First, after a brief overview of the legal basis for the Delaware Court of Chancery’s subject matter specialty in shareholder representative litigation in Part II, Part III summarizes the effect that the growth in shareholder representative actions has had on four areas of Delaware law: (1) awards of attorneys’ fees to plaintiffs’ counsel; (2) certification and removal of lead plaintiffs; (3) motions to stay or dismiss because of concurrent litigation in another jurisdiction; and (4) the interaction of pleading rules, forum shopping, and statutory books and records actions under Delaware General Corporation Law (DGCL) Section 220. Second, the Article attempts to demonstrate how Delaware’s volume of corporate and alternative business entity cases and the responsiveness of its courts, its legislature, and the legal marketplace generally accelerate the development of refined doctrine, measured balance, and valuable predictability. In short, with repeat experience comes ready expertise and real efficiency.21

II. The Importance of Shareholder Representative Litigation to the Substantive Law of Corporations, or a Brief History of Why the Court of Chancery Hears Shareholder Cases

To apprehend the importance of the Delaware Court of Chancery’s procedural rules to Delaware’s substantive law of corporations, some brief background on the origins of equity practice may be useful. Though only an archaic concept for most common law jurisdictions, the distinction between “law” and “equity” is the historical source of the Delaware Court of Chancery’s jurisdiction to hear most corporation cases. Except where enlarged by statute, the Court of Chancery exercises the subject matter jurisdiction that the English High Court of Chancery possessed as of 1776. Legal historians often characterize English common law before the nineteenth century as insufferably rigid. The writ system developed inflexible rules of pleading and evidence that were designed poorly in the first place and, thereafter, rarely modified, despite the recognized hardships they caused. A stock example comes from the “use,” unusual and remarkably effective regulatory machine” over “a broad field of economic activity”). Because the Court of Chancery “sees and has the power to regulate a vast amount of M&A activity, its perspective is not episodic or narrow, but constant and, if not complete, very substantially representative.”

22. See, e.g., DuPont v. DuPont, 85 A.2d 724, 727 (Del. 1951) (“[T]he general equity jurisdiction of the Court of Chancery . . . is defined as all the general equity jurisdiction of the High Court of Chancery of Great Britain as it existed prior to the separation of the colonies . . . .”); Glanding v. Indus. Trust Co., 45 A.2d 553, 558–59 (Del. 1945) (“[T]he Court of Chancery shall . . . exercise that complete system of equity jurisdiction [that the High Court of Chancery possessed as of 1776] in all respects until the Legislature of this State shall provide otherwise.” (emphasis added)).

23. See, e.g., J. H. Baker, An Introduction to English Legal History 53 (4th ed. 2002) (“[T]he formulae through which justice was centralised and administered by the king’s courts in the twelfth and thirteenth centuries were frozen as part of the ‘due process of law’ guaranteed by charters of liberties, and gave rise to a formalistic legal culture which affected legal thought at every turn.”). Baker further notes that “[t]he possibilities of technical failure [under the common law writ system] were legion. And the growing strength of the substantive law could also work injustice, because the judges preferred to suffer hardship in individual cases than to make exceptions to clear rules.” Id. at 102.

24. See id. at 97 (“Although the jurisdictions were adjusted, and the procedures were developed, distorted or evaded in different periods, the
the prototype of our modern trust. In a use, A gave legal title of property to B for the use of C. The evidentiary rules of the common law, however, treated B's legal title as any other form of absolute ownership. Thus, if B absconded with, or otherwise interfered with C's use of, the property, the courts of law could not provide a remedy. Over time, the sovereign's oath to provide "equal and right justice" came to recognize a royal duty to furnish a remedy whenever the law courts' exercise of the crown's judicial powers proved deficient. For that reason, the Lord High Chancellor, essentially a political appointee to an administrative department, was empowered to disregard the formalism of the law, including the requirement of trial by juries and compel

essential premises and outward forms of the common-law system went almost unchanged between the thirteenth and the nineteenth centuries."; id. at 102–03 (noting that the common law favored "strict rules of evidence, rules which might exclude the merits of the case from consideration but which could not be relaxed without destroying certainty or condoning carelessness").

25. See id. at 248–49 (describing the historical evolution of the "use").

26. Id. at 102. Another example comes from the law of debtor-creditor relations. English evidentiary rules treated a sealed bond as incontrovertible proof of the existence of an outstanding debt. Id. Hence, if the debtor repaid the amount borrowed but failed to request back or otherwise destroy the bond issued to the creditor, the creditor could proffer the bond to a court of law and force the debtor to pay on it again. Id.

27. Id. at 98 (quoting the coronation oath of Edward II in the early fourteenth century).

28. Although the absence of juries commonly is considered one of the distinctive features of chancery practice, there is historical evidence that chancellors frequently submitted questions of fact to juries, albeit by use of legal fictions. Blackstone, himself a chancellor, wrote that:

[T]his court is so sensible of the deficiency of trial by written depositions [alone], that it . . . usually directs the matter to be tried by jury . . . . But, as no jury can be summoned to attend this court, the fact is usually directed to [a court of law] upon a feigned issue.

3 WILLIAM BLACKSTONE, COMMENTARIES *452. According to Blackstone, the "feigned issue" might take the form of an action at law for payment on a fictitious bet that, if the contested fact were true, one party owed the other a nominal sum; the parties would concede the existence and validity of the bet and contest only the decisive fact. Id. “[T]hereupon that issue is joined, which is directed out of chancery to be tried: and thus the verdict of the jurors at law determines the fact in the court of equity." Id.

Even today, the Delaware Court of Chancery possesses statutory authority to submit questions of fact to a jury. See DEL. CODE ANN. tit. 10, § 369 (2012). The Court of Chancery has interpreted this authority, however, to be "wholly
that which, in good conscience, ought to be done—in the above example, remedy B’s breach of an extralegal duty to C.\footnote{29} That led to the recognition, and to the enforcement, of fiduciary duties, i.e., extralegal obligations “to act with the highest degree of honesty and loyalty toward another person.”\footnote{30}

Because fiduciary duties are “extralegal,” their enforcement historically became the exclusive province of courts of chancery, in which chancellors administered equity instead of law.\footnote{31} Although most common law jurisdictions long ago merged law and equity in a single court,\footnote{32} Delaware has not.\footnote{33} Therefore, any action against a corporate director for breach of fiduciary duty

\footnereftr{29}{See \textit{BAKER}, supra note 23, at 103 (“The Chancery worked differently. The chancellor was free from the rigid procedures under which such injustices sheltered. His court was a court of conscience, in which defendants could be coerced into doing whatever conscience required in the full circumstances of the case.”).}

\footnereftr{30}{BLACK’S LAW DICTIONARY 581 (9th ed. 2009).}

\footnereftr{31}{Chancery practice also invented procedures and remedies to redress other hardships tolerated by the courts of law, such as the class action mechanism itself (to remedy the inequity of piecemeal adjudication when joinder of numerous interested parties was otherwise impractical) and the equitable remedy of cancellation of a bond (to relieve the burdens discussed \textit{supra} note 26). \textit{See} Amchem Prods., Inc. v. Windsor, 521 U.S. 591, 613 (1997) (noting that Rule 23 “stems from equity practice” (citing Benjamin Kaplan, \textit{Continuing Work of the Civil Committee: 1966 Amendments of the Federal Rules of Civil Procedure (I)}, 81 HARV. L. REV. 356, 375–400 (1967)); \textit{DAN B. DOBBS, DOBBS HORNBOOK ON REMEDIES} § 2.1 (1993) (discussing orders for the cancellation of instruments as among the remedies invented in equity practice)).}


\footnereftr{33}{See \textit{DEL. CODE ANN. tit. 10, § 341} (1953) (“The Court of Chancery shall have jurisdiction to hear and determine all matters and causes in equity.”).}
brought in a Delaware state court must be brought in the Court of Chancery rather than the Delaware Superior Court, Chancery’s sister court of law.34

As a substantive matter, the law of corporations is designed in large part to remedy collective-action problems in commerce.35 Capital markets permit efficient allocation of inexpensive capital to worthy enterprises, but also produce risks problematic for decentralized governance. If an enterprise has many thousands of dispersed owners, only a few possess the economic incentive to inform themselves fully of all material information before participating in daily business decisions.36 Corporation law addresses this risk by vesting control over the company’s day-to-day affairs in a group of professional managers, who act as the agents of their shareholder-principals.37 Thus, for example, the DGCL provides that “[t]he business and affairs of every corporation . . . shall be managed by or under the direction of a board of directors.”38 Because the board’s function is to generate wealth by taking reasonable, good faith risk in deploying the

34. The Delaware Court of Chancery’s general equitable subject matter jurisdiction, though often characterized as engendering a specialization in corporate governance matters, means that the Chancery is actually not exclusively a business court. See generally WOLFE & PITTENGER, supra note 16, § 2.03 (noting the Court’s jurisdiction extends to all matters involving equitable rights—e.g., trusts, estates, and guardianships—as well as all requests for equitable relief—e.g., injunctions, specific performance, and cancellation of bonds).

35. See Jennifer Arlen & Eric Talley, Unregulable Defenses and the Perils Shareholder Choice, 152 U. PA. L. REV. 577, 589 n.32 (2003) (“[T]he very reason publically held corporations exist is to exploit the advantages of vesting control of the firm in professional managers who do not own the firm.” (citing CHARLES R.T. O’KELLEY & ROBERT B. THOMPSON, CORPORATIONS AND OTHER BUSINESS ASSOCIATIONS 158 (3d ed. 1999))).

36. See id. at 589 (“Even sophisticated shareholders generally are not sufficiently informed to manage the firms they own because they own a relatively modest fraction of any one company and thus have neither the incentives to acquire nor the capacity to analyze the information needed to make prudent, everyday business decisions.”).

37. See id. at 588–90 & n.32 (“In addition to being specifically trained for the task and better informed about the firm, managers can make informed decisions free from the costly collective action problems that plague decentralized governance.”).

corporation’s capital, the business judgment rule insulates directors when they make what, in hindsight, turn out to be poor business decisions. 39 Nothing in the DGCL, however, gives directors license to breach the trust they have been given. When a rogue director breaches his or her fiduciary duty, the Court of Chancery has the authority to enforce shareholders’ rights. 40 And, in part because of the sheer number of corporations formed under Delaware law, the Court of Chancery’s docket often has been dominated by such breach of fiduciary duty cases.

With repeat experience comes ready expertise. As Chief Justice Rehnquist remarked at the celebration of its 200th anniversary, the Court of Chancery

has handed down thousands of opinions interpreting virtually every provision of Delaware’s corporate law statute. No other state court can make such a claim . . . . The process of decision in the litigated cases has so refined the law that business planners may usually order their affairs to avoid law suits. This recognition confers on the Court of Chancery one of the highest forms of praise the judiciary can receive. 41

As surely as Rome was built brick-by-brick, so too has Delaware developed its corporate jurisprudence case-by-case. Indeed, the volume of cases that it hears contributes importantly to this


40. See Williams v. Geier, 671 A.2d 1368, 1378 (Del. 1996) (“Only by demonstrating that the Board breached its fiduciary duties may the presumption of the business judgment rule be rebutted . . . .”); McMahon v. New Castle Assocs., 532 A.2d 601, 604 (Del. Ch. 1987) (describing fiduciary relationships as among “the most ancient headings under which Chancery’s jurisdiction falls”).

41. William H. Rehnquist, Chief Justice, United States Supreme Court, The Prominence of the Delaware Court of Chancery in the State-Federal Joint Venture of Providing Justice, Address at the Bicentennial Celebration of the Delaware Court of Chancery (Sept. 18, 1992), in William H. Rehnquist, The Prominence of the Delaware Court of Chancery in the State-Federal Joint Venture of Providing Justice, 48 BUS. LAW. 351, 354 (Nov. 1992); see also Savitt, supra note 21, at 585 (“[T]he five Chancery judges are appointed on the basis of their expertise in Delaware corporate law and cannot help but become even more expert by virtue of their deep and continuous exposure to that law and their obligation to interpret and expound it daily and at length.”).
valuable predictability, even in a dynamic economic and capital marketplace.42

To give just one small example of the organic nature of this process, in December 2011, Vice Chancellor Laster issued an opinion in In re Compellent Technologies.43 In dicta, he questioned the wording of a provision of a merger agreement requiring the target company’s board to give notice to the acquirer if any subsequent, superior offers arose.44 The Vice Chancellor did not question the general validity of this relatively common information rights provision, just the particular verbiage used to express it in the merger agreement at issue in that case.45

Less than two months later, another case—In re Micromet46—challenged a merger agreement containing a nearly identical provision, except for a revision in the language the court had questioned in Compellent.47 The court in Micromet found the revised provision unobjectionable.48 More important than the outcomes of those two cases, however, is what one reasonably can infer from their facts and sequence. Apparently, within a matter of weeks, transactional attorneys had read the Compellent opinion and advised their clients accordingly in connection with a later transaction that, when challenged, survived judicial scrutiny.49 This is a real world example of the predictability Chief

---

42. Regrettably, as expanded upon below, that does not always lead to the avoidance of lawsuits, especially in the M&A arena. See infra Part III.A.
44. Id. at *7 (“The [agreement] required the Board to knowingly breach its fiduciary duties . . . by first requiring the Board to determine that failing to act constituted a breach of its fiduciary obligations and then forbidding the Board to act until subsequent contractual conditions were met.”).
45. See id. (“This last problem could have been avoided by using a pure Superior Offer clause, rather than a hybrid with a Superior Offer trigger and a fiduciary duty determination.”).
47. See id. at *8 (describing the notice provision).
48. Id. at *9 (“[U]nlike Compellent, the recommendation provision here does not restrict the Board’s ability to fulfill known fiduciary duties in a timely fashion. Therefore, the potential problems identified in Compellent do not exist here.”).
Justice Rehnquist praised and the speed with which the Court of Chancery provides it.\footnote{See supra note 41 and accompanying text; see also Savitt, supra note 21, at 583–84 (“The [Court of Chancery’s] willingness to expedite review of any colorable claim of inadequate disclosure, coupled with the recent massive expansion in deal litigation, has produced a new phenomenon: the systematic real-time testing of merger proxies for material deficiencies by court-supervised plaintiffs’ attorneys.”).}

Similar examples of responsiveness can be seen in the Delaware legislature’s occasional passage of legislation to modify or clarify Delaware corporate or alternative entity law as reflected in a relatively recent decision of the Delaware Court of Chancery or the Delaware Supreme Court. One such instance relates to the corporation’s power to indemnify, and to advance legal fees incurred by, an individual as a result of legal proceedings related to that individual’s service as a director of the company.\footnote{See Del. Code Ann. tit. 8, § 145 (2011).} In the 2008 case of \textit{Schoon v. Troy Corp.}, Vice Chancellor Lamb allowed a corporation to amend its bylaws and revoke a former director’s right to advancement even though the bylaws that were in place during his service stated that the right to advancement would continue after his tenure on the board ended.\footnote{See id. at 1165–66 (ruling that a director’s advancement rights are determined by the relevant provisions in the corporation’s governing instruments as of the filing of the action triggering the advancement obligations, and not as of the dates that the director took office or the alleged wrongdoing occurred).}

In early 2009, the General Assembly amended DGCL Section 145(f) to abrogate the holding in \textit{Schoon} by precluding a
corporation from impairing a director’s right to indemnification or advancement after the occurrence of the conduct that is the subject of the indemnification or advancement claim, unless the corporation’s charter or bylaws in effect at the time of the challenged conduct expressly authorized retroactive elimination or impairment.54

Ironically, shareholder litigation also lends itself to the same collective-action governance problems that inhere in the substantive law of corporations it seeks to enforce.55 Two important ameliorating mechanisms are class actions on behalf of all affected shareholders and the derivative action, in which a shareholder sues not on her own behalf personally, but on behalf of the corporation.56 In either case, contingent fees provide incentives for specialized plaintiffs’ firms “to pursue monitoring activities that are wealth increasing for the collectivity (the corporation or the body of its shareholders).”57

---

55. As Chancellor Allen explained,

[a] fundamental condition of the corporate form when stockholders are widely dispersed . . . is that individual shareholders have little incentive to bear the costs associated with activities that monitor board of director (or management) performance . . . . While the conditions that allow investors to be rationally passive are a primary source of utility, they can also lead to inefficiency to the extent centralized management may have incentives that are not perfectly aligned with those of the residual owners of the firm, which is inevitably the case . . . . For that reason some expenditures for shareholder monitoring would be efficient. Such monitoring is, of course, more or less costly to the shareholder who engages in it. In a public company with widely distributed shares any particular shareholder has very little incentive to incur those costs himself in pursuit of a collective good, since unless there is some method to force a sharing of costs, he will bear all of the costs and only a (small) pro rata share of any gains that the monitoring yields.

56. See id. at 403 (explaining that the derivative lawsuit is “[o]ne way the corporation law deals with this conundrum . . . through the . . . recognized practice of awarding to successful shareholder champions and their attorney’s risk-adjusted reimbursement payments (i.e., contingency based attorneys fees”).
57. Id.
The Court of Chancery has broad jurisdiction over disputes concerning the internal affairs of Delaware corporations, and that includes hearing class and derivative actions regarding alleged breaches of fiduciary duties by a company's directors. Consequently, Delaware has a strong interest in ensuring that shareholder representative suits function properly and effectively. Among other things, this interest means that, unlike their Westminster antecedents, the Court of Chancery and the Delaware Supreme Court cannot blind themselves to the complex market dynamics that often engender and affect the shareholder litigation presented to them. This historical and thematic background leads to four specific examples, taken from cases decided within the last two years, of how the Court of Chancery advances the proper and effective functioning of shareholder representative litigation apace of market developments.

III. Recent Issues in Chancery Procedure

The foregoing is intended to introduce two propositions that prompted, and now frame, this Article: (1) the Delaware Court of Chancery's ability to respond rapidly and reliably to market dynamics; and (2) the Court's institutional interest in the proper and effective functioning of shareholder representative litigation. We next examine the juridical interaction between those two propositions, focusing on four specific procedural issues as

58. See Del. Code Ann. tit. 8, § 111 (2010) (expressly granting the Court of Chancery subject matter jurisdiction to interpret, apply, enforce, or determine a number of corporate instruments as well as the provisions of the DGCL itself, all of which is in addition to the Court's traditional equitable jurisdiction over claims of breach of fiduciary duty).

59. For example, Chancellor Chandler once articulated this interest as follows:

It is important for shareholders to bring derivative suits because these suits, filed after the alleged wrongdoing, operate as an ex post check on corporate behavior. . . . When shareholder plaintiffs bring meritorious lawsuits, they deter improper behavior by similarly situated directors and managers, who want to avoid the expense of being sued and the sometimes larger reputational expense of losing in court.

examples: (1) awards of attorneys’ fees to plaintiffs’ counsel; (2) certification and removal of lead plaintiffs; (3) motions to stay or dismiss because of concurrent litigation in another jurisdiction; and (4) the relationship between pleading rules, forum shopping, and statutory books and records actions under DGCL Section 220. These issues evidence collectively a broader theme: the Court’s development of procedural mechanisms that promote effective but fair use by shareholders of representative litigation. Those mechanisms should provide the optimal incentives for shareholders to avoid collective-action problems on the front-end without providing a windfall to plaintiffs’ firms on the back-end by enforcing rules that encourage and support the filing of meritorious suits. In turn, those rules should respect a plaintiff’s choice of forum while maintaining Delaware’s legitimate interest in regulating the internal affairs of entities created under its laws and without causing an arbitrary race to the courthouse.

A. Corporate Benefit on a Contingency Basis: Calculating Attorneys’ Fees

As previously indicated, U.S. capital markets appear to have generated increased litigation in recent years, especially

60. These four examples are not intended to exhaust the ways in which the court’s recent decisions may impact the effective and efficient operation of representative litigation in the corporate context. For example, in Forsythe v. ESC Fund Mgmt. Co., the court recently took a novel, efficiency-promoting approach in approving a proposed derivative action settlement. See Forsythe v. ESC Fund Mgmt. Co., C.A. No. 1091-VCL, 2012 WL 1655538, at *1 (Del. Ch. May 9, 2012). There, the court concluded that the settlement of $13.25 million in cash and other concessions fell “within a range of fairness, albeit at the low end,” but certain objectors asserted that they could recover significantly more if permitted to continue prosecuting the case to trial. Id. Because the proposed settlement was already fair, and because “[p]assing on the current settlement to seek more at trial carries substantial risk,” the court decided to “approve the settlement unless the objectors make the equivalent of a topping bid [by] . . . post[ing] a secured bond or letter of credit for the benefit of [the company on whose behalf the claims were brought] for the full settlement consideration of $13.25 million,” thus permitting the objectors to continue litigating the case for a greater recovery while protecting the company from the risk of recovering nothing after a full trial on the merits. Id.
shareholder representative suits challenging M&A transactions. Preliminary data for 2011 reveals that over 90% of M&A deals over $100 million were challenged in court, up from less than 40% only five years ago in 2007.61 Some evidence suggests that this rise in M&A litigation can be attributed to the Private Securities Litigation Reform Act62 and Securities Litigation Uniform Standards Act,63 both of which were intended to curb strike suits by entrepreneurial plaintiffs’ firms64 but inadvertently may have induced plaintiffs’ firms to bring cases predicated on state fiduciary duty law instead of federal securities law.65 Whatever the cause of this rise in M&A litigation, however, there is no evidence, as yet, that credibly proves that only one out of every ten boards that approved a merger in 2011 managed to do so without breaching their fiduciary duties.66 Unsurprisingly, therefore, most of these cases—approximately 70% of them—settle.67 Moreover, in almost 90% of the cases, the reported settlement consideration that the defendant directors gave in exchange for the release of the shareholders’ claims was a so-called therapeutic benefit, consisting of either supplemental

61. See supra notes 16–19 and accompanying text.
64. See Jennifer J. Johnson, Securities Class Actions in State Court, 80 U. CIN. L. REV. 349, 349–50 (2012) (“The PSLRA . . . was motivated by concerns that plaintiffs’ counsel were untrustworthy and filed frivolous lawsuits primarily for their own advantage . . . . [The] SLUSA . . . ensures that class counsel cannot avoid the PSLRA[] . . . by filing in state court.”).
65. See Brian Cheffins et al., Delaware Corporate Litigation and the Fragmentation of the Plaintiffs’ Bar, 2012 COLUM. BUS. L. REV. 427, 467 (“Various firms that were pushed out of securities class actions . . . responded by gravitating to corporate suits.”); Johnson, supra note 64, at 384 (“M&A objection class actions have replaced traditional stock drop cases as the lawsuit of choice for plaintiffs’ securities lawyers.”).
66. Indeed, “the rise of supermajorities of independent directors on boards, the decline in structural takeover defenses, and the improvement in disclosures [during this period] all suggest that litigation over M&A transactions should have become rarer, not more frequent.” Strine, Hamermesh & Jennejohn, supra note 14, at 16 n.46.
67. Cain & Davidoff, supra note 17, at 33 tbl.2.
disclosures or changes to the nonmonetary deal terms, but not an increase to the deal price offered to the shareholders.68

We next turn from the current state of the markets to the state of the law. Court of Chancery Rules 23(e)69 and 23.1(c),70 like their federal counterparts,71 require court approval before a class action or derivative suit may be dismissed voluntarily or compromised.72 These Rules are “intended to guard against surreptitious buy-outs of representative plaintiffs, leaving other class members without recourse” when the settlement, by design, releases their claims.73 In that approval process, the reviewing court performs at least three separate functions. It decides: (1) whether and how to certify the class; (2) whether, in the court’s independent business judgment, the settlement is in the best interest of the class; and (3) what would be a reasonable award of attorneys’ fees and expenses for the plaintiffs’ counsel.74 What follows addresses only the third of those tasks.

Where a class or derivative action settlement confers a benefit on the plaintiffs, either monetary or therapeutic, Delaware law authorizes an award of attorneys’ fees for counsel’s efforts in creating the benefit.75 “A court’s goal in setting a fee

---

68. Id. By contrast, from “1999 to 2000, 52% of suits filed on behalf of shareholders produced a financial benefit for the class, and only 10% of settlements were ‘disclosure-only.’” Strine, Hamermesh & Jennejohn, supra note 14, at 10.


70. Id. R. 23.1(c).

71. Fed. R. Civ. P. 23(e); id. R. 23.1(c).

72. See Del. Ct. Ch. R. 23(e) (“[A] class action shall not be dismissed or compromised without the approval of the Court . . . .”); id. R. 23.1(c) (same with regard to derivative actions); Fed. R. Civ. P. 23(e) (same); id. R. 23.1(c) (same).


75. Tandycrafts, Inc. v. Initio Partners, 562 A.2d 1162, 1164–65 (Del. 1989). This approach to assessing attorneys’ fees in representative shareholder actions proceeds from a framework of analysis identified in a case referred to as Sugarland and involves consideration of several different factors:

(i) the amount of time and effort applied to the case by counsel for the plaintiffs; (ii) the relative complexities of the litigation; (iii) the standing and ability of petitioning counsel; (iv) the contingent nature
award should be to avoid windfalls to counsel while encouraging future meritorious lawsuits,” but various problems emerge if the benefit is only therapeutic. First, pricing the value conferred by a nonmonetary settlement is an almost metaphysical exercise. What is the dollar value of furnishing more information about the deal decision-making process? Or of the increased chance that relaxing a deal protection term—e.g., reducing a termination fee from $35 million to $30 million—will result in a superior bid? Not the increased value of the topping bid itself (indeed, in many cases there is no topping bid), but rather the therapeutic value of the increased likelihood that, because of counsel’s efforts, another bidder might emerge? To answer these questions, the court draws on its experience.

In *In re Sauer-Danfoss*, decided in early 2011, the court cataloged the fees awarded in numerous disclosure-only cases, and sorted them by the qualitative importance of the disclosures obtained. Its results revealed

_of the litigation; (v) the stage at which the litigation ended; (vi) whether the plaintiff can rightly receive all the credit for the benefit conferred or only a portion thereof; and (vii) the size of the benefit conferred._

_In re Plains Res. Inc. S’holders Litig., C.A. No. 071-N, 2005 WL 332811, at *3 (Del. Ch. Feb. 4, 2005) (citing Sugarland Indus., Inc. v. Thomas, 420 A.2d 142, 149–50 (Del. 1980)). By far, the most important of the Sugarland factors is the benefit conferred by the settlement._

76. _In re Cox Radio, Inc. S’holders Litig., C.A. No. 4461-VCP, 2010 WL 1806616, at *20 (Del. Ch. May 5, 2010), aff’d, 9 A.3d 475 (Del. 2010) (unpublished table decision); but see Ams. Mining Corp. v. Theriault (Southern Peru II), 51 A.3d 1213, 1263 (Del. 2012) (Berger, J., dissenting) (“The trial court’s fee analysis focused on the perceived need to incentivize plaintiffs’ lawyers to take cases to trial. . . . The trial court opined that a declining percentage for ‘mega’ cases would not create a healthy incentive system. . . . That is not a decision based on Sugarland.”).


78. _See, e.g., id. (“In light of this problem, [the Delaware Court of Chancery] attempts to at least achieve consistency, looking at prior decisions to guide future ones.”)._  


80. _Id. at apps. A–C._
a range of discretionary awards with concentrations at certain levels. This Court has often awarded fees of approximately $400,000 to $500,000 for one or two meaningful disclosures, such as previously withheld projections or undisclosed conflicts faced by fiduciaries or their advisors. Disclosures of questionable quality have yielded much lower awards [in the range of $75,000 to $200,000]. Higher awards have been reserved for plaintiffs who obtained particularly significant or exceptional disclosures.81

This is not to say the Court of Chancery applies “a strict formula” in these types of cases.82 Rather, Sauer-Danfoss attempts to streamline the Court’s efforts to “promote[] fairness by treating like cases alike and rewarding similarly situated plaintiffs equally.”83 In that regard, Sauer-Danfoss exemplifies how valuable extensive serial cases involving similar issues really are: both the courts and litigants involved in the review of disclosure-only settlements now often begin with the three “levels” identified in Sauer-Danfoss and then ratchet up or down the ultimate award given (by the courts) or argued for (by the parties) depending on the quality of the disclosures obtained in the particular case. Thus, without aspiring to mathematical precision, taking advantage of the wealth of Chancery precedent can help “anchor this Court’s discretionary fee determinations to something more objective than the boldness of the plaintiffs’ ask and the vigor or passivity of the defendants’ response.”84

Finally, and more by way of a “reality check,” the reviewing court compares the size of the fee award to the number of hours the attorneys actually worked, derives an effective hourly rate based on the fee arrived at by the court, and compares that to the amount that would have resulted from the normal hourly rate of the plaintiffs’ counsel, sometimes referred to as the “lodestar.”85

81. Id. at *18, apps. A–C. Sauer-Danfoss is one of the only opinions the Court of Chancery has issued with appendixes of empirical analysis.


In determining an appropriate fee, the lodestar is of only secondary importance, at best. Still, the reason for considering it is the generally contingent nature of the plaintiffs’ counsel’s involvement—the need to provide appropriate incentives to take meritorious cases that recognize risk and avoid windfalls.

The fact that plaintiffs’ firms usually take their cases on a contingency basis presents another challenge in determining the appropriate amount of attorneys’ fees in settlement of representative shareholder litigation. In contingent fee cases, it serves the shareholders’ interest for the Court of Chancery to provide the right incentives to prosecute meritorious claims. Hence, when parties settle early, courts tend to award attorneys a lower percentage of the benefit conferred than in cases where settlement occurs later in the process, usually in the range of

---


86. See id. (“The time (i.e., hours) that counsel claim to have worked is of secondary importance.”); accord Southern Peru II, 51 A.3d 1213, 1254 (Del. 2012) (“Delaware courts are not required to award fees based on hourly rates that may not be commensurate with the value of the common fund created by the attorneys’ efforts.”).

87. As noted in one Chancery opinion,

fee awards should encourage future meritorious lawsuits by compensating the plaintiffs’ attorneys for their lost opportunity cost (typically their hourly rate), the risks associated with the litigation, and a premium. The fee award, however, can reach a point where it no longer operates as an incentive, and rather morphs into a “socially unwholesome windfall.”

Franklin Balance Sheet Inv. Fund v. Crowley, C.A. No. 888-VCP, 2007 WL 2495018, at *12 (Del. Ch. Aug. 30, 2007) (footnotes omitted) (quoting Seinfeld v. Coker, 847 A.2d 330, 334 (Del. Ch. 2000)). In this context, “risk” means “the risk of losing the case outright, something that every plaintiff must bear. Risk reflects the contingent nature of the work, the financing costs incurred with delaying the attorneys’ compensation until the case is concluded, the inability to diversify away particular risks, as well as other contingencies.” Seinfeld, 847 A.2d at 334 n.11.

88. In re Cox Radio, Inc. S’holders Litig., C.A. No. 4461-VCP, 2010 WL 1806616, at *20 (Del. Ch. May 6, 2010), aff’d, 9 A.3d 475 (Del. 2010) (unpublished table decision); cf. also Seinfeld, 847 A.2d at 333–34 (“It is equally important, however, for plaintiffs to prosecute these lawsuits efficiently. . . . Attorneys should not be encouraged to churn when they can receive a substantial premium in return for a successful result at an early stage of the litigation.”).
10%–15% of the settlement value for early settlements compared to as much as 35% of a post-trial judgment.\textsuperscript{89} Thus, if the benefit is big, so too is the proportional fee award, especially when the plaintiffs’ counsel have vigorously pursued the plaintiffs’ claim. Chancellor Strine recently made this unmistakably clear in \textit{In re Southern Peru Copper Corp.}\textsuperscript{90} In October 2011, he issued a post-trial opinion ruling that the shareholder class was entitled to recover damages, including prejudgment interest, of about $2 billion.\textsuperscript{91} In December 2011, he awarded plaintiffs’ counsel’s attorneys’ fees equal to 15% of the over $2 billion judgment, or over $300 million, equating to an effective rate in the range of $35,000 per hour.\textsuperscript{92} On appeal, the Delaware Supreme Court affirmed that fee, reiterating that the benefit achieved by the litigation is the “first and most important” factor for a court to consider in determining a discretionary fee award.\textsuperscript{93} The Delaware Supreme Court recognized that, although a trial court has discretion to decrease the applicable percentage as the size of the benefit increases, “the use of a declining percentage . . . is not required \textit{per se}.”\textsuperscript{94} Recent precedents like \textit{Southern Peru} make clear that there is enormous upside potential to zealous advocacy of shareholders’ rights—assuming, of course, that the plaintiffs’

\begin{footnotesize}
\footnotespace
\begin{enumerate}
\item [89.] \textit{In re Emerson Radio S’holder Derivative Litig.}, C.A. No. 3392–VCL, 2011 WL 1135006, at *3–4 (“Awarding increasing percentages helps offset representative counsel’s natural incentive to shirk.”).
\item [90.] \textit{In re S. Peru Copper Corp. S’holder Derivative Litig.}, 30 A.3d 60 (Del. Ch. 2011), aff’d, \textit{Southern Peru II}, 51 A.3d 1213.
\item [92.] \textit{Id.} (awarding $304,742,604.45 in attorneys’ fees). Plaintiffs’ counsel represented that they had worked a total of 8,597 hours on the case. Plaintiff’s Petition for Attorneys’ Fees and Expenses at 9, \textit{In re S. Peru Copper Corp.}, 2011 WL 6382006 (C.A. No. 961), 2011 WL 5240160. Based on those two figures, the implied hourly rate of the fee award is $35,477.55.
\item [93.] \textit{Southern Peru II}, 51 A.3d 1213, 1255 (Del. 2012).
\item [94.] \textit{Id.} at 1258; \textit{see also id.} at 1261 (“[W]e decline to impose either a cap or the mandatory use of any particular range of percentages for determining attorneys’ fees in megafund cases.”). Of course, to say that a declining balance “is not required \textit{per se}” in megafund cases still permits reasonable minds to disagree about whether a declining percentage is appropriate, or whether counsel rightly deserves a fee in such an amount, in any particular case.
\end{enumerate}
\end{footnotesize}
claim is meritorious and the harm caused by the defendants' misconduct is extreme, as occurred in *Southern Peru*.

What about the downside? Suppose a plaintiff challenges a deal, engages in some expedited discovery, settles early for a “low quality” supplemental disclosure, and then the plaintiff's attorneys are able to recoup their billable hours at normal rates. Such a system of unlimited upside potential with no downside risk would be unacceptable.95 It would provide incentives for an unwholesome, surreptitious kind of claim that is inimical to the Court of Chancery’s institutional interests in fostering the development and enforcement of a robust and effective body of corporation law.96 From a cynic’s perspective, that means “frequent-filers” who half-heartedly poke around in some expedited discovery, abandon their preliminary injunction motion, and grab quickly for the low-hanging fruit—e.g., a supplemental disclosure or two—still could recoup their time and hope to “hit it big” the next time.

While encouraging zealous advocacy of meritorious claims by fee awards such as in *Southern Peru*, the Court of Chancery also remains vigilant to avoid awarding improper windfalls or

---

95. See In re Revlon, Inc. S'holders Litig., 990 A.2d 940, 959 (Del. Ch. 2010) (“[A] systemic problem emerges when entrepreneurial litigators pursue a strategy of filing a large number of actions, investing relatively little time or energy in any single case, and settling the cases early to minimize case-specific investment and maximize net profit.”); Andrew J. Pincus, U.S. Chamber of Commerce, The Trial Lawyers' New Merger Tax: Corporate Mergers and the Mega Million-Dollar Litigation Toll on Our Economy 1 (2012) (characterizing awards of attorneys' fees for therapeutic-only settlements as “extortion through litigation” and a “litigation tax”).

96. See In re Topps Co. S'holders Litig., 924 A.2d 951, 961 n.38 (Del. Ch. 2007) (“Absent the rational sifting out of non-meritorious cases, stockholders suffer as the costs of litigation exact an undue toll on the procession of transactions valuable to stockholders and cause a harmful diminution in wealth-generating risk-taking by directors.”); Seinfeld v. Coker, 847 A.2d 330, 334 (Del. Ch. 2000) (noting that fee awards larger than necessary to incentivize the bringing of meritorious suits “serv[es] no other purpose than to siphon money away from stockholders and into the hands of their agents”); Strine, Hamermesh & Jennejohn, *supra* note 14, at 5 (“There is strong evidence of excess agency costs in the results of recent corporate representative litigation. Unless a consistent incentive system can be implemented that encourages representative litigation that benefits stockholders, the representative litigation system may on balance hurt investors more than it protects them.” (footnote omitted)).
rewarding defendants’ complacency, which have the potential to be equally problematic. Furthermore, given the volume of stockholder actions that the Court of Chancery adjudicates, it has the ability to speak as loudly in several less-publicized cases as it can in any one highly publicized case. For example, in contrast to *Southern Peru*, the transcript ruling approving a settlement and awarding attorneys’ fees in *In re Inspire Pharmaceuticals*97 attracted relatively modest attention. There, plaintiffs’ counsel requested $500,000 for (as determined by the reviewing court) a handful of questionable, agreed-to supplemental disclosures.98 The defendants, who would have to pay that fee, did not contest it,99 apparently believing, to quote Dickens, that the “issue is only a question of costs, a mere bud on the forest tree of the parent suit.”100 But that “bud” arguably should have gone to the shareholders as deal consideration. More broadly, the costs of expected, perhaps even inevitable,101 deal litigation against directors should not be shoehorned into a transaction intended to maximize shareholder value.102 Therefore, consistent with its


100. CHARLES DICKENS, BLEAK HOUSE 6 (Oxford Univ. Press 1966) (1853).

101. For a discussion of this inevitability, see supra notes 67–68 and accompanying text.

102. Compare Revlon, Inc. v. MacAndrews & Forbes Holdings, Inc., 506 A.2d 173, 183–84 (Del. 1986) (holding corporate directors’ fiduciary duties when negotiating a change-of-control transaction require them to attempt in good faith to obtain the best sales price reasonably available for their shareholders), with *Strine, Hamermesh & Jennejohn, supra* note 14, at 20 (“[I]f the defendants believe that the transaction is good for stockholders, as they should if they are directors recommending the transaction, then settling cases that are an obstacle to the transaction in a cost-effective way is a proper fiduciary act.”). In any event, regardless of whether the intentional pricing of attorneys’ fees to settle
responsibilities, the Inspire Pharmaceuticals court independently examined the therapeutic benefit achieved in light of the numerous cases in Delaware evaluating supplemental disclosures and reduced the fee from the $500,000 requested to $300,000.

Inspire Pharmaceuticals, where requested attorneys’ fees were reduced, is not unusual. Rather, it reflects a particular instance of the procedure the Court of Chancery consistently employs. Aware of the risks of windfalls or complacent litigants, the Court also is mindful of another significant risk of any settlement of a representative action: the possibility that valuable claims will be released for materially less than they are worth. Accordingly, the Court of Chancery vigilantly examines the merits of every settlement of a representative action, whether objected to or not, to determine whether, in the exercise of the reviewing court’s independent business judgment, the settlement is in the shareholders’ interest. The court also reviews carefully perceived strike suits into the overall transaction price is consistent with directors’ fiduciary duties, the fact remains that “the corporations involved in the merger, and therefore their investors, bear those costs. Thus, those costs contribute to an increase in the costs of capital, and a net negative societal impact.” Strine, Hamermesh & Jennejohn, supra note 14, at 19.

103. See supra note 74 and accompanying text (discussing the reviewing court’s function in the approval of settlements of class action and derivative suits).

104. In re Inspire Pharms. Inc. S’holders Litig., No. 6378-VCP, 2012 WL 275115, at ¶ 12 (Del. Ch. Jan. 30, 2012). Even then, however, the effective hourly rate was about $650. McEvilly Affidavit, supra note 99, at ¶ 47 (swearing that the plaintiffs’ counsel “have expended more than 465 hours prosecuting the Action and incurred approximately $12,694.55 in expenses”).


106. See Alon Harel & Alex Stein, Auctioning for Loyalty: Selection and Monitoring of Class Counsel, 22 YALE L. & POL’Y REV. 69, 71 (2004) (“The class attorney’s egoistic incentive is to maximize his or her fees—awarded by the court if the action succeeds—with a minimized time-and-effort investment. This objective does not align with a both zealous and time-consuming prosecution of the class action, aimed at maximizing the amount of recovery . . . .”).

107. See Barkan v. Amsted Indus., Inc., 567 A.2d 1279, 1283 (Del. 1989) (“The Court of Chancery plays a special role when asked to approve the settlement of a class or derivative action. It must balance the policy preference for settlement against the need to insure that the interests of the class have been fairly represented.”).
the related applications for attorneys’ fees. If plaintiffs’ counsel are overcompensated for settlements in which the claimed therapeutic benefits are relatively meager, that might increase the risk that meritorious claims will be settled too cheaply.\textsuperscript{108} Cases like \textit{Inspire Pharmaceuticals} exemplify the Court of Chancery’s continuing efforts to achieve the appropriate balance of incentives.

Contrary to views expressed in recent articles in even the popular press,\textsuperscript{109} the argument that plaintiffs’ firms face no downside risk when pursuing a shareholder representative action on a contingency is overstated. Not every transactional case is put on a fast track. To receive expedited discovery and hearings, plaintiffs must “justify imposing on the defendants and the public the extra (and sometimes substantial) costs” of expedition.\textsuperscript{110} Increasingly, defendants are opposing motions to expedite, a fact reflected in leading treatises and empirical data alike.\textsuperscript{111} Although Delaware courts “traditionally ha[ve] acted with a certain solicitude for plaintiffs” challenging a fast-moving deal

\begin{footnotesize}
\begin{enumerate}[\textsuperscript{108}.]
\item See Forsythe v. ESC Fund Mgmt. Co., C.A. No. 1091-VCL, 2012 WL 1655538, at *5 (Del. Ch. May 9, 2012) (“Because of these incentives, counsel may favor (consciously or not) a bird-in-the-hand settlement over the continuing and costly quest for an uncertain outcome.”).
\item See, e.g., WOLFE & PITTENGER, \textit{supra} note 16, § 10.07[a] (observing that motions to expedite “ha[ve] become an early and sometimes dispositive battleground on the merits of the [preliminary injunction] application itself”); see also infra note 113 and accompanying text (discussing limited empirical review conducted by the authors).
\end{enumerate}
\end{footnotesize}
and, therefore, prefer to “err[] on the side of more hearings rather than fewer,” plaintiffs still must carry the burden of articulating a colorable claim and demonstrating that they are likely to suffer irreparable injury if expedited treatment is not granted.\textsuperscript{112}

A significant number of motions to expedite have been denied in recent years.\textsuperscript{113} Plaintiffs unable to articulate a claim that need be merely “colorable” to survive a motion to expedite later tend to face a serious risk of being unable to plead allegations entitling the plaintiffs to relief under any “conceivable” set of facts—i.e., to survive a motion to dismiss.\textsuperscript{114} Furthermore, plaintiffs effectively lose the ability to settle their claims in exchange for therapeutic changes to the deal terms or supplemental disclosures if the challenged transaction closes before the plaintiffs can take and receive discovery and brief a preliminary injunction motion within the timeframe contemplated by the Court’s usual, nonexpedited rules. In cases where a motion to expedite has been denied, therefore, it may be more difficult for the plaintiffs to achieve a meaningful settlement and reap the attendant premium attorneys’ fees. Indeed, the plaintiffs may not obtain any attorneys’ fees at all. Stated differently, cases that proceed with expedited discovery and briefing toward a preliminary

\textsuperscript{112}. \textit{Giammargo}, 1994 WL 672698, at *2.

\textsuperscript{113}. For example, in an admittedly small sample of all class action or derivative complaints in some way related to M&A transactions filed in the Court of Chancery between January 1, 2011 and June 30, 2012, the authors observed that the plaintiffs filed motions to expedite approximately 50% of the time. Of those motions, defendants challenged approximately 28% (i.e., there were 27 challenges of 98 total motions to expedite). While several motions subsequently were withdrawn by the parties or mooted by further case developments, the Court denied at least 14 motions to expedite, one of which was uncontested, during the observation period. That is, excluding the outlier uncontested motion just mentioned, almost half of the contested motions to expedite were denied (i.e., 13 / 27 = 0.481).

\textsuperscript{114}. \textit{See} Cent. Mortg. Co. v. Morgan Stanley Mortg. Capital Holdings LLC, 27 A.3d 531, 536 (Del. 2011) (observing that “[t]he pleading standards governing the motion to dismiss stage of a proceeding in Delaware . . . are minimal,” and instructing trial courts to “deny the motion unless the plaintiff could not recover under any reasonably conceivable set of circumstances susceptible of proof”). \textit{See also}, e.g., \textit{In re Alloy, Inc. S’holder Litig.}, C.A. No. 5626-VCP, 2011 WL 4863716, at *1, *5 (Del. Ch. Oct. 13, 2011) (granting motion to dismiss in a case where the court earlier had denied a motion to expedite for failure to articulate a colorable disclosure claim).
injunction hearing often are those that already have survived at least one, albeit not especially rigorous, substantive challenge. Hence, the motion to expedite mechanism also operates so as to subdivide shareholder representative cases into at least three general categories: (1) the wholly meritless, for which plaintiffs’ counsel probably will receive no fee at all; (2) the colorable, for which counsel may receive some compensation in the form of an ordinary, hourly rate for raising issues that are readily remedied and that, perhaps with the benefit of discovery, will unearth a more substantive claim; and (3) the clearly meritorious, for which counsel have every incentive to prosecute zealously to trial, if necessary. In this way, the Court of Chancery’s procedural rules help screen out pure “strike suits” but still encourage plaintiffs’ firms to satisfy the watchdog function on which the enforcement of corporation law often depends.

115. In arguing that the possibility of denial at the motion to expedite stage represents downside risk for plaintiffs’ counsel, we are mindful that motions to expedite generally are filed relatively early in the litigation. Therefore, counsel might not yet have invested significant time or other resources in the case, and would have little at risk. Nevertheless, we also appreciate that “more than half of all merger lawsuits settled in 2010 and 2011 were resolved within two months of the filing date” and “[o]nly 15% of lawsuits lasted longer than 100 days before settlement.” Pincus, supra note 95, at 4 (emphasis omitted) (citing Cornerstone Research, supra note 19, at 3). Furthermore, Chancery judges generally are “willing to award substantial attorneys’ fees, even after a relatively quick settlement of the case, [so] that our fee awards are not structured to reward lawyers for needlessly prolonging litigation.” Seinfeld v. Coker, 847 A.2d 330, 333 (Del. Ch. 2000). In this fast-paced environment, therefore, the relatively early stage of the motion to expedite ruling is less relevant than one reasonably might suppose to estimating the investment plaintiffs’ counsel has made in the case.

116. Some scholars have argued that the Court of Chancery’s scrutiny of the fees requested by plaintiffs’ counsel contributes, at least in part, to the growing trend of cases being filed by plaintiffs in courts outside of Delaware. See, e.g., John Armour, Bernard Black & Brian Cheffins, Delaware’s Balancing Act, 87 Ind. L.J. 1345, 1370–72 (2012). For example, although Delaware might be an attractive forum in which to bring a truly meritorious case, the Court’s ability to identify weak cases relatively early arguably encourages plaintiffs’ firms “to bring weak suits in which the only real goal is a quick settlement for fees elsewhere.” Strine, Hamermesh & Jennejohn, supra note 14, at 18 n.49. Hence, the Court’s “continuing efforts to achieve the appropriate balance of incentives” as to any procedural mechanism, Armour, Black & Cheffins, supra note 116, at 1372, necessarily is a multifaceted endeavor in that, for example, “simply cutting fees, denying motions to expedite, and motions to dismiss claims are

The certification of a class and the occasional disqualification of lead plaintiffs as the class representatives is another subject that has received recent attention. Over the last several decades, the identity of lead plaintiffs in shareholder representative actions—and, indeed, the shareholder profile of corporate America\textsuperscript{117}\textsuperscript{117}—has shifted away from individuals to institutions. One anecdotal example is that the Delaware cases reported from the 1980s and 1990s frequently were captioned “Kahn,” “Weinberger,” or another individual’s last name versus the defendant company. Today (assuming the case does not receive a consolidated “In re” caption because of multiple complaints), the caption is more likely to begin with CalPERS, LAMPERS, or some other institutional investor’s name. Constitutional due process requires that lead plaintiffs adequately represent the interests of the shareholder class that they purport to represent.\textsuperscript{118}\textsuperscript{118} In shareholder representative actions, this requirement takes on a what’s-good-for-the-goose-is-good-for-the-gander quality. That is, by filing a complaint alleging breach of fiduciary duty against the directors of a Delaware corporation,


\textsuperscript{118} Phillips Petroleum Co. v. Shutts, 472 U.S. 797, 812 (1985) (“[T]he Due Process Clause . . . requires that the named plaintiff at all times adequately represent the interests of the absent class members.’’); see also \textit{In re MCA, Inc. S’holders Litig.}, 785 A.2d 625, 635 (Del. 2001); Prezant v. De Angelis, 636 A.2d 915, 923 (Del. 1994); Leon N. Weiner & Assocs. v. Krapf, 584 A.2d 1220, 1225 (Del. 1991); Geller v. Tabas, 462 A.2d 1078, 1083 (Del. 1983).
the stockholder-plaintiff “voluntarily assumes the role of fiduciary for the class.”

The shift from individual to institutional shareholders, in turn, has increased the number of institutions assuming fiduciary duties on behalf of similarly situated shareholders. Because decentralized decision-making can be difficult, it may be easier for an individual to comply with fiduciary duties than for an institution. An institution acts only through its human agents. For the institution as a whole to act, or refrain from acting, in a certain way, each person in the decision-making network must participate, increasing the potential for breakdown in the lines of communication and honest mistakes. Even an honest mistake, however, can cause the institution to act in ways that may constitute inadequate representation. Accordingly, where adequacy of representation is challenged, deciding whether an institutional lead plaintiff, in fact, has breached its duties to the class requires in-depth analysis. The following two cases, issued within a few months of each other, make this point.

The first case, Steinhardt v. Howard-Anderson,120 illustrates how straightforward the analysis can be where the lead plaintiffs are individuals. Steinhardt, among other plaintiffs, brought a class action on behalf of the shareholders of Occam Networks, Inc., challenging, and moving to enjoin, Occam’s then-pending acquisition by Calix, Inc. In pursuing a preliminary injunction, “Steinhardt admittedly had been receiving regular written and oral updates about the litigation”121 based on “[i]nformation gleaned from Confidential Discovery Material and counsel’s [nonpublic] litigation assessments.”122 The court later enjoined the transaction for two weeks so shareholders could consider corrective disclosures.123 With the benefit of that nonpublic information, which he received as a fiduciary for the class, Steinhardt was able “to conclude that the Merger would be

120. Id.
121. Id. at *5.
122. Id. at *9.
123. Id. at *6.
consummated with no change in price.”124 To benefit himself individually,

Steinhardt sold Calix short as a way to exit his Occam position. . . . He intended to (and later did) use the shares of Calix stock he would receive when the Merger closed to cover his short sales, even though Steinhardt and his co-plaintiffs were asking the Court to enjoin the closing of the Merger.125

Thus, Steinhardt “both liquidate[d] his Occam position and [took] advantage of the arbitrage spread that existed between Calix and Occam at that time.”126

In Delaware, as elsewhere, use of confidential information by an agent for personal gain constitutes a breach of the fiduciary duty of loyalty.127 Not surprisingly, therefore, the Court of Chancery granted the defendants’ motion for sanctions against Steinhardt, stating:

Consistent with prior rulings by this Court when confronted with representative plaintiffs who have traded while serving in a fiduciary capacity, Steinhardt and the funds [he managed] are dismissed from the case with prejudice, barred from receiving any recovery from the litigation, required to self-report to the Securities and Exchange Commission, directed to disclose their improper trading in any future application to serve as lead plaintiff, and ordered to disgorge profits in the amount of $534,071.45.128

Although that statement could be read to mean that trading by a lead plaintiff is a per se breach of duty, the next case we discuss interpreted Steinhardt to be more dependent on its facts.

124. Id. at *10.
125. Id. at *4.
126. Id. (internal quotation omitted).
127. See RESTATEMENT (THIRD) OF AGENCY § 8.05(2) (2006) (“An agent has a duty not to use or communicate confidential information of the principal for the agent’s own purposes or those of a third party.”).
128. Steinhardt v. Howard-Anderson, C.A. No. 5878-VCL, 2012 WL 29340, at *1 (Del. Ch. Jan. 6, 2012). The court ordered that the disgorged profits be distributed to the class, either immediately or “held in escrow . . . and allocated as part of any future remedy.” Id. at *15. The court, however, expressly declined “to award the disgorged funds to Calix as a proxy for the class” because Calix’s shareholder base includes investors who never held Occam stock before the challenged merger. Id.
In *In re Celera Corp. Shareholders Litigation*, the lead plaintiff, the New Orleans Employees’ Retirement System (NOERS), challenged Quest Diagnostics’ takeover of Celera. On the eve of the preliminary injunction hearing, NOERS settled for supplemental disclosures and modifications to the merger agreement’s deal protection provisions. As with all representative litigation, that settlement was conditioned on confirmatory discovery—if NOERS discovered evidence before the final settlement hearing that suggested its claims were stronger than NOERS had realized, it could rescind the settlement and litigate post-closing damages. Importantly, because of how the deal was structured, there ultimately was a four-day period during which Quest’s acquisition of Celera had become an absolute certainty, yet Celera’s stock continued to trade on the secondary market at around $8.05, five cents more than the merger price. “[A]fter all material information regarding the lawsuit, settlement, and transaction were disclosed to the marketplace,” NOERS’s investment advisor saw “a risk-free arbitrage opportunity” and sold all of its Celera shares at $8.05 before the transaction officially closed. For that and other reasons, however, Celera’s largest shareholder objected to the

---

130. *Id.* at *1.
131. *Id.* at *6–7.
132. *Id.* at *7.
133. *Id.* Specifically, the deal “was structured in two tiers of, first, a tender offer by Quest for any and all Celera shares at $8.00 per share and, second, a back-end squeeze-out merger at the same price . . . .” *Id.* at *1. Moreover, Celera had granted Quest a so-called top-up option, effectively permitting the back-end squeeze-out merger to close quickly and without a shareholder vote pursuant to DGCL Section 253. *Id.; see also* Del. Code Ann. tit. 8, § 253 (2010) (allowing mergers without a shareholder vote if the other merging entity owns at least 90% of the corporation’s outstanding stock entitled to vote). The four-day window mentioned above reflected the time between when Quest publicly announced the success of the front-end tender offer and its intent to exercise the top-up option and when it had taken all the necessary formalities to exercise the top-up option and effect a short-form merger under Section 253. *In re Celera*, 2012 WL 1020471, at *1. For more on top-up options, see Olson v. EV3, Inc., C.A. No. 5583–VCL, 2011 WL 704409, at *1–3 (Del. Ch. Feb. 21, 2011).
settlement, claiming that NOERS no longer had standing to recover monetary damages, therefore lacked any incentive to perform adequate confirmatory discovery, and essentially had abandoned the class in breach of its fiduciary duties. Moreover, between the settlement hearing and the date the opinion was issued, the objector advised the court of no less than three other cases involving similar issues where institutional investors had served as lead plaintiffs.

The Celera court drew on a wealth of precedent—including cases from the 2000s, 1990s, and even the 1930s—in identifying why certain conduct is or is not consistent with a lead plaintiff's fiduciary duties to the class. After discussing those precedents, the court concluded that NOERS's decision to sell all of its Celera stock arguably was “careless and cavalier” because of “the significant waste of the Court’s and the parties' resources caused by th[e] unnecessary side issue.” Nonetheless, the court concluded that NOERS’s decision was neither analogous to Steinhardt nor indicative of insider trading or obvious disloyalty: “notwithstanding its questionable conduct, NOERS still satisfies, if only barely, the requirement for an appropriate class representative,” the court wrote. On that basis, the court certified the class with NOERS as the lead plaintiff, a finding that the Supreme Court upheld on appeal. In its opinion in Celera, however, the Court of Chancery underscored that

the frequency with which Delaware courts have had to address the conduct of lead plaintiffs in recent months is troubling.

---

135. Id. at *1.
136. See id. at *15 (noting that the objecting shareholder cited three recently-decided cases to support its arguments).
137. See id. at *9–16 (citing a variety of cases in analyzing adequacy and typicality under Court of Chancery Rule 23(a)).
138. Id. at *1.
139. Id. at *16; see also id. at *15–16 (distinguishing the three cases cited by the objector).
140. Id. at *20 (certifying the class with NOERS as the lead plaintiff).
141. In re Celera Corp. S'hlder Litig., No. 212, 2012 WL 6707736, at *7-10 (Del. Dec. 27, 2012). The Supreme Court reversed, however, on a narrower issue, ruling that individual class members deserved the right to opt out of the settlement if they wished because of “due process concerns” stemming from, among other things, NOERS's bare adequacy to represent the class. Id. at *13.
When a class representative purports to object on behalf of itself and all others similarly situated only to decide later that the objected-to conduct may not have been all that bad, that representative is prone to appear more concerned about its own interests than those of the class. That appearance undermines the trust shareholders place in lead plaintiffs and, in turn, effaces courts' confidence in the adequacy of the representation that a lead plaintiff is capable of providing. . . . Lead plaintiffs must remain committed to fulfilling their obligations to those they represent throughout the litigation. Among other things, that should include thinking about more than the technical permissibility of their conduct, but also how their conduct is likely to be perceived. Here, NOERS engendered a host of legitimate criticisms to its commitment to this case by choosing to take advantage of a “risk-free arbitrage” opportunity. Technically permissible or not, that choice failed to reflect an appropriate level of regard and respect for NOERS’s position as a fiduciary for the class.142

Suffice it to say that Delaware courts expect representative or lead plaintiffs to take seriously their fiduciary responsibilities to the class of shareholders or company (in a derivative action) they represent and to exercise appropriate care to meet those obligations. Those same requirements apply equally to institutional investors that undertake to represent a class or company.

C. Hercules and the Hydra: The Multi-Jurisdictional Litigation Problem

Growth in deal litigation has led in turn to an increase in the number of jurisdictions where any particular transaction is challenged. For example, over half of the shareholder class actions against Delaware corporations in 2010 were filed in both Delaware and at least one other jurisdiction.143 By comparison, multi-jurisdictional litigation occurred in only 8.6% of litigated

---


143. Johnson, supra note 64, at 39.
M&A transactions in 2005. As even prominent plaintiffs’ attorneys have commented:

An all-too-familiar pattern emerges following the announcement of the acquisition of a Delaware corporation. Some of the plaintiffs’ bar will race to file a class-action lawsuit in the Delaware Court of Chancery. Other plaintiffs’ lawyers will race to a courthouse in the state of the target’s headquarters. A third set of firms will file “tag-along” actions in federal court asserting the same state law claims as the other plaintiffs, but incorporating a proxy claim under Section 14A of the Securities Exchange Act as a basis for the federal court’s jurisdiction. The result is parallel litigation that wastes judicial resources, burdens defendants, and, most importantly, threatens shareholders.

By forcing director-defendants to combat essentially the same adversary with many heads—as Hercules did with the Hydra—multi-jurisdictional shareholder litigation imposes additional burdens on, and amplifies various risks for, all parties involved. For example, “defense counsel are acutely aware [of] which plaintiffs’ counsel are ‘pilgrims’ (i.e., early and easy settlers).”

Permitting defendants to negotiate global releases of all the pending cases with the most supine plaintiffs’ firm harms shareholders. Likewise, “[s]hareholders are also hurt when a plaintiffs’ firm in one of the competing jurisdictions seeks to accelerate the procedural posture of its case by pursuing a temporary restraining order or some other form of extra-expedited relief that is not really appropriate for the facts of the

144. Cain & Davidoff, supra note 17, at 13–14.
146. See MARK P. O. MORFORD & ROBERT J. LENARDON, CLASSICAL MYTHOLOGY 567–68 (8th ed. 2007) (noting that for each head Hercules removed, the Hydra grew two more).
147. Lebovitch, Silk & Friedman, supra note 145, at 2.
148. See supra notes 67–74 and accompanying text (discussing the risks inherent in releasing claims in settlement).
case.”149 Perhaps most problematic is the possibility that two or more jurists in competing jurisdictions will refuse to stay their respective cases in favor of litigation pending in another forum, subjecting defendants to simultaneous litigation in multiple fora, wasting judicial resources, and risking inconsistent findings of fact or interpretations of law.150 Indeed, this very risk materialized in recent years when then-Vice Chancellor Strine and New York State Supreme Court Justice Cahn proceeded forward on parallel tracks with separate actions challenging the same going-private transaction involving the Topps Company.151

With those risks in mind, various members of the Court of Chancery and the corporate bar have proposed several solutions with varying success. One approach relies on traditional comity between sister state courts. Where the authoritative development of Delaware law is at stake in disputes involving entities that Delaware itself has created, Delaware arguably has the strongest

149. Lebovitch, Silk & Friedman, supra note 145, at 2; see also Strine, Hamermesh & Jennejohn, supra note 14, at 89 (noting the “perverse incentives” in multi-forum litigation for “a plaintiffs’ law firm to ‘plant a flag’ in a distinct forum[,] affording ... and risk equal outcomes ... would then leave the law in a confused state and pose full faith and credit problems for all involved”).

150. See In re Allion Healthcare Inc. S’holders Litig., C.A. No. 5022-CC, 2011 WL 1135016, at *4 (Del. Ch. Mar. 29, 2011) (noting “the possibility that two judges would apply the law differently or otherwise reach different outcomes ... would then leave the law in a confused state and pose full faith and credit problems for all involved”).

interest in having its courts articulate those legal rules.\textsuperscript{152} Consistent with the strength of that interest, Delaware provides expedited, direct appeals to its Supreme Court.\textsuperscript{153} In addition, although Delaware courts, for comity reasons, generally “will defer to a first-filed action in another forum and will stay Delaware litigation pending adjudication of the same or similar issues in the competing forum,”\textsuperscript{154} the Court of Chancery has held that a formulaic, “first-filed” rule should “not apply with full force” in the case of competing stockholder representative actions dispersed throughout multiple jurisdictions,\textsuperscript{155} because “less deference [is due] to the speedy plaintiff’s choice of forum.”\textsuperscript{156} Rather, “the chartering state[s] . . . powerful interest in ensuring the uniform interpretation and enforcement of its corporation law” warrants greater consideration.\textsuperscript{157} Consequently, Delaware courts treat competing shareholder representative actions filed closely together in time as contemporaneously filed, allowing a more deliberative analysis of which of the competing courts is best positioned to adjudicate the dispute.\textsuperscript{158} That analysis takes

\textsuperscript{152} See VantagePoint Venture Partners 1996 v. Examen, Inc., 871 A.2d 1108, 1112–13 (Del. 2005) (“[I]n order to prevent corporations from being subjected to inconsistent legal standards, the authority to regulate a corporation’s internal affairs should not rest with multiple jurisdictions.”); In re Topps Co., 924 A.2d at 954–57 (“In a situation like this one, when this court is clearly an efficient and convenient forum prepared to issue a timely ruling, public policy and comity indicate that this state’s courts should answer the question . . . .”); Strine, Hamermesh & Jennejohn, supra note 14, at 85 (arguing that the most important interest affected by shareholder representative suits “is the interest of the corporation’s equity investors,” who, as an otherwise dispersed group united by “the state of incorporation whose law they have chosen to govern their investment”); see also Edgar v. MITE Corp., 457 U.S. 624, 645–46 (1982) (holding no state has a legitimate interest “in regulating the internal affairs of foreign corporations”).

\textsuperscript{153} DEL. CONST. art. 4, § 11, cl. 4.

\textsuperscript{154} WOLFE & PITTENGER, supra note 16, § 5.01 (citing McWane Cast Iron Pipe Corp. v. McDowell-Wellman Eng’g Co., 263 A.2d 281, 283 (Del. 1970)).

\textsuperscript{155} Biondi v. Scrushy, 820 A.2d 1148, 1150 (Del. Ch. 2003).

\textsuperscript{156} Ryan v. Gifford, 918 A.2d 341, 349 (Del. Ch. 2007).

\textsuperscript{157} In re Topps Co. S’holders Litig., 924 A.2d 951, 960 (Del. Ch. 2007).

\textsuperscript{158} See In re Citigroup Inc. S’holder Derivative Litig., 964 A.2d 106, 116–17 (Del. Ch. 2009) (treating actions filed “within the same general time frame” in New York and Delaware as “contemporaneously filed”).
place under a general *forum non conveniens* rubric. At the same time, the Court of Chancery has deferred to sister states where the particular circumstances of a case indicate that that state’s interests are greater than Delaware’s, but rarely will that occur on matters concerning the application of Delaware corporation law.

Another proposed solution is the so-called “One Forum Motion,” where relatively early in the litigation, defendants’

---

159. See *Wolfe & Pittenger*, supra note 16, § 5.01 (“In addition to the comity considerations articulated in *McWane*, the Court of Chancery, in assessing motions to stay Delaware litigation under the first-filed rule, frequently analyzed the same ‘practicality’ factors traditionally applied under the *forum non conveniens* doctrine.”).

160. See, e.g., *TA Instruments-Waters*, LLC v. Univ. of Conn., 31 A.3d 1204, 1210 (Del. Ch. 2011) (“The motion to expedite is denied out of respect for the superior interests of a sister state.”); Transcript of Record at 23–27, *Walton v. Apollo Global Mgmt.*, LLC (Del. Ch. Jan. 29, 2010) (C.A. No. 5216-VCS) (staying action involving a publicly traded Delaware LLC in favor of adjudication in Ohio because the LLC Agreement expressly provided for the application of Ohio fiduciary duty law); *see also* *Diedenhofen-Lennartz v. Diedenhofen*, 931 A.2d 439, 451 (Del. Ch. 2007) (noting that because of Delaware’s strong interest in adjudicating its own corporation law, “Delaware has a related and equally important interest in affording comity to the courts of other jurisdictions when a dispute arises under foreign business law.” (footnote omitted)).

161. That did occur, however, in March and April 2008 when shareholders of Bear Stearns commenced independent actions in Delaware and New York challenging the investment bank’s eleventh-hour merger with JPMorgan, a transaction that had been brokered by the U.S. Treasury and the Federal Reserve to quarantine a severe shock to the national, and even international, financial system. *See generally In re Bear Stearns Cos., Inc. S’holder Litig. (Bear Stearns (Del.)), C.A. No. 3643-VCP, 2008 WL 959992 (Del. Ch. Apr. 9, 2008); In re Bear Stearns Litig. (Bear Stearns (N.Y.)), 870 N.Y.S.2d 709 (N.Y. Sup. Ct. 2008).* Because the “unusual set of facts [were] unlikely to recur,” because the New York cases were first filed and the New York court was attending diligently to them, and because the “extraordinary circumstances” required a rapid response from a competent court free from distractions, the Court of Chancery stayed the Delaware actions in favor of adjudication in New York. *Bear Stearns (Del.),* 2008 WL 959992, at *8. (Ultimately, the New York plaintiffs withdrew their motion for a preliminary injunction and sought only post-closing damages. *See Bear Stearns (N.Y.),* 870 N.Y.S.2d at 724 (reciting the procedural history of the case).) Exigencies such as these, however, are exceptional.

162. The One Forum Motion is referred to in some circumstances as a “Savitt motion” after its apparent inventor, William D. Savitt of Wachtell Lipton Rosen & Katz. *See, e.g.*, *Nierenberg v. CKx*, Inc., C.A. Nos. 5545–CC, 6519–CC, 6624–CC, 2011 WL 2185614, at *1 (Del. Ch. May 27, 2011) (referring to one forum motion as a Savitt motion although neither Savitt nor his firm filed the
counsel file concurrently in all jurisdictions a single motion requesting that the judges communicate directly and confer among themselves about which of the parallel actions should proceed and which should be stayed. It appears that in Delaware the Chancellor and four Vice Chancellors are receptive to such motions, in that these judges “have endorsed the approach and agreed to reach out to their judicial counterpart in the parallel litigation.” The effectiveness of such motions, however, remains somewhat uncertain. Under the doctrine surrounding injunctions in aid of jurisdiction, for example, the Court of Chancery “is empowered to enjoin a party . . . from removing the subject of the controversy to a foreign jurisdiction by filing a later action or proceeding in a foreign forum.” In contrast, some have suggested that, in concurrently filed, multi-jurisdiction cases, each court “lack[s] the power to compel action in the [foreign] litigation absent cooperation from the [foreign] judge”; or, as put more forcefully by some commentators, “they are unnecessary where a foreign court is predisposed to defer to

163. See generally Flinn & McCormick, supra note 162.
164. Id. at 3. Indeed, Chancellor Chandler, who retired from the bench in 2011, indicated that one forum motions were his “personal preferred approach” to address multi-fora deal litigation. Allion Healthcare Inc. S'holders Litig., C.A. No. 5022-CC, 2011 WL 1135016, at *4 n.12 (Del. Ch. Mar. 29, 2011).
166. Flinn & McCormick, supra note 162, at 3. In at least one case, for example, the court expressly “grant[ed] Defendants' motion [to proceed in one jurisdiction] to the extent of confirming that this consolidated action in Delaware shall proceed,” but with the caveat that, “[a]s to whether the related actions before [the foreign judge] should be dismissed or stayed, I express no opinion.” Letter from Donald F. Parsons, Jr., Vice Chancellor, Delaware Court of Chancery, to P. Bradford deLeeuw, Esq., et al., In re Converge, Inc. S'holders Litig., C.A. No. 7368-VCP (Del. Ch. Apr. 13, 2012) (on file with the Washington and Lee Law Review).
the forum whose law is at stake and useless where a judge is determined to keep the case.” 167 Hence, like comity, this solution ultimately may be more aspirational than actual. Nevertheless, the relatively recent invention of the One Forum Motion itself illustrates the virtually real-time engagement between the corporate bar and the Delaware courts in responding to the various dynamics affecting shareholder representative litigation specifically and the law of corporations more generally.

At least one other prospective solution might be more effective, however. Recently, Delaware companies have been amending their certificates of incorporation and corporate bylaws to include “exclusive forum selection” clauses that would require shareholders to litigate so-called “internal affairs” claims in Delaware. 168 There is nonbinding—and arguably conflicting—precedent regarding the enforceability of these forum selection clauses, 169 “and not all courts enforce forum choices even in

169. Compare In re Revlon, Inc. S’holders Litig., 990 A.2d 940, 960 & n.8 (Del. Ch. 2010) (suggesting, in dictum, that a forum selection clause in a corporate charter or bylaw, because both are construed as contracts, could be enforceable), with Galaviz v. Berg, 763 F. Supp. 2d 1170, 1174–75 (N.D. Cal. 2011) (holding, as a matter of federal procedure, that a forum selection clause contained only in a bylaw (as opposed to a corporate charter) and adopted unilaterally by the board of directors (as opposed to with approval from shareholders) is unenforceable). The outcome of the Galaviz decision also might be limited by its facts: an alleged breach of fiduciary duty already had occurred when the board adopted the forum selection bylaw. Galaviz, 763 F. Supp. 2d at 1172. In refusing to enforce the bylaw, the court emphasized that “[p]articularly where, as here, the bylaw was adopted by the very individuals who are named
bilateral contracts.”170 In February 2012, shareholders brought suit in the Court of Chancery to challenge head-on the legality of such charter provisions and bylaws under Delaware law.171 The cases, currently assigned to the Chancellor, have not yet been decided, but are being followed closely by scholars and the practicing bar alike.172

as defendants, and after the alleged wrongdoing took place, there is no element of mutual consent173 necessary to sustain the forum selection clause as a freely negotiated contractual provision. Id. at 1171. It is not clear whether the same result necessarily follows if the alleged misconduct occurs after the bylaw’s adoption.


171. Frank Aquila & Anna Kripitz, Forum-Selection Provisions in Delaware, CORP. COUNS. (Aug. 27, 2012), http://www.law.com/corporatecounsel/PubArticleCC.jsp?id=1202568858164&ForumSelection_Provisions_in_Delaware (last visited Feb. 6, 2013) (on file with the Washington and Lee Law Review). More specifically, the class actions challenged the validity of forum selection bylaws unilaterally adopted by the boards of 12 Delaware companies. While 10 of the complaints have since been dismissed following the repeal of the bylaws in question, the defendants in the mooted cases were unable to stay plaintiffs’ fee petitions, which may be an indication of how the court will rule on the issue of validity in the two cases (against Chevron Corporation and FedEx Corporation) that are still proceeding.

Id.


As of October 2012, “at least 200 companies [had] designate[d] Delaware in their governing documents as the exclusive venue for intra-company disputes.” Aquila & Kripitz, supra note 168, at 50. Only twelve public companies, however, did so by submitting a proposed charter or bylaw amendment to shareholders for a vote; the remainder implemented the provision either before an initial public offering (IPO) or as a board-approved bylaw amendment. Id. at 51. Moreover, both proxy advisory firms and activist shareholders have opposed forum selection provisions in governing instruments. Id. Therefore, outside of the IPO context, companies still might face implementation hurdles as a practical matter regardless of whether the courts uphold exclusive forum selection provisions as a legal matter.
D. 220 or Not 220: Whether 'Tis Nobler to Risk Outrageous Fortune and File Fast or to Take Up the Tools at Hand

Thus far, we have addressed issues that typically arise in “direct actions.” For example, when corporate directors breach their fiduciary duty in the context of a merger, the breach generally harms shareholders directly by divesting them of their stock ownership. For that reason, the shareholders themselves have a claim that they can bring directly against the board and other corporate fiduciaries. Other breaches of duty, however, harm the corporation itself. These breaches harm the shareholders only derivatively as a function of the shareholders’ pro rata investment in the corporation. Examples include excessive or wasteful compensation or a fiduciary’s usurpation of a corporate opportunity. In such cases, the corporation itself possesses, and is entitled to enforce, the claim against the wrongdoers. Normally, the decision whether to initiate a lawsuit on the corporation’s behalf is vested in the company’s board of directors. But when the alleged wrongdoers are the very same directors who must make that decision, potential conflicts of interest necessarily arise. For potential conflicts of this kind, so-called “derivative actions” provide a remedy.

173. As with any general rule, there are, of course, exceptions: [The test for whether a direct cause of action exists] is a two-element test: (1) who was injured and (2) who should recover . . . . [M]entioning a merger does not talismanically create a direct action. Instead, the court must look to all the facts of the complaint and determine for itself whether a direct claim exists. Dieterich v. Harrer, 857 A.2d 1017, 1027 (Del. Ch. 2004).


176. See Schoon v. Smith, 953 A.2d 196, 208 n.45 (Del. 2008) (“Generaly a cause of action belonging to a corporation can be asserted only by the corporation.”).


178. See Schoon, 953 A.2d at 208 (noting the purpose of the derivative action
Since at least the early 1900s, the law of corporations has permitted shareholders to obtain judicial permission to litigate the rights of the corporation, but only if the shareholders first attempt to persuade the board to redress the wrong or are able to show the court why such an attempt would be futile. By empowering a shareholder to enforce the company’s claim against its directors, derivative actions upset the normal balance of rights and obligations between stockholders and directors that are fundamental to the law of corporations. Consequently, the importance of the shareholder’s pre-suit demand requirement is hard to overstate. Rather than impose an ordinary “notice pleading” standard, Court of Chancery Rule 23.1 requires plaintiffs in derivative actions to plead with particularity that either: (1) a majority of the board is so interested or so lacks independence, or (2) the challenged transaction is so beyond the bounds of reasonable business judgment, that the court can infer the directors simply are incapable of deciding objectively whether to pursue or otherwise redress the corporation’s claims.

Pleading a claim with particularity based solely on publicly available information, however, can be difficult. An important tool to make that task more manageable is DGCL Section 220.

is to overcome conflicts of interest and, thus, injustices to the corporation).

179. WOLFE & PITTENGER, supra note 16, § 9.02[b][3][ii]; see also Baker v. Bankers’ Mortg. Co., 129 A. 775, 776 (Del. Ch. 1925) (“A stockholder may under certain circumstances sue in equity in right of the corporation. Where he does so, however, he asserts the right of the corporation. This he is not permitted to do, unless the circumstances be such that the corporation will not sue.”).


181. See Aronson v. Lewis, 473 A.2d 805, 811–12 (Del. 1984) (“[T]he demand requirement . . . exists [to give effect to] the fundamental precept that directors manage the business and affairs of corporations.”), overruled on other grounds by Brehm v. Eisner, 746 A.2d 244 (Del. 2000).

182. DEL. CT. CH. R. 23.1(c); Brehm, 746 A.2d at 253–54; Rales v. Blasband, 634 A.2d 927, 930 (Del. 1993); Aronson, 473 A.2d at 814.

183. See Kenneth B. Davis, Jr., The Forgotten Derivative Suit, 61 VAND. L. REV. 387, 399–400 (2008) (“[P]rospective litigants encounter a practical obstacle to mustering sufficient facts to satisfy Aronson’s ‘particularity’ requirement . . . . One Delaware judge has noted the almost impossible burden that results when the relevant facts are not a matter of public knowledge.” (internal quotation marks omitted)).
That statute permits a stockholder to seek court-ordered inspection of the corporation’s books and records if the stockholder has a “proper purpose . . . reasonably related to such person’s interests as a stockholder.” Investigating corporate malfeasance to satisfy the pre-suit demand requirement qualifies as such a purpose, provided the shareholder can present a credible basis for suspecting wrongdoing. As the Delaware Supreme Court recently reiterated, “Delaware courts have strongly encouraged stockholder-plaintiffs to utilize Section 220 before filing a derivative action . . . .” Numerous other cases similarly exhort prospective derivative plaintiffs to employ the “tools at hand,” referring to a books and records action pursuant to DGCL Section 220, before commencing a plenary action.

184. Del. Code Ann. tit. 8, § 220(b) (2010); see also id. § 220(c)(3) (affording rights of inspection by court order upon showing a proper purpose). An example of relevant “books and records” that a shareholder might seek would be the internal investigation report that a board considered before terminating an officer without cause and with a lucrative severance.

185. See King v. VeriFone Holdings, Inc. (King I), 994 A.2d 354 (Del. Ch. 2010) (“To cite one example, investigating corporate mismanagement . . . is a proper purpose for seeking a Section 220 books and records inspection.”), rev’d on other grounds, 12 A.3d 1140, 1145 (Del. 2011).


187. See Beam v. Stewart, 845 A.2d 1040, 1056–57 (Del. 2004) (noting that the plaintiff had failed to plead sufficient facts to “support a reasonable doubt of independence” and asserting that “had [plaintiff] first brought a Section 220 action seeking inspection of . . . books and records she might have uncovered facts that would have created a reasonable doubt”); White v. Panic, 783 A.2d 543, 556–57 (Del. 2001) (“[T]his case demonstrates the salutary effects of a rule encouraging plaintiffs to conduct a thorough investigation, using the ‘tools at hand’ including the use of actions under [DGCL] § 220 for books and records, before filing a complaint.”); Brehm v. Eisner, 746 A.2d 244, 266–67 (Del. 2000) (“Plaintiffs may well have the ‘tools at hand’ to develop the necessary facts for pleading purposes. For example, plaintiffs may seek relevant books and records of the corporation under Section 220 . . . .”); Sec. First Corp. v. U.S. Die Casting & Dev. Co., 687 A.2d 563, 567 & n.3 (Del. 1997) (noting that DGCL Section 220’s use to gather information is encouraged); Rales v. Blasband, 634 A.2d 927, 934 & n.10 (Del. 1993) (“[A] stockholder who has met the procedural requirements and has shown a specific proper purpose may use the summary procedure embodied in [DGCL Section 220] to investigate the possibility of corporate
because representative litigation—including derivative actions—cannot “optimally serve investors unless suits are actually filed on the basis of a real concern that wrongdoing has occurred and after a proper investigation.”

But where, as has become common today, multiple plaintiffs file multiple complaints in multiple jurisdictions, plaintiffs and their firms risk losing their chance to participate in the leadership structure prosecuting the derivative suit if they dutifully pursue a Section 220 action. To avoid that risk, many plaintiffs engage in a race to the courthouse and file a complaint fast on the heels of a public announcement of some suspicious event without undertaking any prior investigation. Given most jurisdictions’ liberal rules of civil procedure favoring dismissal with leave to amend, competition for litigation arguably wrongdoings.

188. King I, 994 A.2d at 356.
189. As one relatively recent case articulated the dilemma for plaintiffs’ attorneys,

[under a first-to-file system, plaintiffs’ lawyers cannot act as stockholders collectively would want because by proceeding deliberately, a law firm risks losing control of the case to competitors who file immediately. For fast-filing lawyers, the resulting action has the dynamics of a lottery ticket. In most cases, the fast-filing plaintiff will not have pled a derivative claim that can overcome Rule 23.1. But in the rare case, fate may bless the fast-filer with something implicating the board, or a court might be offended by the magnitude of the corporate trauma and allow the derivative action to proceed.

Pyott, 46 A.3d at 346.
190. See, e.g., Fed. R. Civ. P. 15(a)(2) (“The court should freely give leave [to amend a pleading] when justice so requires.”); D.C. Super. Ct. R. Civ. P. 15(a) (“[L]eave shall be freely given when justice so requires.”). Notably, Delaware is not such a jurisdiction, at least not in the Court of Chancery. Court of Chancery Rule 15(aaa) provides, in pertinent part, that dismissal under Rule 12(b)(6) or 23.1 “shall be with prejudice (and in the case of complaints brought pursuant to Rules 23 or 23.1 with prejudice to the named plaintiffs only) unless the court, for good cause shown, shall find that dismissal with prejudice would not be just under all the circumstances.” Del. Ct. Ch. R. 15(aaa) (emphasis added). Thus, in Chancery, “when a plaintiff is confronted with a motion to dismiss under Rules 12(b)(6) or 23.1, he or she must either ‘seek leave to amend [the] complaint or stand on [the] complaint and answer the motion to dismiss,’” but there typically will be no post-dismissal opportunity to buttress and refile a deficient pleading. Zucker v. Andreessen, C.A. No. 6014-VCP, 2012 WL 2366448, at *2 (Del. Ch. June 21, 2012) (alteration in original) (quoting MCG Capital Corp. v. Maginn,
transforms plaintiffs’ difficult task of pleading with particularity into what some defendants view as a form of Sisyphean punishment. According to legend, the Greek gods punished King Sisyphus for his hubris by forcing him to roll a massive stone up a steep hill, only to have it roll back down again as he reached the summit, thus consigning the once-proud king to an eternity of useless effort and unending frustration. 191 Today, a plaintiff can: (1) haphazardly cobble together a complaint that fails to plead demand excusal with particularity; (2) quickly file that complaint in a jurisdiction with liberal pleading rules in an effort to ensure “lead plaintiff” status; (3) perhaps suffer dismissal under Rule 23.1, but with leave to amend; (4) then pursue a Section 220 action in Delaware; and (5) only thereafter, refile a properly researched complaint. In this environment, one could argue, defendants face the wasteful exercise of litigating repeatedly the issue of demand excusal, much like King Sisyphus and his uphill struggle. 192

Chancellor Strine addressed this problem in King v. VeriFone. 193 Sympathetic to Sisyphean frustration, he wrote, “the intended end of the derivative lawsuit is not furthering the interests of fast-filing plaintiffs or their lawyers; rather, the derivative suit is one of several tools that stockholders may use to further the corporation’s best interests.” 194 To “encourage the kind of litigation conduct that is more likely to benefit investors as a class,” 195 he attempted to establish a bright-line rule: once a stockholder files a derivative action complaint, that stockholder is barred from pursuing a related books and record action thereafter. 196 That rule effectively would have forced plaintiffs to

C.A. No. 4521-CC, 2010 WL 1782271, at *5 (Del. Ch. May 5, 2010)).
191. MORFORD & LENARDON, supra note 146, at 354.
194. Id. at 362 (emphasis added).
195. Id.
196. See id. at 356–57 ("[S]tockholders who seek books and records in order to determine whether to bring a derivative suit should do so before filing. . . .


exhaust diligently the “tools at hand” and prepare the strongest possible complaint before filing it.\textsuperscript{197} The Delaware Supreme Court reversed.\textsuperscript{198} While recognizing the potential abuse of Section 220 and explicitly continuing to encourage plaintiffs to employ it early, the Delaware Supreme Court held nevertheless that nothing in the statutory text supported the bright-line rule employed by the Chancellor.\textsuperscript{199} Rather, the decisional law made clear that investigation of corporate misconduct is a proper purpose for bringing a Section 220 action so long as the plenary court has granted the shareholder-plaintiff leave to amend the complaint.\textsuperscript{200}

The applicability of this last example to the thesis of this Article—i.e., that caseload volume involving similar issues leads to refined doctrine, valuable predictability, and measured balance—may not be immediately apparent. Nevertheless, \textit{King} is a salient example of the unique benefits case volume provides in at least two respects. First, case volume enabled the Court of Chancery and the Delaware Supreme Court alike to recognize a problem in the first instance and address it head-on. Indeed, in

\footnote{Once a plaintiff has chosen to file a derivative suit, it has chosen its course . . . .\textsuperscript{\textregistered}}.

\textsuperscript{197.} See \textit{id.} at 356–57 (discussing the policies supported by a bright-line rule).

\textsuperscript{198.} \textit{King II}, 12 A.3d 1140, 1151 (Del. 2011).

\textsuperscript{199.} As the Supreme Court put it,

\begin{quote}
[\textit{a}l\textit{l}though we reject the result reached by the Court of Chancery, and the bright-line rule that drove it, we are sensitive to the policy concerns that animated both. . . . If relief under Section 220 is to be restricted in the manner adjudicated by the Court of Chancery, any such restriction should be imposed expressly by the General Assembly, not decreed by judicial common law decision-making.]
\end{quote}

\textit{Id.} at 1150–51.

\textsuperscript{200.} See \textit{id.} at 1149–50 ("[L]ong-standing Delaware precedent . . . recognizes that it is a proper purpose under Section 220 to inspect books and records that would aid the plaintiff . . . where the earlier-filed plenary complaint was dismissed on demand futility-related grounds without prejudice and with leave to amend."); \textit{accord} Amalgamated Bank v. NetApp, Inc., C.A. No. 6772-VCG, 2012 WL 379908, at *7 (Del. Ch. Feb. 6, 2012) ("\textit{King} stands for the limited proposition that when a plaintiff has been granted leave to amend its complaint a plaintiff may have a proper purpose for demanding such records. When that leave to amend no longer exists, a plaintiff’s proper purpose is extinguished.” (footnote omitted)).
King, the two decisions diverge only as to the solution. In the Delaware Supreme Court’s view, the Chancellor’s proposal required legislative, rather than judicial, action. It is questionable whether jurists in other jurisdictions who only infrequently confront motions to dismiss for failure to make pre-suit demand could have grasped the tension as readily as the Chancellor and the Justices who grappled with how to solve it in King. The relatively small number of jurists involved (ten) and the fact that they all routinely encounter business entity issues of this nature helps to foster the nuanced and well-informed character of Delaware’s decisional law.

Second, case volume affords lower courts successive opportunities to remedy perceived problems in the law to the satisfaction of their appellate courts. The recent South v. Baker decision provides a useful example. After a series of accidents in 2011 at silver mines operated by Hecla Mining Company (Hecla), the United States Mine Safety and Health Administration ordered the company to close one of its mines for

201. King II, 12 A.3d at 1151.

202. Indeed, King itself was a second attempt to address the fast-filing problem. In two earlier cases, the Court of Chancery “declin[ed] to infer ‘the existence of other facts [not explicitly alleged in the complaint] that would have been proved or disproven by a further pre-suit investigation.’” Beam v. Stewart, 833 A.2d 961, 981 n.66, 982 (Del. Ch. 2003) (quoting White v. Panic, 793 A.2d 356, 364 (Del. Ch. 2000)), aff’d on other grounds, 845 A.2d 1040 (Del. 2004). That is, the Beam and White courts essentially proposed that, if a plaintiff has not pursued a books and records action, courts should draw fewer inferences in the plaintiff’s favor from the allegations actually made. See id. at 970 (“[T]he Court accepts as true all well-pled factual allegations contained in the amended complaint, but conclusory statements—those unsupported by well-pled factual allegations—are not accepted as true.”); White v. Panic, 793 A.2d 356, 363 (Del. Ch. 2000) (noting that courts need not make inferences with no support in fact). The Delaware Supreme Court reversed White and affirmed Beam on alternate grounds because

[a] plaintiff’s use of, or failure to use, a books and records inspection does not change the standard to be applied to review of the complaint. Regardless of whether the plaintiff secured any facts alleged in her complaint through a Section 220 inspection, the court must draw all reasonable inferences in the plaintiff’s favor and determine whether those facts create a reasonable doubt of the directors’ independence.

Beam v. Stewart, 845 A.2d 1040, 1057 n.52 (Del. 2004).

safety violations. Consequently, Hecla issued a press release announcing that it was lowering its estimated silver production for 2012. On the heels of that announcement, shareholders brought a so-called Caremark claim. That is, the plaintiffs accused Hecla’s directors of breaching their fiduciary duty by failing to monitor or heed proverbial “red flags” indicating misconduct within the company that, but for the directors’ failure of oversight, could have enabled the company to take remedial measures and avoid legal liability or other corporate losses. The plaintiffs, however, did not avail themselves of Section 220 before filing their complaint. The court ultimately determined that the complaint lacked the particularized allegations necessary to plead demand futility under Court of Chancery Rule 23. Accordingly, the court dismissed the case.

In so doing, the court expressly held that the putative lead plaintiffs provided inadequate representation for failing to utilize Section 220. Specifically, the court observed that, “in a representative action, a trial court has an independent and continuing duty to scrutinize the representative plaintiff to see if

204. Id. at *3–6.
205. Id. at *1.
207. South, 2012 WL 4372538, at *1; see also In re Caremark, 698 A.2d at 971 (“[A] sustained or systematic failure of the board to exercise oversight—such as an utter failure to attempt to assure a reasonable information and reporting system exists—will establish the lack of good faith that is a necessary condition to liability.”).
209. Id. at *7. Although the South court could have dismissed the complaint under Rule 23.1 alone, it felt compelled to engage in this extra analysis because (1) “good faith disagreements exist about whether other stockholders of the corporation are in privity with the named plaintiff such that the with-prejudice dismissal has preclusive effect on other derivative actions,” but (2) even the precedents “that give preclusive effect to a Rule 23.1 dismissal universally recognize that another stockholder still can sue if the first plaintiff provided inadequate representation.” Id. Thus, the court’s judicial finding that the Souths provided inadequate representation preserves the ability of “other stockholders to bring suit, including those stockholders who have attempted to use Section 220.” Id. at *14 (emphasis added).
she is providing adequate representation,”211 and that (as the Delaware Supreme Court suggested in King v. Verifone) “[o]ne possible remedy for a prematurely-filed derivative action might be for the plenary court to deny the plaintiff lead plaintiff status in such a situation.”212 Employing that suggestion, the South court ruled that a plaintiff who hastily files Caremark claims without first conducting an adequate investigation aided by a corporate books and records inspection under Section 220 behaves contrary to the interests of the corporation but consistent with the desires of the filing law firm to gain control of (or a role in) the litigation. The natural and logical inference from this recurring scenario is that the plaintiff is serving the interests of the law firm, rather than those of the corporation on whose behalf the plaintiff ostensibly seeks to litigate.213 That is, the “recurring scenario supports a presumption that the plaintiff has acted disloyally and is not an adequate fiduciary for the corporation.”214

In both developing and applying this presumption of inadequate representation for not utilizing Section 220, the court was careful to note two important qualifications. First, the court explicitly differentiated a Caremark claim—for which “the connection to the board is neither readily apparent nor reasonably inferable from the occurrence of the corporate trauma” and “there usually will not be any need to rush when filing . . . [because t]he claim typically seeks to obtain damages” for a discrete harm that already has occurred—“from other types of derivative actions in which a plaintiff challenges a specific and identifiable board decision.”215

211. Id. at *15.
212. Id. at *13 (quoting King II, 12 A.3d 1140, 1151 (Del. 2011)).
213. Id. at *17.
214. Id.
215. Id. Unlike a Caremark claim, a plaintiff challenging specific and identifiable board decisions may well be able to plead particularized allegations without using Section 220 that are sufficient to survive a Rule 23.1 motion to dismiss, for example by pleading that a majority of the directors were not independent and disinterested (as when directors vote on their own compensation) or that the decision was not entitled to the protections of the business judgment rule (as when a transaction
Second, the court stressed that, under Delaware Rule of Evidence 301, presumptions are rebuttable, meaning that a plaintiff can show that, in fact, he or she did conduct a reasonable investigation or that there was a need to act with celerity in the circumstances of the particular case, i.e., “that acting speedily benefited the corporation and not just the plaintiffs’ law firm.”

Despite these express qualifications, the presumption of inadequate representation very well might undermine the incentive to race to file in the first instance in future cases—assuming, of course, the Delaware Supreme Court condones its use.

Another solution, proposed by Chancellor Strine in February 2012, is when determining which complaint was first-filed for purposes of consolidation or appointment of lead counsel, the relevant filing date should relate back to the earlier Section 220 action, if one was filed. That approach would put the plaintiff’s counsel who first pursue a statutory books and records action before filing their complaint on an equal footing with the “fast filers,” thereby undercutting the incentive to file quickly. Courts that are asked to appoint lead plaintiffs and counsel from several competing parties and firms generally will consider, among other things, the quality of the pleadings as an indicator of the type of representation the putative lead plaintiffs and their counsel likely will provide.

Although the Chancellor’s “relation back” approach is an intriguing notion, it may face its own multi-jurisdictional issues. That is, even if the determination of lead counsel in Delaware requires relating back to the filing date of the Section 220 action, would judges in other jurisdictions meet the onerous standard for waste).

Id. (footnote omitted).

216. Id. (citing Del. R. Evid. 301).

217. See Transcript of Record at 34–36, La. Mun. Police Empls.’ Ret. Sys. v. Page (Del. Ch. Feb. 6, 2012) (C.A. No. 7041-CS) (“Perhaps we should start thinking about a diligent books and records demand about wrongdoing as having somewhat of the same dignity as a first-filed complaint but done by people who are doing it the right way.”).

confronting whom to designate as lead counsel in representative litigation in their own courts be persuaded to do the same?

Lastly, there is the possibility that someone will embrace the Supreme Court’s suggestion in King. If nothing in the statutory text of Section 220 supports the Chancellor’s rebuffed bright-line rule from King, the Delaware Corporate Law Council could urge the Delaware General Assembly to amend the statute. As alluded to earlier, in a small state like Delaware, amending a statute is considerably easier than one might think. The corporate bar regularly engages with both the courts and the legislature because all of them have an important shared interest in maintaining a General Corporation Law that is appropriately balanced, fair, and effective for each legal issue that may arise in contemporary and dynamic capital markets.

IV. Conclusion

The overarching theme of this Article has been that Delaware’s volume of corporate and alternative business entity cases, the fact that those cases are litigated before the relatively small, but expert, Delaware Court of Chancery and Supreme Court, and the responsiveness of its courts, its legislature, and the marketplace generally accelerate the development of refined doctrine, measured balance, and valuable predictability. This Article, of course, is hardly the first to recognize or to describe that theme, but the matters reviewed here demonstrate that

219. See, e.g., Savitt, supra note 21, 585–86 (describing the Court of Chancery “as an unusual and remarkably effective regulatory machine” in part because of the “vast amount of M&A activity” it sees); Aquila & Kripitz, supra note 168, at 50 (“Delaware is known as the premier forum for resolving . . . internal [corporate] disputes . . . .”); Jill E. Fisch, The Peculiar Role of the Delaware Courts in the Competition for Corporate Charters, 68 U. Cin. L. Rev. 1061, 1080–81 (2000) (“Delaware courts aggressively adopt and modify corporate law doctrine, exhibiting a degree of activism that more closely resembles the legislative process[,] . . . driven by policy considerations, including an effort to respond on an ongoing basis to developments in the business world.”); Rehnquist, supra note 41, at 354; Bernard S. Black, Is Corporate Law Trivial?: A Political and Economic Analysis, 84 Nw. U. L. Rev. 542, 588–89 (1990) (arguing that Delaware’s prominence as a state of incorporation is due to the expertise of its judiciary).
the Court of Chancery’s unique characteristics facilitate its ability to respond to challenging new issues of substantive and procedural law alike. Indeed, institutional success for the Court of Chancery, because of its de facto specialization in a substantive field that lends itself to collective-action problems, depends on refined doctrine and measured balance as to the procedural mechanisms that are central to the development of Delaware commercial law. This overarching theme also demonstrates the relatively common phenomenon in Delaware that at any one time multiple disputes pertaining to some particular area of corporation law will be before the Court at once. In 2011 and 2012, one of the areas that came more directly under the judicial microscope encompassed important aspects of representative shareholder actions. Although not all of the issues in that area have achieved the level of predictability, efficiency, or stasis to which the law aspires, the special nature of the Court of Chancery and its docket enabled it to accomplish quite a lot in a short time span. The result is greater clarity regarding important aspects of representative litigation and, we hope, more effective representation of shareholders’ interests in the future. If recent history provides an accurate barometer, many of the legal issues we have addressed will be almost passé by the time this Article is published because they will have been replaced by clusters of new cases presenting various manifestations of the different issues du jour for 2013 and beyond. Whatever the future issues may be, we are confident that, for the reasons discussed in this Article, the Delaware courts and the Court of Chancery, in particular, together with the Delaware legislature and bar, will remain adept in meeting the inevitable challenges and innovations the marketplace brings before them.
Risk-Based Student Loans

Michael Simkovic*

Table of Contents

I. Introduction ................................................................. 529

II. Government Support for Higher Education ................. 531
   A. Higher Education as an Investment in Human Capital .............................................. 532
   B. The Demand for Skilled Labor and Social Mobility .............................................. 548
   C. Values-Based Arguments for Higher Education Funding ........................................... 551
   D. Education Funding and Independent Research .................................................... 553
   E. Higher Education Funding Options: Student Debt or General Taxes ...................... 555

III. U.S. Federal Student Loan Programs ......................... 560
   A. Student Eligibility Criteria Are Generally Not Risk-Based ..................................... 561
   B. Only Exceptionally Poorly Performing Institutions Are Excluded .......................... 562
   C. Borrowing Limits Depend on Grade Level and Dependent Status ............................ 564
   D. Federal Student Loan Pricing is Statutory, Not Risk-Based .................................... 565

---

* Associate Professor of Law, Seton Hall University School of Law. Thanks go to Professors Peter Arcidiacono, Douglas G. Baird, Dan Black, Jean Braucher, David G. Carlson, Jonathan Glater, Rachel Godsil, Thomas H. Jackson, Stephen J. Lubben, Frank McIntyre, John Morley, Chrystin Ondersma, Rafael Pardo, Frank Pasquale, Mark Poirier, John A.E. Pottow, Robert K. Rasmussen, and Richard Squire for helpful comments and suggestions. Thanks also to Michael Amalfe, Cynthia Furmanek, William Pugh, Timothy Malacrida, and Arkell Weygandt for research assistance, and to Luke Spreen and Steve Hipple at the U.S. Department of Labor, Bureau of Labor Statistics for assistance with access to data.
IV. Higher Education and Labor Market Coordination

Problems ........................................................................................................................................ 567

A. Students as Customers Create Pressure to Reduce Academic Standards ........................................ 567

B. Educational Institutions May Have Incentives to Funnel Students into Areas That Do Not Maximize Students’ Future Incomes or Employment Prospects ....................................................................................... 570

C. Many Students May Not Understand the Connection Between Their Chosen Course of Study and Future Income and Employment Prospects .................................................................................................. 583

D. Skilled Labor Markets Suffer from Periodic Booms and Busts ...................................................... 587

E. Skills Mismatches Reduce Employment and Output ......................................................................... 588

V. The Theory of Risk-Based Credit Pricing ...................................................................................... 589

A. A Simple Risk-Based Credit Pricing Model ...................................................................................... 594

B. Data-Driven Risk-Based Pricing ...................................................................................................... 596

C. Risk-Based Pricing Is Not Necessarily the Same as Market-Based Pricing ...................................... 597

D. Government’s Incentives Are Uniquely Well Aligned with Students’ ................................................ 599

E. Forecasting Change Is Both Necessary and Perilous ..................................................................... 602

F. Private Lenders Could Be Enlisted to Correct Forecast Errors ...................................................... 606

G. Limited Dischargeability of Student Loans in Bankruptcy Reduces Lender Incentives to Monitor Students and Risk-Adjust Student Loan Pricing ...................................................................................... 607

H. Income-Based Repayment Plans .................................................................................................. 617

VI. Ethical Considerations and the Limits of Risk-Based Pricing ..................................................... 620

A. Factors That Predict Student Loan Defaults .................................................................................. 620

B. Preserving Equal Opportunity, Social Mobility, and Individual Choice ........................................... 623

C. Risk-Based Pricing and Choice of Major ....................................................................................... 625
I. Introduction

Credit markets serve a vital function in capitalist economies: evaluating the riskiness of a range of possible investments and channeling resources toward those investments that investors believe are most likely to prove successful. This process is known as the “risk-based pricing” of credit. Ideally, risk-based pricing should lead to lower cost of capital for lower risk investment choices with larger rewards, and therefore more investment in such promising activities. Conversely, risk-based pricing should lead to higher costs of capital—and therefore less investment—in high-risk activities with relatively low rewards. If creditors are well informed and analytic, and borrowers respond to financial incentives, then risk-based pricing—compared to uniform credit pricing—leads to a more efficient allocation of society’s limited resources.

Although risk-based pricing is standard in business-loan markets, and may be increasingly common in consumer-credit markets such as mortgages and credit cards, risk-based pricing is seldom used in the market for student loans. Most student loans are extended under Federal Student Loan programs administered by the Department of Education. These federal programs have historically offered loans at rates lower than those offered by most private lenders, on terms that are more attractive to student borrowers, and without adjusting the pricing on loans.

---

according to the risks inherent in different courses of study or lending to different types of borrowers.

The Federal Student Loan programs—first established in the mid-twentieth century to increase the supply of skilled labor, promote economic and technological development, and provide upward socioeconomic mobility—are broadly successful. They have provided low-cost credit to millions of students, helped increase educational attainment, held administrative costs lower than those of the private sector, and generated a profit for the federal government.

However, Federal Student Loan programs have not incorporated many recent insights from financial, developmental, and labor economics that distinguish between different types of education. Because of this, Federal Student Loan programs, and more broadly, U.S. labor markets, are not performing at their full potential. There is a large mismatch between the skills workers have and employers’ needs, and this mismatch contributes to structural unemployment, reduced output, and student loan defaults.

This Article argues that introducing risk-based pricing in federal student loans would advance the interests and values that Congress articulated when it first established federal support for higher education. Risk-based pricing of student loans would signal the long-term financial risks inherent in different courses of study. This price signal would likely improve students’ ability to make informed decisions about the course of study that would best balance their innate abilities and individual preferences with postgraduate economic opportunities. Similarly, price signals would enhance postsecondary educational institutions’ abilities to adjust their programs to improve their students’ postgraduate prospects.

Allocating educational resources more efficiently would not only benefit individual students and their families. It would enhance the productivity and competitiveness of the U.S. labor force, with beneficial consequences for both the private sector and public finances. Over the long term, such efficiencies could increase the resources available for further investment in education and research.

Transparent, risk-based student loan pricing could greatly benefit students and educational institutions, particularly if it
were data-driven and sensitive to the values of equal opportunity and independent research that are central to the academic enterprise. This Article discusses legal and policy reforms that could facilitate risk-based student loan pricing, potential hazards from a shift toward risk-based pricing, and safeguards that could help protect students and educators from abuse.

This Article focuses primarily on the economic consequences of education rather than on moral or philosophical views about the ideal purpose of education or its proper role in society. The economic focus of this Article is not intended to deny the intellectual merit of philosophical views about education, but rather to reflect the fact that government support for higher education in the United States has primarily been driven by economic considerations, particularly during the mid-twentieth century when Federal Student Loan programs were established.

Part I of this Article discusses rationales for government support for higher education, with an emphasis on Human Capital Theory. Part II discusses the U.S. federal student loan system. Part III discusses coordination, information, and incentive problems in the higher education and skilled labor markets. Part IV explains the theory of risk-based credit pricing and how risk-based pricing of federal student loans could ameliorate some of the coordination problems discussed in Part III. Part V discusses predictors of income, employment, and student loan default, and also considers ethical and moral considerations that might limit or preclude the use of certain predictors to risk-adjust student-loan pricing.

II. Government Support for Higher Education

In most developed economies, government provides some form of public support for higher education, either through grants or loans. Rationales for government support for higher education

generally relate to positive externalities beyond the direct benefits to the individual student. These externalities may be economic in nature, or may relate to more subjective values espoused by a given polity. Values-based rationales in the United States often cite the role of public investment in education in reducing inequality or providing socioeconomic mobility.

A. Higher Education as an Investment in Human Capital

Economic benefits of higher education are well known: education increases wages and reduces the risk of

---


4. See E. DIGBY BALTZELL, THE PROTESTANT ESTABLISHMENT: ARISTOCRACY AND CASTE IN AMERICA 351 (1964) (“[T]he campus community has now become the principal guardian of our traditional opportunarian ideals.”); JOHN A. DOUGLASS, THE CALIFORNIA IDEA AND AMERICAN HIGHER EDUCATION: 1850 TO THE 1960 MASTER PLAN 1–2 (2000) (“California was not alone in its efforts to nurture higher education as both a tool for socioeconomic mobility and an engine for economic growth.”); OECD, supra note 2, at 13 (“During the past 50 years, the expansion of education has contributed to a fundamental transformation of societies in OECD countries. In 1961, higher education was the privilege of the few . . . .”); Lani Guinier, The Supreme Court, 2002 Term: Comment: Admissions Rituals as Political Acts: Guardians at the Gates of our Democratic Ideals, 117 HARV. L. REV. 113, 137 (2003) (“I identify four important values associated with access to higher education: individualism, merit, democracy, and upward mobility. Of these four, the value that seems to integrate the other three with higher education is upward mobility.”).

5. See GARY S. BECKER, HUMAN CAPITAL: A THEORETICAL AND EMPIRICAL ANALYSIS WITH SPECIAL REFERENCE TO EDUCATION 246 (1994) (“The rate of return to an average college entrant is considerable, of the order of 10 or 12 percent per annum”); id. at 247 (“[A]bility explains only a relatively small part of the [earning] differentials [between high school and college educated workers] and college education explains the larger part.”); Orley Ashenfelter & Alan Krueger, Estimates of the Economic Return to Schooling from a New Sample of Twins, 84 AM. ECON. REV. 1157, 1157 (1994) (estimating from a sample of identical twins that an additional year of schooling increases wages by 12% to 16%, and reporting that this is probably not due to differences in innate ability); Thomas Lemieux, Postsecondary Education and Increasing Wage Inequality, 96 AM. ECON. REV. 195, 196 (2006) (“By 2003–2005 . . . the return to post-secondary education is much
unemployment, presumably by increasing labor productivity. In addition to benefiting the student by facilitating higher future income, education may also lead to positive financial externalities such as increased tax revenues, reduced burdens on public services.

higher than the return to elementary and secondary education."); id. at 199 ("[P]ost secondary education plays a crucial role in explaining [increasing wage inequality]. By contrast, labor market experience, primary and secondary education, and the position of workers without postsecondary education in the wage distribution play a small role in explaining changes in the wage structure over the last 35 years."); OECD, supra note 2, at 13 ("Among the 34 OECD countries, most of those in which college enrolment expanded the most over the past decades still see rising earnings differentials for college graduates . . . .").


7. See Mincer, supra note 6, at 22; David A. Wise, Academic Achievement and Job Performance, 65 AM. ECON. REV. 350, 364 (1975) (providing evidence that college education increases productive ability); cf. Samuel Bowles & Herbert Gintis, Schooling in Capitalist America Revisited, 75. SOC. EDUC. 1, 1 (2002) ("[T]he contribution of schooling to individual economic success could be explained only partly by the cognitive development fostered in schools . . . . [S]chools prepare people for adult work rules by socializing people to function well and without complaint in the hierarchical structure of the modern corporation."); Joseph Stiglitz, The Theory of “Screening,” Education, and the Distribution of Income, 65 AM. ECON. REV. 283, 298 (1975) (arguing that education acts to provide information to employers about the innate abilities and characteristics of prospective employees and that education may not in and of itself improve labor productivity); Paul J. Taubman & Terence J. Wales, Higher Education, Mental Ability, and Screening, 81 J. POL. ECON. 28, 43 (1973) (supporting the screening hypothesis).

8. OECD, supra note 2, at 165 ("Investments in education also generate public returns from higher income levels in the form of income taxes, increased social insurance payments and lower social transfers.").

9. See id. at 193

A large body of literature suggests that education is positively associated with a variety of social outcomes, such as better health, stronger civic engagement and reduced crime. A small but increasing number of studies further suggest that education has a positive causal effect on these social outcomes. There is also research suggesting that education can be a relatively cost-effective means to improve health
and more rapid technological innovation and economic growth.10


Just as accumulation of personal human capital produces individual (income) growth, so do the corresponding social or national aggregates. . . . growth of human capital is both a condition and consequence of economic growth. . . . [h]uman capital activities involve . . . the production of new knowledge which is the source of innovation and of technical change which propels all factors of production.
Figure 1.1: Educated Workers Earn More and Are Unemployed Less

Educational attainment and median weekly earnings, 2011
2011 USD, workers age 25 or over


Educational attainment and average unemployment rates, 2011
Percent of workers age 25 or older who were unemployed

Figure 1.2: Decades of Data Show that Educated Workers Are Less Likely to be Unemployed

Average annual unemployment rates, age 25 or older, 1992-2011
Percent of workers age 25 or older who were unemployed

Figure 1.3: Decades of Data Show that Educated Workers Earn More, and the Wage Premium Has Increased over the Last Thirty Years

Median usual weekly earnings of full-time workers 25 years and over by educational attainment, 1979-2011

Real 2011 USD

This perspective—known as “Human Capital Theory”\(^\text{11}\)—is the leading economic explanation for the higher wages of educated workers. An alternate view that developed during the 1970s, “Signaling Theory,” claims that education leads to a more efficient allocation of talent by sorting workers according to innate ability.\(^\text{12}\) Risk-based pricing of student loans is compatible with either a Human Capital or Signaling view, although the case for subsidized education is stronger under Human Capital Theory.

Empirical evidence in favor of Human Capital Theory has mounted over the last thirty-five years, including many studies of wage differences of identical twins who differed with respect to the number of years of education.\(^\text{13}\) In addition to the twin studies, there have been many careful econometric studies that controlled for various measures of innate ability.\(^\text{14}\) These studies


\(^{12}\) See Stiglitz, supra note 7, at 283 (discussing a “screening” process that allows individuals to be labeled by their productivity); Taubman & Wales, supra note 7, at 43–49 (suggesting using education as a screening device). Under Signaling Theory, education can create value because it enables the employers who value skilled workers the most to identify those workers and bid for their services, leading to a more efficient allocation of skilled labor. Signaling Theory implies that labor market outcomes should not depend on what students study, but only on how well they perform academically relative to other students with similar standardized test scores, or perhaps whether they demonstrate a strong work ethic by choosing a challenging major.


\(^{14}\) See David Card, The Causal Effect of Education on Earnings, in 3
suggest that a college degree on average increases wages by 40%.\textsuperscript{15}

Figure 1.4: Education Boosts Wages After Controlling for Student Ability

Economists' estimates of increase in wages caused by one year of schooling

Percent increase in lifetime wages from an additional year of school, midpoint estimate

Early Human Capital Theory focused on the number of years of schooling or the completion of a degree, while more recent studies have focused on differences between fields of study. These studies generally conclude that choice of field of study affects wages and employment, even after controlling for ability

\textsuperscript{15} Id. at 1802.
sorting. Figures 2.1 through 2.3 below show differences in earnings and employment by college major, both at graduation and later in life.

Figure 2.1: Some Academic Majors Have a Higher Initial Labor Market Value Than Others

Recent graduates' median starting salary offer by major, 2011
2011 USD thousands

- Engineering: $59.1
- Computer Science: $56.2
- Mathematics: $49.2
- Accounting: $48.4
- Economics: $47.9
- Healthcare: $42.2
- Business Administration: $40.9
- Liberal Arts/Humanities: $34.2
- Communications: $32.9
- Biology: $31.1
- History/Political Science: $30.7
- English: $30.3
- Education: $29.7
- Psychology: $28.7
- Sociology/Social Work: $28.5
- Visual & Performing Arts: $24.1

Source: National Association of Colleges and Employers, The Class of 2011 Student Survey Report 36 Figure 30.
Note: Bachelor's degree recipients only.

16. See Amanda Thorson, The Effect of College Major on Wages, 13 THE PARK PLACE ECONOMIST 45, 48 (2005), https://www.iwu.edu/economics/PPE13/thorson.pdf ("Every study on the matter... shows that at least some gap remains even after controlling for human capital variables when looking at either specific majors or aggregated major groups.")
Figure 2.2: Some Academic Majors Are More Likely to Lead to Employment at Graduation Than Others

**Job offer rate by major, 2011**

Percent of recent graduates with job offers at graduation

Source: National Association of Colleges and Employers, The Class of 2011. Student Survey Report 34 Figure 28.
Figure 2.3: Over the Long Term, College Graduates in Some Fields Earn More and Are More Likely to Work Full Time

Annualized median earnings by bachelor degree field, 2009
Population age 18 and over where highest degree is bachelor’s
2011 USD thousands

<table>
<thead>
<tr>
<th>Field</th>
<th>Full time workers</th>
<th>All workers with earnings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering</td>
<td>$75</td>
<td>$80</td>
</tr>
<tr>
<td>Computer</td>
<td>$72</td>
<td>$70</td>
</tr>
<tr>
<td>Business</td>
<td>$58</td>
<td>$53</td>
</tr>
<tr>
<td>Mathematics</td>
<td>$59</td>
<td>$49</td>
</tr>
<tr>
<td>Communications</td>
<td>$52</td>
<td>$49</td>
</tr>
<tr>
<td>Health sciences</td>
<td>$57</td>
<td>$48</td>
</tr>
<tr>
<td>Other</td>
<td>$56</td>
<td>$48</td>
</tr>
<tr>
<td>Natural science</td>
<td>$54</td>
<td>$47</td>
</tr>
<tr>
<td>Agriculture</td>
<td>$52</td>
<td>$46</td>
</tr>
<tr>
<td>Architecture</td>
<td>$50</td>
<td>$43</td>
</tr>
<tr>
<td>Liberal arts</td>
<td>$50</td>
<td>$42</td>
</tr>
<tr>
<td>Social science</td>
<td>$48</td>
<td>$42</td>
</tr>
<tr>
<td>Philosophy</td>
<td>$50</td>
<td>$39</td>
</tr>
<tr>
<td>Literature</td>
<td>$52</td>
<td>$39</td>
</tr>
<tr>
<td>Education</td>
<td>$43</td>
<td>$38</td>
</tr>
<tr>
<td>Foreign language</td>
<td>$45</td>
<td>$37</td>
</tr>
<tr>
<td>Psychology</td>
<td>$47</td>
<td>$37</td>
</tr>
</tbody>
</table>

Source: US Census Bureau, Survey of Income and Program Participation, 2008 Panel, Table 4G.
Note: Bachelor’s degree recipients only; annualized earnings calculated by multiplying monthly earnings by 12.
Human Capital Theory helps explain wage and employment differentials between theoretical and applied majors. For example, although math majors on average have higher standardized test scores than engineering or computer science majors,\textsuperscript{17} math majors are less likely to be offered employment at graduation and receive lower starting salary offers than students who majored in computer science or engineering.\textsuperscript{18} Similarly, business majors, who have relatively low average standardized test scores,\textsuperscript{19} have better labor market outcomes than higherscoring social science or humanities majors.\textsuperscript{20}
Figure 3.1: Differences in Earnings by Major Do Not Appear to Be Due Solely to Differences in Student Ability

2011 median starting salary offer and 2007 mean SAT score by college major

Real 2011 USD thousands


Note: Differences in SAT scores may be underestimated because SAT scores are for intended majors and salaries are for completed majors. There is some evidence that lower ability students switch from challenging majors such as Engineering and Computer Science into less challenging majors such as Business, English, and other social science and humanities fields.
Although these observations could be interpreted in various ways, the differences appear to reflect the value of

21. It is possible that low-ability students are “signaling” commercialism rather than developing practical skills, but presumably entering the work force at a younger age would signal commercialism more forcefully than studying something “commercial.”

Another possibility is that high-ability math, social science, and humanities majors may opt out of the labor market by going to graduate school, and that
field-specific skill development rather than differences in ability levels. Even within engineering, there are large starting wage differences by specialty.22

Human Capital Theory also helps explain higher average per-capita productivity and wages in states and nations with higher levels of educational attainment. If education only sorted workers according to ability, it would presumably only increase the variance of wages (i.e., income inequality), while leaving the mean unaltered.23

Further, Human Capital Theory helps explain the willingness of many employers to pay for professional degree programs for successful employees.24 Employers’ willingness to educate workers whom employers already know to be of high quality suggests that employers believe that professional education has skill-development value rather than mere sorting value.

Just as corporations depend on the productivity of their employees, workers’ productivity and wages are an extremely important source of revenue for central governments. Labor is less mobile than capital, and therefore easier to tax.25

only the students in these fields with relatively low abilities may enter the labor market at college graduation, driving down reported wages and employment.


23. Signaling Theory can explain these findings either by assuming that sorting creates collective as well as private benefits, or under strained interpretations of the data—for example assuming that prosperity causes education, or that a third unidentified variable consistently causes both high levels of education and high levels of prosperity. See, e.g., Andrew Weiss, Human Capital vs. Signaling Explanations of Wages, 9 J. ECON. PERSP. 133, 145–46 (1995) (addressing objections to the “Sorting Approach”).


25. It is far more difficult to learn a new language and emigrate than to convert capital to a new currency and invest across borders. See, e.g., OECD, TAX POLICY REFORM AND ECONOMIC GROWTH 19 (2010) (“Globalization may . . . increase the opportunities for tax avoidance and evasion especially as concerns
In a country such as the United States, which taxes wages at much higher rates than capital, public expenditures that increase wages are more likely to benefit public finances through higher future tax revenues than public expenditures that increase the return on private capital.26

Whereas the capital gains tax rate is typically fifteen percent, the average effective tax rate on human capital—that is, the tax on the increase in wages attributable to education—will often be around thirty to fifty percent because the wage premium will fall into high federal, state, and local income tax brackets and will often also be subject to payroll taxes.27 In addition, education is not treated as favorably under the Tax Code as other forms of investment with respect to the ability to recover investment costs, deduct interest on loans, or smooth income across tax years.

In sum, a large proportion of the benefits of human capital redound to public finances rather than to the educated worker. Education is generally a profitable public investment, not a

26. Wages are subject to both federal income taxes and federal payroll taxes, whereas capital gains and dividends are subject only to income taxes. Income tax rates for capital gains and dividends are much lower than income tax rates for wages. The difference in tax treatment of income from wages and income from capital is so extreme that although the income tax is nominally progressive, in practice extremely wealthy individuals who derive most of their income from investments have much lower average federal tax rates than middle class workers who derive most of their income from wages. See Martin A. Sullivan, Economic Analysis: At the Helmsley Building, the Little People Pay the Taxes, 130 TAX NOTES 855, 855–56 (2011) (discussing how tax rates are lower for the very wealthy than for the average person). Some have countered that corporate income taxes should be counted as additional taxes on capital, but whether the incidence of corporate tax is primarily on investors, employees, or customers remains hotly debated. See, e.g., Charles E. McLure, Jr., The Elusive Incidence of the Corporate Income Tax: The State Case, 9 PUB. FIN. Q. 395, 395–98 (1981) (discussing the controversies and approaches to the corporate income tax).

27. See, e.g., OECD, TAXING WAGES 2008–2009, at 109 (2010) (estimating total 2009 marginal tax burden on labor in the U.S. to be between 30% and 60% for workers earning at least 50% of the average wage, with the highest tax burdens on singles).
mere expenditure. In fact, the public benefits from higher education in the United States are the highest in the developed world, while public costs are among the lowest, suggesting that public investment in higher education in the United States could be profitably increased.

**B. The Demand for Skilled Labor and Social Mobility**

Whereas private higher education in the United States was originally a form of luxury consumption—training for the financially secure children of the upper class that emphasized cultural refinement and social grace over technical skill—federal government support for higher education emerged with a belief by business leaders that education can and should promote economic development by training skilled labor and supporting applied research. This emphasis on economic development is evident in the requirements of the Northwest Ordinances of 1785 and 1787 and the Morrill Act of 1862.

28. OECD, supra note 2, at 158–60 (reporting that public and private benefits of education in OECD countries, including the U.S., greatly exceed public and private investment in education).

29. Id. at 165–67.

30. DOUGLASS, supra note 4, at 2; OECD, supra note 2, at 13.

31. DOUGLASS, supra note 4, at 2–3, 33–34. The strongest proponents of practical education were Northeastern business interests, while the principal opposition came from Southern conservatives. Id.

32. See the full text transcripts of Land Ordinance 1785, available at http://memory.loc.gov/cgi-bin/query/r?ammem/bdsdcc:@field(DOCID+@lit(bdsdcc13201)), and 1787, available at http://www.ourdocuments.gov/doc.php?doc=8 (last visited Feb. 3, 2013) (on file with the Washington and Lee Law Review); DOUGLASS, supra note 4, at 20 (“The constitutions of existing states provided for one or more state-supported institutions of higher learning as a means to further social and economic progress and as a legal mechanism for securing federal land grants for education under the Northwest Ordinances of 1785 and 1787.”).

33. Officially known as the Agricultural College Land Act. 7 U.S.C. §§ 301–49 (2012). See Bok, supra note 10, at 62 (“Americans tended to look on higher education as a means for providing the knowledge and trained man power that a rapidly developing society required. In 1862, Congress embodies this spirit in the Morrill Act . . . .”); DOUGLASS, supra note 4, at 34 (“As a condition for accepting Federal scrip, by 1866 each state would need to charter either existing or new institutions to fulfill the purpose of the act: namely, to provide
under which the federal government granted land to state governments to fund public institutions of higher learning that would teach labor-market-relevant skills.34 Similarly, Congress emphasized the need for a technically skilled labor force, particularly in areas of science and technology, when it implemented the first federal student loan program through the National Defense Education Act of 1958 (NDEA).35 The need for greater central government support for higher education was made salient in 1957 by the Soviet Union’s launch of Sputnik I and II, the first man-made satellites.36 These early Soviet technological triumphs over the United States were generally attributed in the U.S. to the Soviet Union’s seemingly superior system of education.37 The Soviet educational system, compared to the U.S. system, was believed to be more meritocratic, to focus more on science and technology, and to more closely coordinate its efforts with national economic and military priorities.38

34. DOUGLASS, supra note 4, at 33–34.
The guaranteed student loan program [established by the Higher Education Act of 1965] took as its model . . . loans offered under the National Defense Education Act (NDEA), a law passed in 1958 in reaction to the launch of the Sputnik satellite by the Soviet Union. At that time, lawmakers encouraged Americans to educate themselves in scientific and technical fields.
36. See DOUGLASS, supra note 4, at 198 (“[T]he substantial increases in direct student aid under post-Sputnik federal legislation initiated a new era of federal involvement in higher education.”); id. at 234 (“Sputnik was a technological marvel. It was the first intercontinental missile, opened the space age, and marked the beginning of satellite communications. It was also a profound political event.”).
37. See id. at 234 (“American popular opinion credited the Soviet Educational System with Sputnik’s success. Here was the source for its scientists and research. Conversely, the reason for America’s apparent second place position . . . was its faltering schools and universities.”).
38. See id. (“The quick conclusion of many was that America’s system of education was disorganized, it failed to provide sufficient training in the
race that followed, the U.S. shifted toward a centralized, taxpayer-funded, and government-coordinated model of university-based scientific and technical research, coupled with increased education subsidies.

In approving subsequent federal student loan programs, such as the guaranteed loan program established by the Higher Education Act of 1965, Congress emphasized the need for greater equality of opportunity and social mobility as well as the need for a skilled labor force.

Recently, state governments have renewed their insistence that public support for higher education should be conditional on higher education serving the needs of the labor market and economic growth. And educational leaders have recognized the legitimacy of government efforts to coordinate universities’ activities with economic priorities.

---

40. Glater, supra note 35, at 20, 35–38.
41. See Erin Sparks & Mary Jo Waits, Nat’l Governors Ass’n Ctr. for Best Practices, Degrees for What Jobs? Raising Expectations for Universities and Colleges in a Global Economy 40 (2011), http://www.nga.org/files/live/sites/NGA/files/pdf/1103DEGREESJOBS.PDF (“Governors and state policymakers are increasingly recognizing the importance of ensuring that students who graduate from institutions of higher education . . . are equipped with the skills to fill good, high-paying jobs that are in high demand by employers, thereby boosting the state’s economic growth.”).
42. See Bok, supra note 10, at 40

We cannot assume that . . . market forces will automatically lead colleges and universities to train physicians or doctoral students in numbers corresponding to society’s needs. If the government is subsidizing university programs or if these programs are important enough to the public, officials will naturally wish to intervene whenever the results stray too far from the nation’s interests.
C. Values-Based Arguments for Higher Education Funding

In addition to financial benefits, many commentators have argued that education provides some ethical, spiritual, or political benefits, not only to the individual student, but also to society at large.\textsuperscript{43} Purported benefits of education range from promoting equality or social mobility, to safeguarding liberty, to reinforcing moral and ethical behavior, to fostering informed participation in democratic processes, to encouraging voluntarism and civic virtue.\textsuperscript{44}

In the nineteenth century, private colleges—in contrast to state universities—often saw their role as the ethical and moral development of good parishioners and good citizens. However, by the mid-twentieth century, this moralistic view was largely supplanted even at elite private colleges by a focus on the role of higher education in promoting individual and collective economic advancement.\textsuperscript{45}

In the United States, social mobility, equality of opportunity, and material progress were viewed not only as private goods, but also as public benefits that legitimized the United States’ political and economic systems, brought more talented individuals into leadership positions, and dampened the appeal of communism. Even as the Cold War has receded into distant memory, the prospects of equal opportunity and social mobility continue to be

\textsuperscript{43} See OECD, supra note 2, at 192 (“Adults aged 25 to 64 with higher levels of educational attainment are, on average, more satisfied with life, engaged in society and likely to report that they are in good health, even after accounting for differences in gender, age and income.”).

\textsuperscript{44} See Glater, supra note 35, at 12–13, 16–19; Guinier, supra note 3, at 115–33, 137.

\textsuperscript{45} See Bok, supra note 10, at 3–4

[In the early 1900s,] the American University was evolving from a church-oriented college into a larger, more diverse institution with stronger graduate and professional programs capable of serving the needs of a developing economy. . . . [B]usinessmen and financiers quickly replaced the clergy as dominant figures on the boards of leading universities.

\textit{See also id.} at 62–66; \textit{id.} at 121 (“[B]y the mid-twentieth century, little remained of the earlier efforts of colleges and universities. Catalogues continued to speak of moral development as a prominent aim of the institution, but there was scant evidence of any serious effort to pursue this objective.”).
cited as justifications for inequality—in effect, the prospect of social mobility is a substitute in U.S. political discourse for equality.46


There is also a conservative justification for both high inequality and low levels of social mobility, which rests on an assumption of very high heritability of talent or ability, and assumes that those who are poor are deficient in ways that are heritable and largely immutable. See, e.g., Richard Herrnstein & Charles Murray, The Bell Curve: Intelligence and Class Structure in American Life (2004) (arguing that genetically heritable intelligence determines class structure in the United States).

Most labor economists, demographers, sociologists, and psychologists reject this view as inconsistent with the data. See, e.g., Michael Hout et al., Inequality by Design: Cracking the Bell Curve Myth (1996) (reanalyzing the data used in the Bell Curve and arguing that the authors overestimated the role of intelligence in setting wages and underweighted the role of manipulable factors such as education); Lisa Barrow & Cecilia Rouse, The Economic Value of Education by Race and Ethnicity, 2006 Econ. Persp. 14, 23 (analyzing data and concluding that returns on education do not differ by race); James J. Heckman, Lessons from the Bell Curve, 103 J. Pol. Econ. 1091, 1091–1120 (1995); Orley Ashenfelter & Cecilia Rouse, Schooling, Intelligence, and Income in America: Cracks in the Bell Curve (Nat’l Bureau of Econ. Research, Working Paper No. 6902, 1999), http://www.nber.org/papers/w6902.pdf (reviewing the econometric literature and concluding that the economic returns on schooling do not differ significantly by family background or by measures of ability of the student); Christopher Winship & Sanders Korenman, A Reanalysis of the Bell Curve 1, 21–22 (Nat’l Bureau of Econ. Res., Working Paper No. 5230, 1995), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=225294 (arguing that the measure of parental socioeconomic status used in the Bell Curve did not capture important family characteristics such as single-parent family structure at age fourteen, and therefore overestimated the effects of intelligence).


However, whatever the heritability of intelligence, there is substantial evidence from randomized controlled studies and quasi-experimental designs that early childhood interventions, smaller class sizes, career training programs, and college completion can improve educational and economic outcomes. See
D. Higher Education Funding and Independent Research

Governments may also fund educational institutions because of the benefits of unbiased research conducted by experts who are insulated from political and market pressures. Notable leaders of educational institutions have expressed concerns that external funding can corrupt academic research. For example, industry funding of research is affiliated with scientifically questionable pro-industry conclusions in pharmaceutical research, nutritional research, and environmental research.


47. See Charles I. Jones, Sources of U.S. Economic Growth in a World of Ideas, 92 AM. ECON. REV. 220, 228 (2002) (arguing that 30% of U.S. growth between 1950 and 1993 is attributable to the rise in educational attainment and 50% is attributable to the rise in worldwide research intensity).


When government funding comes with direct control by political leaders, such funding also creates the risk of attempts to politicize education or enforce a rigid ideology. Federal is very high confidence that the global average net effect of human activities since 1750 has been one of warming . . . .\); id. at 39 ("Most of the observed increase in global average temperatures since the mid-20th century is very likely due to the observed increase in anthropogenic [greenhouse gas] concentrations."); NAOMI ORESKES & ERIK M. CONWAY, MERCHANTS OF DOUBT: HOW A HANDFUL OF SCIENTISTS OBSCURED THE TRUTH ON ISSUES FROM TOBACCO SMOKE TO GLOBAL WARMING 2–9 (2010) (addressing the doubt cast upon scientific research in regard to the tobacco industry and global warming); JAMES LAWRENCE POWELL, THE INQUISITION OF CLIMATE SCIENCE 64–65 (2011) (noting the criticism of scientists who dissent from the findings of the IPCC); Riley E. Dunlap & Aaron M. McCright, Climate Change Denial: Sources, Actors, and Strategies, in ROUTLEDGE HANDBOOK OF CLIMATE CHANGE AND SOCIETY 240, 240–45 (Constance Lever-Tracy ed., 2010) (examining how uncertainty regarding climate change has been manufactured over time).


government control over education curricula and personnel decisions is now restricted by statute.\textsuperscript{53} Tuition—and indirectly, student loans—can provide a neutral source of funding for unbiased research because students are unlikely to have a personal financial or partisan interest in the outcome of their professors’ research.

\textit{E. Higher Education Funding Options: Student Debt or General Taxes}

To the extent that one accepts the existence of one or more positive externalities of education, education may be a natural public good that should be subsidized by government.\textsuperscript{54} However, a government’s ability to benefit from educating its citizens may be constrained when government-funded education provides portable skills and workers can readily seek employment across political borders.\textsuperscript{55} A government that generously funds education with the expectation of higher future tax revenues may fall prey to another government that actively seeks educated immigrants and can charge lower taxes because it does not provide as much public funding for education.\textsuperscript{56} Governments can reduce the financial risk of

\textsuperscript{53} Limits on U.S. government control over education are codified at 20 U.S.C. § 1232a.

\textsuperscript{54} Guinier, \textit{supra} note 4, at 129–30 (“[T]he shift in funding priorities [away from education] was driven in part by an ideological shift during the Reagan era. Higher education was presented as a private benefit to be financed by the individual, instead of a public good to be funded by the government.”).

\textsuperscript{55} See Demange, Fenge & Uebelmesser, \textit{supra} note 2, at 248 (arguing that the mobility of students has made educational competition between countries more intense); Poutvaara, \textit{supra} note 2, at 663 (stating that the training government does not fully realize return of educational investment for emigrants).

\textsuperscript{56} The United States has been particularly successful at attracting technically skilled immigrants educated in—and often at the expense of—other countries, but not particularly successful at providing technical education to its native population. \textit{See} Frederic Docquier & Abdeslam Marfouk, \textit{International Migration by Education Attainment, 1990-2000, in International Migration, Remittances and the Brain Drain} 151, 152–53, 187 (Caglar Ozden & Maurice
emigration by structuring public funding for education as loans rather than as outright grants. 57

Higher education in the United States—a country where relatively few graduates have internationally transferable technical skills 58 and out migration rates are relatively low 59—is unusual because of heavy reliance on private funding rather than public funding. 60 In much of the rest of the developed


57. See Poutvarra, supra note 2, at 680–82 (proposing income-contingent loans as one solution to problems facing European public education).

58. The percentage of U.S. graduates with science, math, computer science, or engineering degrees is very low compared to the rest of the developed world. See Jeffrey J. Kuenzi, Cong. Research Serv., Science, Technology, Engineering, and Mathematics (STEM) Education: Background, Federal Policy, and Legislative Action, at CRS-1 (2008) (“When compared to other nations, the math and science achievement of U.S. pupils and the rate of STEM degree attainment appear inconsistent with a nation considered the world leader in scientific innovation.”); OECD, supra note 2, at 80 (showing that the United States lags behind other countries in the number of tertiary graduates in science-related fields). A disproportionately large share of awarded and commercialized U.S. patents are authored by immigrants who were educated elsewhere. See Jennifer Hunt & Marjolaine Gauthier-Loiselle, How Much Does Immigration Status Boost Innovation?, 23 (Nat'l Bureau of Econ. Research, Working Paper No. 14312, 2008), http://www.nber.org/papers/w14312.pdf (“We find that a college graduate immigrant contributes at least twice as much to patenting as his or her native counterpart. The difference is fully explained by the greater share of immigrants with science and engineering education.”).

59. See Docquier & Marfouk, supra note 56, at 168–72 (estimating that North America had an emigration rate for skilled labor of only 0.9%, by far the lowest of any region studied).

60. See OECD, supra note 2, at 165 (“Direct costs for education are generally borne by the public sector, except in Australia, Japan, Korea, and the United States, where private direct costs such as tuition fees constitute over half of the overall direct investment costs.”); id. at 231–34 (showing the United States as having a much greater reliance on private funding for higher education than the average country); Demange, Fenge & Uebelmesser, supra note 2, at 253–54 (stating that unlike in the European Union, private sources of funding for education are more important than public sources in the United
world, governments primarily finance higher education through general tax revenues. Students are expected to pay minimal tuition and fees while they are in school, and as a result, recent graduates are burdened with minimal debt. Access to university education may be allocated through competitive examination, but inequality in family financial resources generally has a limited impact on educational attainment. Because the government provides much of the funding for education, the government can readily prioritize certain fields of inquiry by devoting more resources to those subject areas, and can try to match educational offerings to employment opportunities.

By contrast, in the United States, federal government support to students is generally in the form of loans that must be repaid with interest, and students therefore graduate with high debt burdens. Although some state governments support

61. See Demange, Fenge & Uebelmesser, supra note 2, at 253 (stating that public funding based on tax revenues is dominant in European Union countries).

62. OECD, supra note 2, at 48 (“[I]n Finnish higher education . . . the number of entry places is restricted.”); Demange, Fenge & Uebelmesser, supra note 2, at 264–65 (discussing access restrictions in Germany and France).

63. See OECD, ECONOMIC POLICY REFORMS: GOING FOR GROWTH 2010, at 194 (hereinafter OECD, ECONOMIC POLICY REFORMS) (“In some countries, there exist social transfer programmes that are specifically directed to paying part of [parent costs in poor households of investing in the education of their children]. Such redistributive policies could thus reduce current income inequalities across parents so that their descendants’ income would converge more quickly.”); Charlene Marie Kalenkoski & Sabrina Wulff Pabilonia, Parental Transfers, Student Achievement, and the Labor Supply of College Students, 23 J. POPULATION ECON. 469, 494–95 (2010) (finding that students who receive less support from their parents work longer hours while in school, and that longer work hours reduce these students’ GPAs).

64. As of 2010, federal loans exceed federal grants by a factor of more than two to one. Earlier in the decade, the proportion of loans was even higher. COLLEGE BD., TRENDS IN STUDENT AID 2011, at 10 tbl.1 (2011), http://trends.collegeboard.org/sites/default/files/Student_Aid_2011.pdf. Total grants and total loans—not just federal—each account for roughly half of the aid to undergraduates, but for graduate students, loans exceed grants by a factor of two to one. Id. at 17, tbls.8A & 8B. Government support has declined as a share of U.S. educational institutions’ revenue since the early 1980s. Michael S. McPherson & Morton O. Schapiro, U.S. Higher Education Finance, in HANDBOOK OF ECONOMICS OF EDUCATION 1403, 1403–34 (Eric Alan Hanushek & Finis Welch eds., 2006).

65. See Elizabeth Warren, Sandy Baum & Ganesh Sitaraman, Service
public universities that offer lower tuition to residents, state support for higher education has been eroding for decades, and public universities increasingly resemble private universities in their dependence on tuition revenues. Parental financial resources are a strong predictor of educational achievement, and intergenerational social mobility is low by developed-world standards.

Pays: Creating Opportunities by Linking College with Public Service, 1 HARV. L. & POL’Y REV. 127, 127 (“[S]tudents . . . are leaving college deep in debt.”); id. at 129 (stating that most United States college graduates take on debt to pay for college); Guinier, supra note 4, at 130 n.67. In 2009–2010, 56% of students who attended four-year public colleges borrowed money to do so, and they each borrowed an average of $22,000. Sixty-five percent of students who attended four-year private colleges borrowed, and they each borrowed an average of $28,000. The percent who borrowed and the average dollar value of debt (adjusted for inflation) have both increased over the last decade. COLLEGE BD., supra note 64, at 4, 19 figs.10A & 10B.

European governments have also slightly reduced the proportion of public support for higher education, but public support still accounts for a much larger share than in the United States. See Demange, Fenge & Uebelmesser, supra note 2, at 265–66.

Parental or socio-economic background influences descendants’ educational, earnings and wage outcomes in practically all countries for which evidence is available. . . . The influence of parental socio-economic status on students’ achievement in secondary education is particularly strong in Belgium, France and the United States . . . . Inequalities in secondary education are likely to translate into inequalities in tertiary education and subsequent wage inequality.

See also Warren, Baum & Sitaraman, supra note 65, at 127 (“[A]lmost 20% of low income-high school graduates with high test scores do not manage to enroll in college at all within two years of graduating high school.”).

See OECD, ECONOMIC POLICY REFORMS, supra note 63, at 185 fig.5.1, 187 (finding that intergenerational wage mobility as measured by father-son pairs is lower in the United States than in Denmark, Australia, Norway, Finland, Canada, Sweden, Germany, Spain, or France; only Italy and the U.K. had less social mobility than the United States). The OECD notes that genetic heritability of innate ability should be constant across countries, but wage mobility seems to be higher in countries with more generous taxpayer funded...
social welfare and education policies. *Id.* at 186, 196. The OECD also notes substantial evidence that early childhood education and care programs improve labor market outcomes for children from poorer backgrounds. *Id.* at 193–94. It is unclear if government funding of higher education has as large an impact on mobility as funding early childhood education.

Relatively low social mobility in the United States may be surprising to many; the statistical reality of a stable class structure starkly contrasts with perceptions of the United States as a dynamic, open, and fluid society. However, within the OECD, social mobility and equality appear to be complementary—the least equal societies tend to have the least mobility. *Id.* at 195 fig.5.10. See also *supra* note 46 for different perspectives on the desirability of mobility.
III. U.S. Federal Student Loan Programs

The overwhelming majority of the U.S. student loan market consists of federal government loans.69 Prior to the emergence of federally-backed student loans, student loans were rare and expensive, and higher education was generally only available to the children of the wealthy.

Historically, the largest category of government-backed loans was federal guaranteed loans, which were guaranteed, subsidized, and regulated by the government, but originated and owned by private financial institutions or sold to private investors through securitization.70

However, Congress eliminated guaranteed loans in 2010 and shifted all lending to the government’s direct loan program.71 Guaranteed loans were eliminated because of a widespread

69. As of 2011, 90% of new student loans were federal government loans. Government loans have been a majority of the market for many years, but increased dramatically after 2008 as private lending collapsed. Jonathan Riber & Maxim Berger, U.S. Private Student Loan Landscape, 7 U.S. STRUCTURED FIN. NEWSL. (DBRS, Toronto, Ont.), Oct. 5, 2011.


perception that the guarantees and subsidies—which reduced the riskiness of the loans to only slightly higher than U.S. government Treasuries, but enabled private lenders to profit by charging far higher interest rates—represented a subsidy to private financial institutions and their investors rather than a benefit to students or taxpayers.72

Federal direct government loans are administered by the U.S. Department of Education. The government retains ownership of these direct loans, and can therefore profit when the interest rate spread above Treasuries exceeds losses from defaults and administrative costs. Federal student loans are generally less expensive than private loans, but the federal direct loan program is still a moneymaker for the federal government.73

A. Student Eligibility Criteria Are Generally Not Risk-Based

The borrower eligibility criteria for federal student loans are fairly minimal, and generally not risk-based. A student must be

72. See Glater, supra note 35, at 39–40 (discussing the preferences of Senator Edward Kennedy and President Bill Clinton for direct lending on grounds of cost effectiveness and aggressive lobbying by private lenders against direct lending); id. at 57 n.224 (discussing costs savings from ending the guaranteed loan program); see also Deborah J. Lucas & Damian Moore, Guaranteed versus Direct Lending: The Case of Student Loans, in Measuring and Managing Federal Financial Risk 163, 164 (Deborah J. Lucas ed., 2010)

[T]he guaranteed program appears to be fundamentally more expensive than the direct program. . . . [G]uaranteed lenders are paid more than is required to induce them to lend at statutory terms. . . . To the extent that the market is not perfectly competitive, guaranteed lenders presumably are able to retain some of the surplus.

73. See Deborah Kalcevic & Justin Humphrey, Cong. Budget Office, CBO March 2012 Baseline Projections for the Student Loan and Pell Grant Programs, tbls.2 & 3 (Mar. 13, 2012) (projecting a negative subsidy, i.e., profit, for federal student loans originated in 2013 of around 32% of lending volume, or $36.5 billion in profit). After subtracting $1.6 billion in administrative costs (equal to 1.4% of lending volume), projected 2013 profits are $34.9 billion. Id. at tbl.4. The student loan program remains profitable in every year projected, although profits decline to around $10 billion in later years. See also Dep’t of Educ., Student Loans Overview, Fiscal Year 2013 Budget Request, at R-11 to R-12 (providing similar profit estimates for 2012 and noting that the federal student loan program was profitable in 2009–2012).
enrolled in a program at an accredited higher educational institution “leading to a recognized educational credential” such as a degree or certificate,74 must maintain “academic standing consistent with the requirements of graduation”75 unless there are “special circumstances,”76 must not currently be in default on a federal student loan,77 must be a U.S. citizen or on the path toward citizenship,78 and if previously convicted of defrauding the federal student loan program, must have made restitution.79 Eligibility can be suspended or terminated for drug offenses.80 The use of more restrictive eligibility criteria than those provided for by statute is generally prohibited.81

B. Only Exceptionally Poorly Performing Institutions Are Excluded

Similarly, the eligibility criteria for educational institutions are fairly minimal, with the Department of Education relying heavily on state accreditation agencies.82 Two sets of regulations have been established over the past two and a half decades to cull some of the worst performing institutions from student loan eligibility, but regulations do not seek to make fine performance-based distinctions among eligible institutions.

First, in response to high student loan default rates at some “proprietary” or “for-profit” educational institutions in the 1980s, Congress passed the Student Loan Default Prevention Initiative Act of 1990 (SLDPA).83 Under the SLDPA, institutions lost their

75. Id. § 1091(c)(1)(B).
76. Id. § 1091(c)(3)(C).
77. Id. § 1091(a)(3).
78. Id. § 1091(a)(5).
79. Id. § 1091(a)(6).
80. Id. § 1091(r).
81. Id. § 1077A(e); id. § 1077A(f).
82. Braucher, supra note 71, at 446, n.19.
eligibility for student loans if their cohort default rate (CDR) exceeded twenty-five percent for three years in a row.\footnote{84}{The CDR was the percent of students entering repayment in a given year who defaulted during the subsequent year, and CDR was therefore roughly a one-year default rate because students typically entered repayment in October or November. In 2008, the CDR measure was extended by one year (so that it captures students who enter repayment in one year and default by the end of the subsequent two years), and the maximum CDR was increased to 30%. \cite{Braucher supra note 71, at 464–65 & n.121}}

The CDR measure helped eliminate some small and poorly performing institutions, but sophisticated educational institutions increasingly manipulated the CDR statistic by moving recent students into deferment\footnote{86}{Deferment refers to a postponement of payment on a loan that is allowed under certain conditions and during which interest does not accrue for subsidized loans. Deferment is available for student borrowers who are enrolled at least half time in an eligible postsecondary school or studying full time in a graduate fellowship program or an approved disability rehabilitation program. It is also available for up to three years of unemployment and economic hardship, or for active duty military service. Direct Loans, Deferment and Forbearance, DEP'T OF EDUC., http://www.direct.ed.gov/postpone.html (last visited Feb. 3, 2013) (on file with the Washington and Lee Law Review).} or forbearance\footnote{87}{Forbearance refers to a postponement of payment on a loan, typically if the borrower does not qualify for a deferment and is unable to make payments for a reason such as poor health. Interest continues to accrue during forbearance. \cite{Id.}} so that they would not count as defaulters.\footnote{88}{Braucher, supra note 71, at 465 n.127.} CDR had a positive but limited effect.\footnote{89}{\textit{Id.} at 464–65 (finding that CDR regulation helped shut down some, but not all of the worst performers).}

Recently, largely in response to another wave of high defaults at some proprietary educational institutions, the Department of Education established a Gainful Employment Rule (GER) that again attempts to cull the worst performing institutions.\footnote{90}{The Department of Education published its final Gainful Employment Rule on June 13, 2011. The Rule was scheduled to go into effect on July 1, 2012, and the earliest any educational institution might lose eligibility under the GER Rule is 2015. \textit{Id.} at 466.} GER may be more difficult to manipulate than the older CDR measure.
because GER measures performance based on repayment rates rather than default rates.91 GER uses two tests—one that looks at whether former students are in fact repaying their loans92 and another that looks at debt-service-to-income ratios to determine whether graduates have sufficient income to enable them to have a reasonable chance of repaying their loans.93 An educational institution may remain eligible for student loans if it passes either test in at least two out of four consecutive years.

C. Borrowing Limits Depend on Grade Level and Dependent Status

Federal student loan borrowing limits are set by statute.94 The loan limits are determined by the students’ grade level and, for undergraduates, students’ status as dependents. Annual Stafford Loan limits increase as undergraduates progress from year one to year three, and are higher for students who are “independent.”95 Stafford and Perkins Loan limits are higher for graduate students than for undergraduates.96 The federal loan limits are less than the total cost of attendance at most private colleges and many flagship public

91. Id. at 467–69.
92. A program can remain eligible if at least 35% of students are repaying at least some portion of the principal on their federal loans. Id. at 467–68.
93. A program can remain eligible if either the mean or the median annual student loan payment of graduates of their program is either 12% or less of annual earnings or 30% or less of discretionary income. Id. at 468.
94. 20 U.S.C. § 1078(b)(1)(A)&(B) (2006) (setting forth loan limits for Subsidized Stafford Loans); id. § 1078–8(d) (setting forth loan limits for Unsubsidized Stafford Loans); id. § 1087dd(a)(2) (setting forth loan limits for Perkins Loans); id. § 1087E(a)(1) (noting that limits on direct loans are the same as limits on guaranteed loans).
95. Undergraduate students are “independent” if they meet any of the following criteria: the student is at least age twenty-four, is married, is a veteran or on active duty in the military, is an orphan or a ward of the state, or has legal dependents other than a spouse. Direct Loans, Glossary, DEP’T OF EDUC., http://www.direct.ed.gov/glos.html#anchor388090 (last visited Feb. 3, 2013) (on file with the Washington and Lee Law Review).
96. See sources cited supra note 94.
colleges, and students with financial need may therefore turn to private loans to help make up the difference. Graduate students and the parents of dependent undergraduate students with good credit histories also have access to PLUS Loans, under which borrowing is limited by the students’ financial need rather than a fixed dollar amount.

The federal student loan limits could at best be described as crudely risk-based: as students exceed loan limits for less expensive lending programs, such as subsidized Stafford Loans and Perkins Loans, they will move on to more expensive programs such as PLUS Loans and their borrowing costs will increase. Students may also turn to higher cost private student loans or credit card debt when federal student loans are inadequate to finance their education.

A fully risk-based approach to loan limits would focus on expected debt-service-payment-to-income ratios. The relevant question is not simply how much students borrow each year. Instead, the relevant question is whether students’ incomes at graduation and beyond will be sufficient to repay their debts over the next ten to thirty years.

D. Federal Student Loan Pricing is Statutory, Not Risk-Based

In theory, rather than cut off access to credit entirely to poor performing institutions and ignore risk differences above a minimal threshold, the Department of Education could embrace a more nuanced approach by incorporating risk levels into loan pricing. However, in practice, federal student loan pricing is largely uniform and not risk-based.

Interest rates on government loans are set by statute at the same level for all eligible borrowers under a particular loan program. Federal student loan rates are currently set at a fixed rate between 3.4% and 7.9%, with lower rates available to

---

98. 20 U.S.C. § 1078-2(a)(1) (discussing eligibility for PLUS Loans); id. § 1078-2(b) (limiting PLUS Loan borrowing to a student’s estimated cost of attendance, minus other financial aid).
undergraduates than to graduate students. The rates have changed over time, but have historically been either a fixed interest rate, or a capped variable rate determined by adding a spread to a variable Treasury bill rate.

The interest rates are not risk-based. A successful medical student with virtually no risk of becoming unemployed or defaulting on her debts would pay the graduate student rate—between 6.8% and 7.9%—while a struggling art history major with rather less secure employment prospects would pay the undergraduate rate of 3.4%.

100. The interest rate for new loans made under the William D. Ford Federal Direct Loan Program (DLP), made on or after July 1, 2006, is a fixed rate, generally either 6.8% or 7.9%, depending on the loan program. 6.8% is the rate for Federal Direct Stafford Loans and Federal Direct Unsubsidized Stafford Loans; 7.9% is the rate for PLUS Loans. 20 U.S.C. § 1087E(b)(7) (2006). Between July 1, 2008, and July 1, 2013, a lower interest rate was available for undergraduate student borrowers under DLP Loans—this rate was 3.4% for undergraduate DLP Loans originating between July 1, 2011, and July 1, 2013. Id. § 1087E(b)(7)(D). The interest rate for Perkins Loans made after October 1, 1981, is 5%. Id. § 1087dd(c)(1)(D).

101. For example, the interest rate for Stafford Loans dispersed between October 1, 1998, and July 1, 2006, was 2.3% plus a 91-day Treasury bill rate, capped at a maximum rate of 8.25%. Id. § 1087E(b)(6).

The interest rate for federal Perkins Loans is established by 20 U.S.C. § 1087DD(c)(1)(D). The interest rates for DLP Loans are established by 20 U.S.C. § 1087E(b)(7). Maximum interest rates for federal guaranteed loans made under the Federal Family Education Loan Program (FFELP), including PLUS Loans, were established under 20 U.S.C. § 1077A (“applicable interest rates”). Lenders were permitted to charge less than the maximum, but rarely did so. Glater, supra note 35, at 40 & n.139.

When student loan rates were variable, the most commonly used reference rate was the 91-day (3-month) Treasury bill rate, although the 52-week (1 year) Treasury bill rate was also used. The interest rate the borrower actually paid might have in some instances been lower because of interest subsidies described in 20 U.S.C. § 1078. Students can convert their variable-rate loans into fixed-rate loans through consolidation.

The same terms and conditions generally apply to loans made under the FFELP and DLP. 20 U.S.C. § 1087E(a)(1).
IV. Higher Education and Labor Market Coordination Problems

Resources are increasingly allocated to U.S. educational institutions through a market-based process. Students with admissions offers from multiple schools resemble customers who pay for the education they receive and can choose from several different options. Educational institutions, like suppliers in any competitive market, offer discounts to preferred customers. Students with financial resources or access to credit decide where they will study, and what they will study. This market-based approach is consistent with values that emphasize autonomy for individual students and political independence for academic institutions.

However, this freedom may come at a steep price to employers, lenders, and ultimately to the students themselves. As discussed above, government support for mass higher education has always been intended to supply skilled labor, boost economic growth, and encourage social mobility through increased wages and employment. However, skewed incentives and information asymmetries have increasingly shifted educational resources away from human capital investment and toward present consumption.

A. Students as Customers Create Pressure to Reduce Academic Standards

The student-as-customer approach creates pressures toward grade inflation and lower educational standards. Because

102. See David D. Dill, Allowing the Market to Rule: The Case of the United States, 57 Higher Ed. Q. 136 (2003) (noting the United States’ distinctly market-based approach to higher education, and expressing concerns that the United States’ market-based approach may lead to inefficiency because of information asymmetries and reliance on reputation as a gauge of quality).

103. Warren, Baum & Sitaraman, supra note 65, at 128 (“Over the past decade, the federal government, state governments, and colleges and universities have all directed increasing portions of their funds toward high-achieving middle- and upper-income students in order to influence their choices about where to go to college.”).

104. See generally supra Part II.

105. See infra notes 107–10. The same dynamic observed in U.S. higher
universities depend on enrollments for revenue, and students decide where to enroll, administrators and faculty have strong incentives to ensure that students enjoy their time at the university. Universities can and do cater to students’ appetites by offering amenities such as luxury dorms and athletic facilities—amenities many students appear to value more than good instruction. Emphasis on keeping the student-customer happy also extends into the classroom.

One of the ways that universities can encourage faculty to focus on student enjoyment is by linking departmental funding and professors’ promotion to student enrollment numbers and course evaluations. Professors can increase enrollments and


106. See Brian Jacob, Brian McCall, & Kevin Stange, The Consumption Value of Postsecondary Education 33 (2011), http://www.rand.org/content/dam/rand/www/external/labor/seminars/adp/pdfs/2011/stange.pdf (presenting evidence from student enrollment decisions of the high school classes of 1992 and 2004 that most students are more willing to pay for spending on amenities like student activities, sports, and dormitories than on college instruction, but noting that high-achieving students tended to focus more on academic quality).


[T]he effects of market competition on academic behavior compromise the capacity of universities to maintain and improve academic standards. . . . Many universities have responded to the more competitive market by linking academic promotion to student evaluations of teachers and tying departmental budget allocations to student enrolments [sic]. [This] provides the opportunity for instructors to increase the demand for their individual courses and programmes by inflating grades and/or lowering academic standards rather than by actually improving student learning.
RISK-BASED STUDENT LOANS

boost course evaluations by assigning better grades for less work. Students actively shop for classes with professors who give generous grades.¹⁰⁸ Students also give better course evaluations to professors who grade more generously and who flatter students—and worse evaluations to professors who demand more work and more substantive learning.¹⁰⁹

Presumably, many professors respond to such incentives by reducing the rigor of their classes and inflating grades.¹¹⁰ The result is that student-customers study less, learn less,¹¹¹ and are more satisfied with the experience—at least until it is time to

¹⁰⁸ Talia Bar, Vrinda Kadiyali & Asaf Zussman, Grade Information and Grade Inflation: The Cornell Experiment, 23 J. ECON. PERSP. 93, 101–02 (2009) (presenting evidence “that the provision of [average] grade [and grade distribution] information [to Cornell students] led to increased enrollment into leniently graded courses” and that students of high ability were less likely than students of lower ability to pursue courses with more lenient grading).


¹¹⁰ See Donald L. Crumbley et al., What Is Ethical About Grade Inflation and Coursework Deflation?, 8 J. ACAD. ETHICS 187, 187 (2010) (arguing that course evaluations have “caused grade inflation, coursework deflation, and a reduction in student learning as a result of unethical behavior of professors and administrators”); Kiridaran Kanagaretnam et al., An Economic Analysis of the Use of Student Evaluations: Implications for Universities, 24 MANAGERIAL & DECISION ECON. 1, 1–13 (2003) (stating that excessive weight on student evaluations can have negative consequences); David A. Love & Matthew J. Kotchen, Grades, Course Evaluations, and Academic Incentives, 36 E. ECON. J. 151, 151 (2010) (modeling professor behavior and suggesting that increased emphasis on course evaluations can lead to grade inflation); Charles E. Snare, Implications of Considering Students as Consumers, 45 C. TEACHING 122, 122 (1997) (stating that the student-as-customer approach leads to grade inflation, reduced rigor, and less substantive learning); James J. Wallace & Wanda A. Wallace, Why the Costs of Student Evaluations Have Long Since Exceeded Their Value, 13 ISSUES IN ACCT. EDUC. 443, 445 (1998) (same); see also Brenda S. Sonner, A is for “Adjunct”: Examining Grade Inflation in Higher Education, 76 J. EDUC. BUS. 5, 7 (2000) (presenting evidence that adjunct instructors give higher grades than full-time faculty and suggesting that this may be because adjuncts face greater pressure to obtain high course evaluations so that their teaching contract will be renewed).

¹¹¹ See generally Philip Babcock, Real Costs of Nominal Grade Inflation? New Evidence from Student Course Evaluations, 48 ECON. INQUIRY 983 (2010) (providing evidence that higher nominal grades (i.e., grade inflation) can dramatically reduce student effort and study time).
enter the labor market. Potential employers may not be so satisfied with the quality of the education job applicants have received, and recent graduates may not be very satisfied with the employment opportunities that are available to them.112

B. Educational Institutions May Have Incentives to Funnel Students into Areas That Do Not Maximize Students’ Future Incomes or Employment Prospects

Some college majors are more challenging than others. Grade inflation is generally more prevalent in humanities and social sciences and less prevalent in science, technology, engineering, and mathematical subject areas (STEM).113 STEM majors spend

112. BYRON AUGUSTE ET AL., MCKINSEY GLOBAL INST., AN ECONOMY THAT WORKS: JOB CREATION AND AMERICA’S FUTURE 57 (2011) (“[E]mployers still have trouble finding workers with specific skills. And many students lack a clear picture of which jobs and skills will be in high demand.”); ERIN SPARKS & MARY JO WAITS, NAT’L GOVERNORS ASS’N CTR. FOR BEST PRACTICES, DEGREES FOR WHAT JOBS? RAISING EXPECTATIONS FOR UNIVERSITIES AND COLLEGES IN A GLOBAL ECONOMY 8 (2011)

Currently, businesses and states are not getting the talent they want—and students and job seekers are not getting the jobs they want. There are problems with quality. For instance, employers responding to a recent survey estimated that 40 percent of college graduates available to them do not have the necessary applied skills required to meet their needs.

113. See VALEN E. JOHNSON, GRADE INFLATION: A CRISIS IN COLLEGE EDUCATION (2003) (reporting the most generous grading at Duke University in the humanities, intermediate grading in the social sciences, and the most stringent grading in sciences and engineering); Alexandra C. Achen & Paul N. Courant, What Are Grades Made Of?, 23 J. ECON. PERSP. 77, 81–82, 90 (presenting evidence that science and math classes at the University of Michigan generally offer lower average grades than social science and humanities classes, particularly for introductory level, required classes); Patrick D. Larkey, Comment: An Alternative to Traditional GPA for Evaluating Student Performance, 12 STAT. SCI. 269, 270 (1997) (“The few studies that have been done all indicate that there has been relatively more grade inflation in ‘softer’ subjects.”); Kevin Rask, Attrition in STEM Fields at a Liberal Arts College: The Importance of Grades and Pre-collegiate Preferences, 29 ECON. EDUC. REV. 892, 894 (2010) (“Grades given in the sciences are often among the lowest.”); Richard Sabot & John Wakeman-Linn, Grade Inflation and Course Choice, 5 J. ECON. PERSP. 159, 161–63 (1991) (presenting data from Williams College showing that humanities departments generally give the highest grades while science, math, and
more hours studying and have fewer hours for paid work or leisure,114 and often take longer to complete their degrees.115 Yet STEM majors receive lower grades. A large proportion of incoming students report that they intend to study a STEM field. But before graduation, many students switch from STEM to the humanities or social sciences.116 Students switch even though their future employment prospects and wages might be higher if they majored in certain select STEM fields.117 This flight from economics departments generally give lower grades); see also, e.g., ARCIDIACONO ET AL., supra note 17, at 19 (presenting additional evidence from Duke University that grades are higher in the humanities and social sciences than in natural sciences, engineering, and economics).

114. See NATIONAL SURVEY OF STUDENT ENGAGEMENT, ANNUAL RESULTS 15 (2011) (reporting that engineering and science majors study more hours per week than humanities, education, or business majors, and that business, social science, and humanities majors spend more hours per week working).


117. See Mark C. Berger, Cohort Size Effects on Earnings: Differences by College Major, 7 ECON. EDUC. REV. 375, 381 (1988) (reporting that wages for science and liberal arts majors are depressed more by an increase in the size of their college cohort than are wages for engineering or business majors); Mark C. Berger, Predicted Future Earnings and Choice of College Major, 41. INDUS. & LAB. REL. REV. 418, 426 (1988) (reporting relatively high earnings for U.S. business and engineering graduates and relatively low earnings for liberal arts and education graduates); Black et al., supra note 18, at 365 (finding that engineers earn more than economics majors, who earn more than most other social science, business, and humanities majors, and that MBAs with chemical engineering undergraduate majors earn more than other MBAs and that "the measured differentials reflect in part real differences in the market returns to different fields of study"); Charlotte Christiansen et al., The Risk Return Trade-off in Human Capital Investment, 14 LABOUR ÉCON. 971, 984–85 (2007) (reporting relatively high risk-adjusted returns for engineering and health sciences degrees and relatively low risk-adjusted returns for humanities and education degrees for a sample of Danish graduates); Scott L. Thomas, Deferred Costs and Economic Returns to College Major, Quality, and Performance, 41
STEM is likely due at least in part to university-created grading incentives.  

However, not all STEM graduates fare well. There is evidence of an “oversupply” of science Ph.Ds seeking professorships within universities, and declining pay and working conditions for Ph.Ds in many fields. See B. Lindsay Lowell & Hal Salzman, Into the Eye of the Storm: Assessing the Evidence on Science and Engineering Education, Quality and Workforce Demand 2 (2007), http://www.urban.org/UploadedPDF/411562_salzman_Science.pdf (presenting evidence that there are three times as many U.S. science and engineering graduates as available job openings in these fields); B. Lindsay Lowell et al., Paper Presented at the Annual Meetings of the Association for Public Policy: Steady as She Goes? Three Generations of Students through the Science and Engineering Pipeline 31–32 (Nov. 7, 2009) (presenting evidence that starting in the early 1990s, top performing students increasingly opted out of science and engineering employment) (on file with the Washington and Lee Law Review); Beryl Lieff Benderly, The Real Science Gap, MILLER-MCCUNE, July/Aug. 2010, at 30, 33 (“[B]ecoming a scientist now entails a penurious decade or more of graduate school and postdoc positions before joining the multitude vainly vying for the few available faculty-level openings.”); Romer, supra note 116, at 241 (arguing that there is a glut of scientists within universities but a shortage of scientists who can work in industry).

118. See Valen E. Johnson, An Alternative to Traditional GPA for Evaluating Student Performance, 12 STAT. SCI. 251, 251 (1997) (“[D]ifferences in grade distributions result in a substantial reduction in the number of courses taken by students in subjects like mathematics and the natural sciences, as well as other challenging upper-level undergraduate courses.”); Sabot & Wakeman-Linn, supra note 113 (providing evidence that students respond to grades as incentives and choose to study humanities rather than science and math, even though there is likely to be greater employment opportunity in STEM fields, because universities offer higher average grades and a narrower grading distribution in humanities classes); Larkey, supra note 113, at 270

[T]here has been relatively more grade inflation in “softer” subjects. Grade inflation has become an important edge for some fields in competing for students as core curricula have waned and student choices have waxed. It is a perverse form of price competition; they have been able to offer higher grades for equivalent or lesser amounts of work.
Figure 5.1: Compared to Other High-Income Countries, the U.S. Produces Relatively Few STEM Degrees

STEM college degrees as percent of total by country, 2008

<table>
<thead>
<tr>
<th>Country</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korea</td>
<td>33%</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>33%</td>
</tr>
<tr>
<td>Germany</td>
<td>29%</td>
</tr>
<tr>
<td>France</td>
<td>27%</td>
</tr>
<tr>
<td>Austria</td>
<td>27%</td>
</tr>
<tr>
<td>Finland</td>
<td>27%</td>
</tr>
<tr>
<td>Japan</td>
<td>24%</td>
</tr>
<tr>
<td>Sweden</td>
<td>24%</td>
</tr>
<tr>
<td>Switzerland</td>
<td>23%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>23%</td>
</tr>
<tr>
<td>Italy</td>
<td>22%</td>
</tr>
<tr>
<td>Canada</td>
<td>22%</td>
</tr>
<tr>
<td>Ireland</td>
<td>22%</td>
</tr>
<tr>
<td>Belgium</td>
<td>20%</td>
</tr>
<tr>
<td>New Zealand</td>
<td>20%</td>
</tr>
<tr>
<td>Denmark</td>
<td>19%</td>
</tr>
<tr>
<td>Australia</td>
<td>19%</td>
</tr>
<tr>
<td>Norway</td>
<td>15%</td>
</tr>
<tr>
<td>United States</td>
<td>15%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>14%</td>
</tr>
<tr>
<td>Iceland</td>
<td>14%</td>
</tr>
</tbody>
</table>

Source: Organization for Economic Cooperation and Development, Graduates by Field of Education.

See also Ost, supra note 116, at 923–34 (presenting evidence that science majors are pushed away by low grades in their major and pulled away by higher grades in classes in other fields); Rask, supra note 113, at 892–900 (estimating that roughly 2% to 4% more students would complete STEM education if the grading distribution in STEM fields and liberal arts were equal).
Figure 5.2: Students Who Take Courses in High Value Fields Receive Lower Grades, Especially in the Early Years of College When They Select a Major, Even Though . . .

Grades by course type and school year for a sample of Duke undergraduates
Noncumulative within-year grade point average

Figure 5.3: Students in High Value Fields (Other Than Business) have slightly higher ability levels and ...

SAT scores by pre-college intended major and completed major for a sample of Duke undergraduates

Mean SAT points

- Humanities / Social Sciences (excluding economics) degree completers
- Natural Science / Engineering / Economics degree completers

Figure 5.4: Students in High Value Fields (Other Than Business) Spend More Time Studying

**Hours per week spent preparing for class by full time college seniors, by major**

- Engineering: 19 hours
- Physical sciences: 18 hours
- Biological sciences: 17 hours
- Arts & humanities: 17 hours
- Education: 15 hours
- Social sciences: 14 hours
- Business: 4 hours

Source: Indiana University Center for Postsecondary Research, National Survey of Student Engagement 15 (2011) Figure 7.

**Percent of full time seniors who spend more than 20 hours per week preparing for class, by major**

- Engineering: 42%
- Physical sciences: 36%
- Biological sciences: 34%
- Arts & humanities: 31%
- Education: 26%
- Social sciences: 23%
- Business: 19%

Source: Indiana University Center for Postsecondary Research, National Survey of Student Engagement 15 (2011) Figure 8.
Why might universities wish to use high grades and low workloads to channel students away from fields that are in demand in the labor market? In some cases, it may not be a university policy so much as a series of decisions by individual departments or professors.

One less than completely satisfying explanation is that professors who teach “softer,” more subjective material find it more difficult to distinguish between students of high and low
A higher cost to professors of sorting students by ability would explain a narrower grading distribution, but not necessarily higher average grades or lower workloads. Some STEM professors might wish to “weed out” weaker students because of concern about the harm such students might cause if they graduate with satisfactory grades and then err on the job—imagine a poorly designed bridge—whereas liberal arts professors may be less concerned about the quality of graduates engaged in nonlethal endeavors. Of course, low grades, large class sizes, and limited academic support may deter even some high-ability students from pursuing STEM educations, and many STEM careers involve little risk to public safety.

Another possibility is that humanities and social sciences professors feel compelled to compete for student enrollments by offering better grades for less work, because they cannot compete by offering equally attractive post-graduation employment opportunities. STEM professors may feel less pressure than humanities professors to compete for student enrollments because STEM professors can generate revenue for the university by bringing in research grants or developing commercially useful patents.

But why do universities tolerate such interdepartmental grade-based competition for enrollments? Universities could standardize grading by creating a mandatory grading curve across disciplines or requiring professors to rank students on a

119. See Achen & Courant, supra note 113, at 78 (“Departments that evaluate student performance using interpretive methods will tend to have higher grades, because using these methods increases the personal cost to instructors of assigning and defending low grades.”); id. at 87–88; Larkey, supra note 113, at 270 (“There is apparently more resistance to [grade] inflation in domains with more sharply and logically defined right and wrong answers.”).

120. See Achen & Courant, supra note 113, at 78 (“Grades can be used in conjunction with other tools to attract students to departments that have low enrollments and to deter students from courses of study that are congested.”); Larkey, supra note 113, at 270.

121. Bok, supra note 10, at 74 (“Teaching loads... have dropped furthest... in scientific disciplines, such as mathematics and experimental physics, where the interests of undergraduates have steadily given way to the demands of pure research.”); id. at 140–42 (discussing efforts to commercialize university-developed technologies).
percentile scale. Universities could make grading even more meaningful by using rankings adjusted for the difficulty of the course and the level of competing students. Indeed, one such proposal—complete with a workable methodology—was put forth by a professor of statistics at Duke University, considered, and ultimately rejected by his university.

One possibility is internal university politics and interdepartmental rivalries. Another possibility is that such changes are simply not in the overall interests of universities.
Instruction in any area with attractive career opportunities outside academia—whether in medicine, law, business, or certain STEM fields—will be relatively expensive. Those with talent in such fields have attractive employment opportunities in the private sector or government, and those who agree to forgo those opportunities to teach at universities must therefore be offered higher wages than other professors.126

Undergraduate students usually pay the same tuition price per credit regardless of what they study,127 so students who are willing to study subjects with limited value in the labor market channel students away from classes that are overcrowded relative to supply of qualified instructors and toward courses in which teaching capacity is underutilized).

126. See, e.g., Richard B. Freeman, The Market for College-Trained Manpower: A Study in the Economics of Career Choice 165–67 (1971) (documenting shortages of faculty in high-demand fields in the United States in the 1960s when academic norms called for roughly equal pay across disciplines); Gary Rhodes, Managed Professionals: Unionized Faculty and Restructuring Academic Labor 75 (1998) (“The pattern of salary dispersion by field suggests that academic managers have been willing and able to respond to field-defined labor markets.”); Ronald Ehrenberg, Hirschel Kasper & Daniel Rees, Faculty Turnover at American Colleges and Universities: Analyses of AAUP Data, 10 Econ. Educ. Rev. 99, 99–110 (1991) (presenting data confirming the importance of salary to faculty retention); Faculty Salaries Vary by Institution Type, Discipline, Chron. of Higher Educ., Apr. 11, 2011 (presenting data showing that faculty salaries are highest in law, business, economics, computer science, engineering, and health science, and lowest in performing arts, education, and the humanities); Mike Horsley, et al., Dept of Educ., Emp’t & Workplace Relations, Salary Relatives and the Academic Labour Market 47–79 (2003) (documenting Australian universities’ difficulty hiring faculty when candidates were qualified for more highly paid private sector work, especially in finance and business, science, engineering, and information technology).

127. A few public universities charge a higher price for courses or majors in high cost, high value areas such as business or engineering, in part because these areas are expensive and state legislatures have refused to approve general tuition increases. There is some anecdotal evidence suggesting that differential tuition has reduced the likelihood that low-income students will study toward high value degrees. Universities that charge differential tuition or fees include the University of Wisconsin, Rutgers, the University of Illinois, the University of Kansas, and the University of Nebraska. See Jonathan D. Glater, Certain Degrees Now Cost More at Public Universities, N.Y. Times, July 29, 2007, http://www.nytimes.com/2007/07/29/education/29tuition.html?pagewanted=all&r=0 (last visited Feb. 3, 2013) (on file with the Washington and Lee Law Review).
may be more profitable to teach than students who insist on instruction with higher value in the labor market (i.e., engineering, business, or health science). Universities can use grades or higher student-to-faculty ratios to channel students away from fields in which instruction is expensive.

Higher student-to-faculty ratios in STEM fields detract from the quality of instruction in these fields. Students in STEM majors are less likely to complete their degrees in four or five years, and frequently complain about inadequate academic support.

As discussed in greater detail below, the government could counteract these perverse incentives through risk-based pricing, which would create constructive financial incentives for both students and universities. Students who persisted in challenging, high-value majors would be rewarded with lower interest rates and a lower total cost of education. Universities that channeled more students toward fields with high value in the labor market would find their “institutional interest rate” decrease, which would encourage prospective students to matriculate and increase university revenues. By contrast, students or universities that devoted resources toward low-value instruction would find their interest rates increase and their resources shrink.

128. See Bok, supra note 10, at 41 (“Universities may offer worthless instruction because of . . . financial pressure.”).

129. See Achen & Courant, supra note 113, at 78, 89 (arguing that universities adjust their grading distributions to channel students away from classes that are overcrowded relative to supply of qualified instructors and toward courses in which teaching capacity is underutilized); Freeman, supra note 125, at 344 (same).

130. See Hurtado et al., supra note 115 (stating that STEM majors have lower rates of degree completion than other college majors); Assia Boundauli, Why Would-be Engineers End Up as English Majors, CNN (May 21, 2011, 10:16 PM), http://www.cnn.com/2011/US/05/17/education.stem.graduation/index.html (last visited Feb. 3, 2013) (“Science and math programs are designed and taught to winnow down the number of students. University tenure systems often reward professors who conduct research and publish their work, but not those who teach well.”) (on file with the Washington and Lee Law Review); Christopher Drew, Why Science Majors Change Their Mind, N.Y. TIMES, Nov. 6, 2011, at ED16 (citing difficult coursework, intense competition, and grade inflation in non-STEM majors as reasons for the high attrition rate in STEM degree programs).
The federal government could also help universities obtain the resources they need to hire more qualified instructors in high demand, labor-market-relevant fields by boosting federal student loan limits. At a minimum, loan limits could be increased for students who are studying toward high-income, low unemployment careers, so that universities can obtain the revenues they need to offer competitive salaries to qualified instructors in these fields.\(^\text{131}\)

But if universities are student-customer driven, why do students not already demand more instruction in fields that will prepare them better for the labor market? One possibility is that students prefer to enjoy a leisurely college experience and are relatively indifferent to the long-term financial consequences.\(^\text{132}\)

A second possibility is that liberal arts majors correctly perceive that maximizing their undergraduate GPA will improve their chances of admission to an elite graduate program that will ultimately boost their long term income, and are more willing or able than STEM majors to delay entering the labor force.\(^\text{133}\) And indeed, some labor economists have suggested that studies


\(^{132}\) See supra note 97 (presenting evidence that many students place high value on leisure and personal enjoyment both during and after college).

\(^{133}\) Eric Eide & Geetha Waehrer, The Role of Option Value of College Attendance in College Major Choice, 17 ECON. EDUC. REV. 73, 73 (1998) ("Anecdotal evidence suggests that many students choose to major in fields never intending to terminate their education with an undergraduate degree, but rather they intend to enroll in professional or academic graduate programs."); Kimberly A. Goyette & Ann L. Mullen, Who Studies the Arts and Sciences? Social Background and the Choice and Consequences of Undergraduate Field of Study, 77 J. HIGHER EDUC. 497, 524–27 (2006) (reporting that lower socioeconomic status students are more likely to choose vocational majors and more likely to be employed in higher paying jobs four years after graduation, while higher socioeconomic status liberal arts majors are more likely to enroll in graduate school, which may lead to higher long-term incomes); Johnson, supra note 118, at 251–52 ("For those students whose primary objective is to gain admittance to medical school or law school . . . 'grade shopping' may represent an optimal career strategy.").
focusing on wages and employment at graduation undervalue humanities and social science degrees relative to engineering and business degrees because they ignore the option value nonvocational degrees confer by facilitating enrollment in graduate school.\textsuperscript{134} A third possibility is that students do not have good information about the labor market until it is too late.

\textbf{C. Many Students May Not Understand the Connection Between Their Chosen Course of Study and Future Income and Employment Prospects}

Students—who are by definition still learning about their chosen field—may not have accurate information about post-graduation employment prospects and wages in their own or other fields. Empirical studies have documented systematic mistakes in undergraduate students’ perceptions about prospective wages by college major and occupation. According to

\begin{quote}
\textsuperscript{134} Eide & Waehrer, \textit{supra} note 133, at 74, 77 (arguing that humanities and social science degrees may prepare students for a wider range of graduate schools than engineering, computer science, or business degrees and therefore have higher option value). The investigators only analyzed the probability of attending graduate school conditional on a certain choice of major, not the type of graduate school attended or wages after completion of graduate school.

The assumption that social science and humanities degrees are undervalued on a relative basis may not hold in the absence of grade inflation (which may improve chances of admission to elite programs) or with more nuanced analysis of different types of graduate degrees or post-graduate-school wages. At least with respect to professional schools, graduate students with undergraduate backgrounds in business, economics, or STEM seem to have considerable advantages. \textit{See, e.g.}, Black et al., \textit{supra} note 18, at 365–66 (finding that economics majors make more money than most other majors when they enter the work force directly after graduation, and economics majors who attend law or business schools have higher wages than those who majored in other fields, except for MBAs with undergraduate degrees in Chemical Engineering).

Law students with technical backgrounds may find it easier to specialize in high-wage, high-demand areas of legal practice, such as patents, tax, bankruptcy, commercial law, healthcare regulation, energy regulation, financial regulation, or other technical fields. Business students who can manage engineers, evaluate technical companies, or analyze large data sets seem to be in higher demand than those with soft skills such as communications or marketing. And medical school requires that students have scientific backgrounds.
\end{quote}
one study of undergraduates at the University of California at San Diego, students typically mistake expected wages by 20%.\textsuperscript{135} According to another study of male undergraduates at Duke, students typically overestimate wages in their own field, and 7.5% of students would switch majors if they optimally forecasted wages.\textsuperscript{136} It should be noted that this study controlled for the influence of students’ abilities and career preferences on choice of major.\textsuperscript{137}

Studies also suggest that students generally learn about labor market prospects a year or so before graduation—too late to easily change majors.\textsuperscript{138} Students are better informed of starting salaries than about potential increases,\textsuperscript{139} and might overestimate the importance of starting salary relative to lifetime earnings. Disconcertingly, there is evidence that the students who know the least about major and occupational wage differences are those from poor families.\textsuperscript{140} This is not due to lower ability levels of less wealthy students; the investigators suggest it may be because such students have fewer college-educated relatives and friends who can inform them about the

\begin{itemize}
\item [Students think the market premium is higher for their own major compared to those for majors they did not choose. . . . We estimate that over 7.5\% of students would switch majors if this forecast error was not present. Thus, our results suggest an important role for informational differences in modeling the choice of college major.]
\item \textsuperscript{137} See id. (“Our model-based estimates clearly indicate that expected earnings do matter for student’s choice of major, even after controlling for ability and career preferences.”).
\item \textsuperscript{138} See Betts, supra note 135, at 47–48 (reporting that students do not acquire most of their knowledge about wages until their final year of education).
\item \textsuperscript{139} See id. at 39 (“Students’ knowledge of salaries of younger workers is quite good, but becomes progressively worse as the experience of the worker in question increases.”).
\item \textsuperscript{140} See id. at 37–38, 43. All students surveyed were enrolled at UCSD. \textit{Id.} at 29. The authors controlled for ability as measured by GPA, and even after controlling for GPA, students from lower income families were less knowledgeable about occupational wages. \textit{Id.} at 35–36, 43.
\end{itemize}
labor market for college-educated workers. Students are not simply uninformed, but in some instances, they are actively misled by aggressive and deceptive recruiting efforts.\textsuperscript{141}

Studies have demonstrated that students choose their major based largely—but not exclusively—on expected post-graduation wages.\textsuperscript{142} Many students would switch to majors linked to higher


It is unclear if marketing abuses also occur at non-profit institutions. Sales and marketing efforts appear to be more aggressive in the for-profit sector, which spends an average of 23% of revenue on sales and marketing compared to 0.5% in the nonprofit sector. See Emma Roller, \textit{Senate Bill Would Bar Colleges from Using Federal Student Aid for Marketing}, CHRON. OF HIGHER EDUC., Apr. 18, 2012.

A study by the Government Accountability Office suggested that after controlling for student characteristics, graduates of for-profit institutions generally have worse outcomes, although some for-profit programs performed well. See \textit{GOV’T ACCOUNTABILITY OFF., POSTSECONDARY EDUCATION: STUDENT OUTCOMES VARY AT FOR-PROFIT, NONPROFIT, AND PUBLIC SCHOOLS} 5–8 (2011) (comparing graduation rates, employment outcomes, student loan debt, default rates, and licensing exam results of students graduating from for-profit, nonprofit, and public institutions).

\textsuperscript{142} See Richard J. Cebula & Jerry Slopes, \textit{Determinants of Student Choice of Undergraduate Major Field}, 19 AM. EDUC. RES. J. 303, 303 (1982) (presenting evidence that “earnings differentials among fields and differences in the rate of change in earnings among fields are the most important factors in the student’s decision”); Claude Montmarquette, Kathy Cannings & Sophie Mahseredjian, \textit{How Do Young People Choose College Majors?}, 21 ECON. EDUC. REV. 543, 554
post-graduation wages if students had more accurate and timely information. Expected wages play a particularly important role in major choice for students from low socioeconomic status backgrounds. Choosing the right major could help poor students use education to achieve upward socioeconomic mobility.

Recently, there have been several attempts to make the education and labor markets more transparent. These include data supplied for free by the U.S. Departments of Labor and Education, new “scorecards” and “shopping sheets” for educational institutions, and entrepreneurial efforts to make labor market information more accessible.

As discussed in greater detail below, risk-based pricing of student loans could help make wages and employment prospects more transparent and salient to students at an earlier stage in their educational careers.

(2002) (“[C]hoice of college concentration depends decisively on the expected earnings in a particular concentration.”). But see Peter Arcidiacono, Ability Sorting and the Returns to College Major, 121 J. ECONOMETRICS 344, 372 (2004) (finding that the wage premium for certain college majors does not drive students of higher ability to select those majors as much as student subject matter preferences).

143. See Arcidiacono, Hotz & Kang, supra note 136, at 2–3 (stating that many students would switch their majors if they had more accurate salary information).

144. See Kimberly A. Goyette & Ann L. Mullen, supra note 133, at 524–27 (reporting that low socioeconomic status students are more likely to choose vocational majors and more likely to be employed in higher paying jobs four years after graduation, while higher socioeconomic status liberal arts majors are more likely to enroll in graduate school); Yingyi Ma, Family Socioeconomic Status, Parental Involvement, and College Major Choices—Gender, Race/Ethnic, and Nativity Patterns, 52 SOC. PERSP. 211, 211 (2009) (finding that lower socioeconomic students favor more lucrative majors compared to higher socioeconomic status students).


D. Skilled Labor Markets Suffer from Periodic Booms and Busts

Students generally do not have good information with which to forecast the likelihood of employment in a given field. Without a good forecast of decisions by employers (the demand side) and knowledge of how many other students are studying toward a particular field (the supply side), such prediction is difficult if not impossible. Instead, students appear to assume that the future will resemble the present.

This assumption and the long production lag for skilled labor leads to boom and bust cycles in the labor market known as “cobweb” cycles. At the start of the cycle, many students seek to study toward a high-income occupation. Years later, when they all simultaneously try to enter the labor force, the large supply of labor causes wages to crash in their occupation. In the second stage of the cycle, students choosing an occupation at the time of the crash then avoid training for the newly low income occupation, and years later, there will be a shortage of labor for that occupation, causing wages to rise and the cycle to repeat.

Cobweb cycles have been demonstrated in a wide range of skilled labor markets, including markets for engineers, lawyers, scientists, and professors. The number of years in the cycle is

147. See infra note 148 and accompanying text.

typically proportional to the number of years of advanced training required to enter the profession.

E. Skills Mismatches Reduce Employment and Output

Mismatch between skills and employment opportunities lead to an inefficient reduction in productivity and higher unemployment. Economists at the Federal Reserve have estimated that the mismatch between worker skills and employer needs accounts for more than one-fourth of unemployment in the United States. When employment opportunities are available but workers lack the required skills, the resulting reduction in employment is called “structural unemployment.”

As will be discussed below, even educational institutions themselves may not always have good information about the labor market or the resources to acquire that information. Some of the best information, such as confidential tax or social security records—which may have more accurate wage data than surveys—may only be available to or acquirable by the government.

Because the U.S. government’s role in the higher education market is primarily as a lender, risk-based pricing of federal student loans may offer the best approach to reconcile traditional academic values with the goals of a more transparent and

149. See Auguste et al., supra note 112, at 57 (“[E]mployers still have trouble finding workers with specific skills. And many students lack a clear picture of which jobs and skills will be in high demand.”); Robert Shimer, Mismatch, 97 Am. Econ. Rev. 1074, 1074 (2007) (arguing that geographic attachment and skills mismatch contribute to structural unemployment).

150. See Narayana Kocherlakota, President, Fed. Res. Bank of Minneapolis, President’s Speech at Missoula, Montana: Back Inside the FOMC (Sept. 8, 2010), available at http://www.minneapolisfed.org/news_events/pres/speech_display.cfm?id=4532 (last visited Feb. 3, 2013) (“How much of the current unemployment rate is really due to mismatch? The answer seems to be a lot. [With better matching] we would have an unemployment rate closer to 6.5 percent, not 9.6 percent.... [O]ver 2.5 percentage points of the current unemployment rate is attributable to mismatch.”).

efficient labor market, higher wages, and lower structural unemployment.

V. The Theory of Risk-Based Credit Pricing

Lenders are not like ordinary retail merchants. Whereas a merchant can complete a transaction at the time of sale, a lender will have an ongoing relationship with the borrower for the entire duration of the loan. The lender continues to bear risk that at some point the borrower will become unable to repay and will default on the debt.152

Risk-based credit pricing involves adjusting the interest rate on loans so that the interest rate compensates the lender not only for the time value of money,153 but also for the risk that borrowers will default on their debts and cause the lender to incur losses.

152. Default refers to late payment or partial or complete failure to repay. Defaults can be distinguished from one another by their severity, or the losses to the lender in the event of default, which can range from 0% to 100% of the loan value. See Understanding Default, FEDERAL STUDENT AID, http://studentaid.ed.gov/repay-loans/default (last visited Feb. 3, 2013) (defining and explaining default and its consequences) (on file with the Washington and Lee Law Review).

153. The time value of money is the idea that the value of money is higher in the present than in the future, either because money today could be invested to produce profit in the future, or because people tend to prefer instant gratification and would generally rather consume the same good or service now rather than in the future. See Michael Simkovic & Benjamin Kaminetzky, Leveraged Buyout Bankruptcies, the Problem of Hindsight Bias, and the Credit Default Swap Solution, 2011 COLUM. BUS. L. REV. 188, 188 n.214 (explaining the time value of money concept).

The time value of money is broader concept than inflation, which refers to general changes in price levels, so that a dollar at one point in time will not purchase the same basket of goods and services in the future. Inflation is generally positive (i.e., the value of money declines over time). However, even with zero inflation, there would still be a positive time value of money because of the prospect of real (after-inflation) profitable investments and because of preferences for present consumption.

The time value of money is typically measured as the interest rate charged to a hypothetical debtor with no risk of defaulting, also called the risk-free rate. Although no debtor is completely risk-free, the risk-free rate for dollar-denominated loans is typically assumed to be the rate of interest on U.S. government securities. See id. at 188–89; Joost Driessen, Is Default Event Risk Priced in Corporate Bonds?, 18 REV. FIN. STUD. 165, 169 (2005) (noting that U.S. government securities are considered to be risk-free).
From the perspective of risk-based credit pricing, uniform student loan pricing is a redistributive policy. Uniform pricing subsidizes the riskiest borrowers while profiting from the safest borrowers. In the student loan context, uniform credit pricing is a subsidy to students who are studying fields with the lowest value in the labor market and a tax on students who are studying fields with the highest value in the labor market and the best employment prospects. This subsidy creates perverse incentives—discouraging the most able students and most economically valuable programs, while encouraging the highest risk and least economically valuable programs.

Risk-based student loan pricing should reduce moral hazard by forcing student borrowers to internalize the risks created by their own decisions, encouraging students to study toward high-value occupations. Risk-based student loan pricing should reduce adverse selection by discouraging students with poor prospects from borrowing heavily to attend expensive education programs of dubious value, while encouraging the most promising students to borrow what they need to complete valuable degrees. Moreover, risk-based pricing would clarify the differential economic value of different courses of study, and help students make choices that are in their own long-term best interests.

Over time, risk-based student loan pricing should cause colleges to shift educational resources toward teaching subjects

---

Treasury bonds are assumed to be default-free); Francis A. Longstaff, Sanjay Mithal & Eric Neis, Corporate Yield Spreads: Default Risk or Liquidity? New Evidence from the Credit Default Swap Market, 60 J. FIN. 2213, 2223 (2005) (“[T]he Treasury curve . . . is the standard benchmark riskless curve in most empirical tests in finance.”); Robert C. Merton, On the Pricing of Corporate Debt: The Risk Structure of Interest Rates, 29 J. FIN. 449, 449 (1974) (noting that government bonds are essentially default-risk-free). For example, for a 10-year student loan, the yield on a 10-year Treasury bond might be used as the risk-free rate.

154. To be more precise, uniform pricing is a subsidy to those with the highest risk of default and a tax on those with the lowest risk of default. However, because students who need to borrow generally come from middle class backgrounds, and will therefore rely on their future labor income (or their spouse’s labor income) to repay their student loans, risk will in practice generally turn on career prospects. See JAQUELINE E. KING, FINANCING A COLLEGE EDUCATION: HOW IT WORKS, HOW IT'S CHANGING 27 (1999).
and skills that are most valued in the labor market. This should improve the risk profile of borrowers, and reduce student loan defaults and structural unemployment.

Some may argue that wages in the labor market do not reflect the social value of certain occupations. For example, progressives frequently claim that teachers are undercompensated relative to their education levels and social contribution. Indeed, the original NDEA emphasized the importance of training more teachers as well as STEM specialists, and current student loan programs include special loan forgiveness provisions for teachers. By contrast, many economists have suggested that an individual's willingness to sacrifice income to pursue less lucrative education and linked career paths suggests that such education or career paths may be more enjoyable and constitute a form of nonmonetary compensation.

The transition may be gradual because the institution of tenure limits the flexibility of academic staffing. However, to the extent that untenured faculty, adjunct faculty, or graduate students teach classes in low-value fields, universities could adjust and reallocate resources toward higher value fields fairly rapidly.

See, e.g., EILIS LAWLOR, HELEN KERSLEY, SUSAN STEED & MARTIN COTTINGHAM, NEW ECON. FOUND., A BIT RICH: CALCULATING THE REAL VALUE TO SOCIETY OF DIFFERENT PROFESSIONS 27–28 (2009) (arguing that not all highly compensated occupations generate social value); Warren, Baum & Sitaraman, supra note 65, at 132–36 (advocating the value of military and civil service).

See LAWRENCE MISHEL, SYLVIA A. ALLEGRETTI & SEAN P. CORCORAN, ECON. POLY INST., THE TEACHING PENALTY: AN UPDATE THROUGH 2010 (2011) (arguing that teacher salaries are too low to attract top-performing graduates). There is little controversy that teacher compensation in the United States is not sufficient to attract and retain many top-performing college graduates. Whether the current workforce of teachers is “undercompensated” relative to its contributions is a separate question.

See National Defense Education Act of 1958, Title VIII, § 802, Pub. L. No. 85-864, 72 Stat. 1580, 1597–98 (codified as amended in scattered sections of 20 U.S.C.) (“Funds paid to a State under this title for area vocational education programs may be used, in carrying out such programs . . . for . . . maintenance of adequate programs of administration, supervision, and teacher-training . . . .”).


See Annette Alstadsæter, Measuring the Consumption Value of
However, if public sector workers are undercompensated, questions remain about whether public sector compensation should be increased through taxpayer-funded loan subsidies or through a taxpayer-funded increase in starting salaries for college graduates entering public service.

Risk-based pricing does not demand a rigid embrace of laissez-faire. In every developed economy, both markets and governments set wages. If wages for public sector occupations are below socially optimal values, then this reflects a failure by governments to set taxes and spending at appropriate levels. If wages in socially destructive occupations are high, this reflects a failure of regulation and taxation to punish socially destructive behavior and channel profits toward productive activity. Risk-based student loan pricing does not create any of these failures—it simply reflects the current reality of U.S. political economy, and provides information and incentives to help college students navigate a challenging labor market.

To reject risk-based pricing outright in favor of uniform student loan pricing, one would have to believe that all low compensation occupations have positive externalities, all high compensation occupations have negative externalities, and that these externalities precisely match differences in wages and employability. In other words, one would have to believe that wage differentials are not only inefficient, but are the exact opposite of the efficient distribution of wages, and everyone with a college degree should earn the same income.

Furthermore, one would have to assume that supply-side subsidies such as low student loan interest rates are preferable to targeted demand-side subsidies such as higher taxpayer-funded wages for specific occupations.161

---

161. Whereas demand-side subsidies would help workers to capture the
There are many reasons to favor demand-side subsidies (higher wages) over lower interest rates. Loan subsidies would effectively increase public sector pay only for workers with student loans. Public sector employers might try to capture part of this subsidy by reducing entry-level base salary, thereby making public sector careers relatively less attractive to graduates who have no debts. Loan subsidies could unintentionally funnel the children of the poor into low paying public sector careers and affluent graduates into private sector careers with better long-term earnings. Such career sorting risks exacerbating class divisions and reducing social mobility.

By contrast, increased public service starting salaries would equally encourage students with debts and those without to enter public service. This could improve the quality of public service personnel by making such jobs attractive to a wider range of qualified applicants, increasing competition for public service jobs. It could also help prevent public service careers from becoming segregated along class divisions.

If certain occupations are “undercompensated,” then governments should increase compensation using the tax and spending power. Risk-based student loan pricing simply reflects the willingness—or unwillingness—of market participants and governments to pay for services, and channels human capital to where it is valued most. Indeed, governments may have no choice but to increase public sector compensation if college students become savvier about prospective wages in different occupations and refuse to accept low-wage (but supposedly valued) positions.

---


162. See U.S. CONST. art. I, § 8, cl. 1 (“The Congress shall have Power To lay and collect Taxes, Duties, Imposts and Excises, to pay the Debts and provide for the common Defence and general Welfare of the United States . . . .”).
A. A Simple Risk-Based Credit Pricing Model

Consider the following simplified example of risk-based credit pricing, adapted from Michael Simkovic & Benjamin Kaminetzky, Leveraged Buyout Bankruptcies, the Problem of Hindsight Bias and the Credit Default Swap Solution.\(^{163}\)

Assumptions are listed below.

- Each group wants to borrow $100,000
- Each borrower borrows the same amount of money
- There are no transaction costs or administrative costs
- All payments and defaults are made at the end of year 1
- The relevant risk-free rate is 3% per year
- The lender is risk-neutral and not subject to liquidity risk
- The lender can borrow at the risk-free rate
- Group A
  - Borrowers in Group A will repay their loans in full.
- Group B
  - Ten percent of the borrowers in Group B will default on their loans in the first year, while 90% of the borrowers in Group B will repay their loans in full.
  - The Group B borrowers who default will repay 70% of the balance of their loans.

Group A is risk-free. If the lender were to lend to Group A borrowers at 3%, the lender would end the year with $103,000 ($100,000 in principal repayment plus $3000 in interest payments). This would exactly match the lender’s cost of capital and—in the absence of any administrative costs—would cause

the lender to break even. At any price above 3% interest, the lender would make a profit, and at any price below 3%, the lender would lose money.

Lending to Group B involves credit risk, but there is an interest rate at which the lender should be indifferent between lending to Group A or Group B. Specifically, there is an interest rate at which the lender would expect to have the same $103,000 at the end of the year after lending to Group B. The expected value of the loan to Group B is the sum of the repayments by the 90% of creditors who will repay their loans in full and the repayments by the 10% who will default on their debts and only repay 70% of their loans. For the 90% who will repay, the lender will receive $90,000 \times (1 + \text{the interest rate}). For the 10% who will default, the lender will receive $10,000 \times 0.7 \times (1 + \text{the interest rate}). Solving for the interest rate that will enable the lender to earn $103,000, we get ($103,000 - $97,000) / $97,000 = 6.19%.

Under simplified assumptions, the formula for calculating the break-even interest rate could be rewritten as:

$$i = \frac{1 + r}{1 - (D \times L)} - 1$$

(Eq. 1)

in which:

- $i$ = break-even interest rate
- $r$ = risk-free rate
- $D$ = probability of default in year 1
- $(1 - D)$ = probability of no default in year 1
- $L$ = loss rate given default; $(1 - L)$ = recovery rate

164. In practice, the formula for calculating the break-even interest rate will be somewhat more complicated for a number of reasons. First, every loan program will entail some administrative and other transactions costs, and these costs must be recovered either through fees, interest spreads, or some other source of revenue. Second, student loans last for more than one year. The probability of default may be higher in some years than in others, and the loss rate given default for any given loan will generally go down over time as the borrower repays the principal. Third, there is an interaction between the interest rate charged and the probability of default; all else being equal, a higher interest rate will increase the probability that the borrower will default, while a lower interest rate will reduce the probability of default by making payments more affordable.
The break-even interest rate is an important concept because it helps us understand whether a student loan, priced at a particular level, is a budget-neutral transaction, a profit center for the government/taxpayer, or a subsidy to the student borrower. Risk-based pricing does not rely on perfectly accurate predictions for individuals, but rather predictions that more or less hold statistically true for groups of borrowers.

B. Data-Driven Risk-Based Pricing

At a technical level, the question is as follows: what readily observable characteristics can be used, individually or in combination with one another, to explain variation in loan defaults and loss rates? Multivariate statistical methods such as ordinary least squares regression (OLS) can be used to build a model that predicts loan losses.¹⁶⁶

An analyst would load panel data consisting of individual loans into statistical software. Each loan would be associated with observable characteristics of the borrowers at the time the loan was extended, called predictors—for example college major, class rank, standardized test scores, geographic location, type of school attended, expected debt-to-income ratios at graduation—and the eventual outcome—that is, did the borrower default on the loan, and if so, what percent of the loan did the lender lose.

The result of OLS is a mathematical equation of the following form:

\[ E(DL) = b_0 + b_1 x_1 + \ldots + b_n x_n + e_n \]  

(Eq. 2)

¹⁶⁵. The Department of Education estimates its recovery rate on defaulted loans, after taking into account collection costs and the time value of money, as between 75% and 82%. Dep’t of Educ., Student Loans Overview, Fiscal Year 2013 Budget Request, at R-31 (2012). This high recovery rate may be due in part to the limits on bankruptcy discharge and extensive mechanisms available to collect defaulted federal student loans. In Equation 1, \( L \) therefore can be assumed to be between 0.18 and 0.25.

¹⁶⁶. Alternately, a logit or probit model could be used because the value of \( E(DL) \) should always be between 0 and 1.
in which:

\[ E(DL) \] is the predicted loss as a percent of the loan (after taking into account the time value of money)
\[ x \] is the value of each predictor (1 through \( n \))
\[ n \] is the number of predictors
\[ b \] is the value of each coefficient (1 through \( n \)) which describes the direction and magnitude of the relationship between the predictor and losses.
\[ e \] is the error term.

The value determined in equation 2 can be plugged into equation 1 as \((D \times L)\) to price the interest rate on the loan. This analysis is purely technical and data driven. The only assumption is that the future will look like the period of time during which the loans in the panel data were extended and tracked.

Empirically validated predictors, and ethical considerations regarding the use of certain predictors, are discussed in detail in Part V. Because some predictors of default are beyond students’ control, or relate to hallmarks of disadvantage, using them to price student loans would undermine the federal student lending programs’ policy goals of providing equal opportunity and upward mobility.

C. Risk-Based Pricing Is Not Necessarily the Same as Market-Based Pricing

Risk-based pricing as used in this Article has a precise definition that mathematically connects pricing to risk; it is not simply the price that a private lender would charge in an unregulated market. The goal of private lenders is to maximize profit, and this entails charging the highest price possible without losing too much volume or taking on too much risk. To the extent that consumer credit markets are less than perfectly efficient and price competitive, and demand for credit is inelastic, private lenders should have opportunities to charge prices that are higher than those required to compensate them for risk. This is what Alan M. White refers to as “opportunity pricing.”\(^{167}\)

\(^{167}\) Alan M. White, Risk-Based Mortgage Pricing: Present and Future
Consumer credit markets may be less than perfectly efficient because of complicated contractual terms or fee structures, information asymmetries, limited mathematical skill of borrowers, liquidity issues, or transaction or search costs. Empirical studies have demonstrated limited price competition in the mortgage and credit card markets, and it is probably safe to assume that pricing is also less than perfectly competitive in the private student loan market.

Journalists have documented efforts by some private student lenders to pay college financial aid officers to steer students toward specific lenders—that is, efforts to compete other than by providing credit to borrowers at the lowest price and on the best terms. Complaints collected by the Consumer Financial Protection Bureau suggest that some debtors who borrowed from private student lenders were not aware that they were eligible for lower cost federal loans.

---


168. See id.; Patricia A. McCoy, Rethinking Disclosure in a World of Risk-Based Pricing, 44 HARV. J. ON LEGIS. 123 (2007) (arguing that application fees in subprime mortgage markets can prevent comparison shopping and reduce price competition); Alan M. White & Cathy L. Mansfield, Literacy and Contract, 13 STAN. L. & POL'Y REV. 233, 266 (2002); see also infra note 193.

169. See, e.g., Andra C. Ghent & Marianna Kudlyak, Recourse and Residential Mortgage Default: Evidence from U.S. States, 24 REV. FIN. STUD. 3139, 3139–86 (2011) (finding that although mortgage default rates are lower in states with lender friendly collections laws, mortgage interest rates are not any lower in such states); Michael Simkovic, The Effect of BAPCPA on Credit Card Industry Profits and Prices, 83 AM. BANKR. L.J. 1, 7–22 (2009) (finding that although lender friendly bankruptcy reforms reduced losses to credit card lenders, credit spreads and fees on credit cards increased, likely because of industry consolidation and market inefficiencies).

170. See Glater, supra note 35, at 48–51 (describing the student loan scandal of 2007).

Because opacity and limited price competition create a favorable business environment for lenders, private lenders may not volunteer transparent, simple information that would make credit markets more efficient and price competitive. And without such clear, simple disclosures, risk-based pricing may not influence students’ behavior.

Even in relatively efficient credit markets such as corporate bond markets, credit spreads will typically exceed default risk because credit spreads also incorporate factors such as liquidity risk, systemic risk, and investor risk aversion. However, the U.S. government is generally not subject to such risks—the U.S. government’s borrowing costs actually go down during financial crises as investors flood into Treasuries and the Federal Reserve can provide liquidity in the event that investors are unwilling to do so. The U.S. government can therefore safely lend at lower interest rates than private lenders, and by lending directly to students, avoid leaky subsidies to private lenders such as guarantees and liquidity injections in times of crisis.

**D. Government’s Incentives Are Uniquely Well Aligned with Students’**

The discussion in Parts IV.A–B above considers only the role of government as student lender, and ignores the role of government as tax collector and provider of social insurance. Taking into account the broader role of government suggests that the government should be more risk-tolerant. Under some scenarios, the government might even benefit by running a risk-based student loan program at a loss—that is, as a subsidy program.

---

172. See Simkovic & Kaminetzky, supra note 153, at 194–99 (discussing the components of credit spreads in the corporate bond market).

173. See id. at 198 (noting that banks’ funding costs increase relative to the federal governments’ in times of crisis).

174. See Simkovic, *Competition and Crisis in Mortgage Securitization*, supra note 70, at 61–64 (arguing for direct mortgage lending by the federal government to avoid moral hazard and subsidies to private lenders in times of crisis).
If the government were only a lender, its interests would diverge from those of student borrowers, who resemble equity holders. However, as a tax collector, the government is in effect an equity investor in the student borrowers’ future income, and the government’s interests are therefore more closely aligned with those of student borrowers.

Pure lenders wish to minimize defaults and loan losses; they are sensitive to downside risk, but indifferent to upside potential, because any upside will go to the borrower, not the lender.¹⁷⁵ Student loan losses depend not only on average occupational income levels, but also on income distributions.¹⁷⁶ Lenders prefer narrower income distributions, because lenders face asymmetric payoffs—the best lenders can hope for is full repayment of each loan.¹⁷⁷ If an income distribution widens, a larger fraction of borrowers will have low incomes and will default on their loans, but the borrowers with higher incomes will not pay the lender any more than they are contractually required to pay.¹⁷⁸

In other words, there is a difference in interests between student lenders—who should prefer lower risk occupations—and student borrowers, who should be more willing to take greater educational and occupational risks in return for higher expected returns.

Unlike a private lender, the government can benefit from student borrowers’ upside potential because the government collects payroll and income taxes. The government—like student

¹⁷⁵. See Simkovic & Kaminetzky, supra note 153, at 214–18 (discussing the differing interests of equity and debt investors).


¹⁷⁷. Id.

¹⁷⁸. Id.
borrowers and unlike private lenders—should therefore be more willing to trade off greater risk (in the form of a wider distribution of incomes) in return for greater rewards (in the form of higher average income).

The extent to which the government should embrace risk will depend on a variety of factors, including labor tax rates, loan-to-income ratios, the likelihood of emigration, and social insurance levels.

Under some scenarios, the government can improve public finances and overall welfare by intentionally running a student loan program at no profit, or even at a loss when measuring direct revenue against direct costs. This is because a subsidy may cause more workers to pursue education that will increase their lifetime incomes and therefore their lifetime tax contributions. Higher tax revenues from the additional student borrowers who succeed may more than offset higher loan losses from the additional student borrowers who fail. A subsidy is most likely to be welfare-enhancing when individual students are risk-averse—for example, because they cannot easily diversify their investment in human capital—and the government is risk neutral because it is more diversified.

The government should ideally view itself not as a mere lender, but rather as a diversified investor in the global competitiveness of the United States labor force.

179. All else being equal, the government should be more risk-tolerant if labor tax rates are higher or more steeply progressive because tax revenue will be higher as average income increases and the distribution of incomes increases (more income will be taxed at higher rates).

180. All else being equal, the government should be more risk-tolerant if the size of the loan is smaller relative to the average lifetime income because the government’s potential loan losses will be smaller relative to the government’s potential tax revenues.

181. All else being equal, the government should be more risk-tolerant if the probability of skilled emigration is lower because expected future tax revenue will be higher.

182. All else being equal, the government should probably be more risk tolerant if social insurance levels are lower because downside risk for the student borrower will have less of an impact on public finances. However, to the extent that education can raise incomes or reduce unemployment, higher social insurance levels may weigh in favor of a larger subsidy for education.

183. The U.S. government should promote the interests of the U.S. work
Operationalizing such a view would require a complex behavioral and financial model that is beyond the scope of this Article. For our purposes, it should simply be noted that credit spreads calculated using the methods described in Part IV.B do not take into account positive externalities associated with higher education, and could therefore be interpreted as an upward bound on the interest rate that the government should charge student borrowers.

E. Forecasting Change Is Both Necessary and Perilous

A purely data-driven analysis of student loan performance and labor markets is appealing because it is objective and ameliorates concerns about political influence or corruption. Any labor market forecasts used to price student loans might face political pressure. Employers in many industries may wish to channel students toward their field to drive down labor costs and upgrade their workforces. Conversely, skilled workers might wish to channel students away from their own fields to reduce competition and keep wages high. Employers may claim labor force because the U.S. government cannot tax the work forces of other countries. The U.S. government is also elected to represent the interests of U.S. citizens, most of whom depend on wages as their primary source of income.


186. The “scamblog” literature, in which ostensibly disgruntled lawyers and doctors advise others not to follow in their career footsteps, may reflect efforts by the shrewdest and most highly paid skilled workers to reduce competition and increase wages in their own fields. Department of Labor and Census data
“shortages” and workers may claim “surpluses,” but these claims are amorphous in a market system in which supply-and-demand imbalances can be corrected by changes in wages.\textsuperscript{187} Risk-based lending policies should be based on hard numbers such as real wages and unemployment rates in different occupations, not on empirically unverifiable claims by interested parties. As suggested by the data in Figure 6 below, long-term historic wages by major are often a reasonably good predictor of future wages by major.

\textsuperscript{187} As discussed above, labor and education markets are not perfectly efficient competitive markets that flawlessly and quickly adjust to reach equilibrium. \textit{See supra} note 188. However, there is evidence that students respond to changes in starting salaries and that labor markets can thereby adjust, however imperfectly. \textit{See, e.g.}, Richard B. Freeman, \textit{The Overeducated American} 62–63, 98–108 (1976) (explaining that the surplus of doctorate students in the 1970s led to a market drop in job prospects, which led to a decline in the number of students seeking a Ph.D. in the 1980s).
Figure 6: Past Starting Salary for Graduates with a Certain Major Is a Reasonably Good Predictor of Future Starting Salaries by Major

Unfortunately, the assumption that the past accurately predicts the future in the student loan and labor markets—the continuity assumption—might not always be the wisest, especially during periods of rapid technological change or crisis. The introduction of risk-based pricing could itself alter labor markets by channeling students toward certain occupations and away from others.\textsuperscript{188} Forecasting change based on leading indicators rather than historic data could be critical to avoid an

\textsuperscript{188} Id. at 62–63.
overshoot. For example, there is some evidence that labor market policies to encourage more students to study science in the 1960s may have contributed to high unemployment rates and low wages for scientists such as physicists in the 1970s.

The continuity assumption could be relaxed by incorporating additional information about the labor market to forecast changes, for example, by using a cobweb model including data about future labor supply (i.e., the total number of students in a city or state, across the United States or around the world studying toward each occupation and the number of years until they complete their training) and future demand (the number of job openings for each profession according to employer surveys or industry growth estimates combined with input-output tables). The Department of Education could be offered limited discretion—that is, it could be permitted to use a forecasting model to set an interest rate within a range of the rate suggested by the historic data. Or perhaps the Department of Education should establish the historic baseline and the discretionary forecast should be under the aegis of the U.S. Department of Labor’s Bureau of Labor Statistics.

Ideally, risk-based pricing could be used to help labor markets adjust more quickly and with less volatility than seen in traditional cobweb cycles. Reliable, detailed, and timely data is essential. Forecasting would benefit from greater integration of dispersed data sources and timelier reporting.

Existing data sources include government data—U.S. Bureau of Labor Statistics’ (BLS) Occupational Employment Statistics (OES), the BLS National Compensation Survey (NCS), the BLS Job Openings and Labor Turnover Survey (JOLTS), the Bureau of the Census Current Population Survey (CPS), Internal Revenue Service (IRS) data, Social Security Administration (SSA) data, and a variety of surveys compiled by the National Center

189. Id.
190. Id. at 62–73, 98–108.
191. For a discussion of limitations of existing employment forecasting models, see Richard B. Freeman, supra note 185, at 13 (arguing that global rather than national or local labor markets, market and technological changes, factor substitution, and other issues dramatically limit the accuracy of employment forecast models).
for Education Statistics (NCES); foundation data such as the College and Beyond Database (C&B) and Panel Study of Income Dynamics (PSID); and proprietary data from payroll processors or the National Association of Colleges and Employers (NACE). 192

F. Private Lenders Could Be Enlisted to Correct Forecast Errors

The government could provide an additional safety-valve to correct for erroneous forecasts by providing private employers or other investors an opportunity to fund student loans at lower interest rates than the government if the investors believe that government forecasts are too conservative for particular fields of study. 193 Such an approach would be less susceptible to


193. For price competition to be most effective, it would probably be necessary to create a standard student loan contract that could vary only with respect to one or a few numeric price terms (i.e., the interest rate) and feature clear and easy-to-understand disclosures. Complex pricing schemes and varied contractual terms can reduce price competition in consumer credit markets by exploiting bounded rationality or cognitive biases, confusing borrowers, and
RISK-BASED STUDENT LOANS

manipulation than forecasts based on surveys of employers, because private lenders would face loan losses if they used overly optimistic forecasts.194

Indeed, two small lenders are already engaged in “cherry picking,” funding loans to low risk students at elite MBA programs at slightly below the federal government’s rates.195

G. Limited Dischargeability of Student Loans in Bankruptcy Reduces Lender Incentives to Monitor Students and Risk-Adjust Student Loan Pricing


194. The incentive to make accurate forecasts could be enhanced by making student loans more readily dischargeable.


196. See Robert Clark, CORPORATE LAW 6–10 (1986) (discussing rationales
limited in U.S. education finance. Instead, risks are imposed primarily on student borrowers.\textsuperscript{197} Risks are shifted from lenders to student borrowers by restricting student-loan debtors’ access to a bankruptcy discharge and curbing exemptions from collections for income sources such as social security retirement and disability benefits.\textsuperscript{198}

This risk-shifting toward student borrowers implicitly reflects an assumption that individual student borrowers are in a better position to assess and bear the risks of education than is the government or a private lender.\textsuperscript{199} However, as discussed above in Parts III and IV, there are strong theoretical and practical reasons to believe that the government as creditor may often be in a better position to evaluate the risk of education and spread that risk, and risk should therefore be shared more equally.\textsuperscript{200}

\begin{itemize}
\item \textsuperscript{197.} \textit{Understanding Default}, supra note 152; \textit{see also infra} notes 210–24 and accompanying text (discussing limits on dischargeability in bankruptcy); \textit{cf.} Section V.H. (discussing a recent shift toward greater debt forgiveness for federal student loan borrowers through income based repayment plans).
\item \textsuperscript{198.} \textit{Facing Loan Default}, supra note 197.
\item \textsuperscript{199.} \textit{See Jackson, supra} note 196, at 229 (“Recent scholarly treatments of discharge law have focused on whether the debtor or the creditor is the superior risk bearer and whether discharge should be presumptively available.”); Theodore Eisenberg, \textit{Bankruptcy Law in Perspective}, 28 UCLA L. Rev. 953, 982 (arguing that debtors are more able to control their financial activities and judge their financial circumstances than lenders, thus debtors are presumably more able to bear risks); \textit{id.} at 981 (“A discharge system provides a technique for allocating the risk between a debtor and his creditors.”); \textit{id.} at 982 (arguing that a debtor is presumably better able to bear risks than creditors because a debtor is “in greater control of [his] financial activities than any particular lender” and therefore a better judge of his own circumstances).
\item \textsuperscript{200.} \textit{See, e.g.}, \textsc{Barr-Gill & Warren, supra} note 193, at 69–74 (discussing the need for safety regulation in various consumer credit markets because of limitations on information and rationality); Katherine Porter, \textit{College Lessons: The Financial Risks of Dropping Out}, in \textsc{Broke: How Debt Bankrupts The}
Thomas Jackson explains the policy goals behind the bankruptcy discharge as one of nuanced paternalistic regulation of credit.\textsuperscript{201} The goal is to protect debtors by enlisting sophisticated creditors:

In each case [of risky activities] society decides—or should decide—at what point the expected costs of a given activity outweigh the prospective benefits to the individual and society. Borrowing, however, cannot be regulated by means of the rough and general rules . . . . Discharge policy provides an alternative . . . . Discharge . . . heightens creditors' incentives to monitor; by providing for a right of discharge, society enlists creditors in the effort to oversee the individual's credit decisions . . . . The availability of the right of discharge induces creditors to restrict the individual's credit intake and thus to assist in ensuring that he does not seriously underestimate his future needs.\textsuperscript{202}

In theory, restricting access to discharge for student borrowers should reduce (though not eliminate) lenders' risk of loss\textsuperscript{203} and thereby reduce lenders incentives to monitor borrowers and to ration and price credit according to risk.\textsuperscript{204}

\textsuperscript{201} Jackson, \textit{supra} note 196, at 248–52.
\textsuperscript{202} \textit{Id.} at 248–49.
\textsuperscript{203} See Ronald J. Mann, \textit{Bankruptcy Reform and the “Sweat Box” of Credit Card Debt}, 2007 U. Ill. L. Rev. 375, 379–81 (arguing that delaying bankruptcy discharge benefits creditors by extending the time period during which they can collect); Simkovic, \textit{supra} note 169, at 1 (finding that bankruptcy reforms that restricted consumer borrowers' access to a Chapter 7 discharge reduced credit card lenders' losses).
\textsuperscript{204} See also Stephen J. Lubben, \textit{Derivatives and Bankruptcy: The Flawed
These predictions hold true with respect to the student loan market. Government and private lenders could protect themselves from losses due to default by engaging in good underwriting—205—that is, by assessing the riskiness of different student borrowers and different courses of education, adjusting the pricing and availability of credit accordingly, and subsequently accepting and spreading losses. However, rather than undertake the socially useful task of evaluating students’ career and income prospects based on their academic performance and chosen field of study—and sharing this information with students through disclosures of differential loan pricing—lenders price loans uniformly and transfer as much of the risk as possible to student borrowers.206

There are few credit-market signals to warn students of danger when they decide where to enroll and what to study. The government does not restrict loans to students who wish to attend relatively high-risk educational institutions.207 Nor does the government differentially price student loans according to borrower-specific or program-specific risks.208 Student borrowers, however, are sorely in need of good information. The

---

205. See Simkovic, Competition and Crisis in Mortgage Securitization, supra note 70 (explaining the underwriting process for mortgages).

206. Although many private student lenders purport to engage in “risk-based” pricing, they often do so based on credit scores, which may reflect the credit history of the student’s parents rather than the future earning potential and creditworthiness of the student. See Credit Scores, FINAID.ORG, http://www.finaid.org/loans/creditscores.html (last visited Feb. 3, 2013) (on file with the Washington and Lee Law Review); see also supra Part IV.C (discussing “opportunity pricing”). There are, however signs that at least some private lenders are beginning to price differences in earnings risks across programs. See supra notes 160, 173.

207. See supra Part II.B.

208. See supra Part II.D.
consequences to student borrowers of making the wrong choice can be severe and lifelong.\textsuperscript{209}

Whereas the U.S. Bankruptcy Code generally gives individual borrowers an “insurance policy” against failure in the form of a bankruptcy discharge,\textsuperscript{210} student loans are somewhat more difficult to discharge than most kinds of debt.\textsuperscript{211} Student loans have become more difficult to discharge through a series of amendments enacted over the last forty years, ostensibly to protect taxpayers by shifting risks onto students.

Federally-backed student loans have been available for many decades, first under the NDEA in the late 1950s and early 1960s, and subsequently under the HEA starting in 1965.\textsuperscript{212} Until 1976,

\begin{itemize}
\item \textsuperscript{209} See Atkinson, supra note 9, at 5 n.13 (explaining that student loan debt can have negative consequences); Porter, supra note 200, at 85–100 (arguing that educational debt is risky because, although completing a four-year college degree reduces the risk of bankruptcy by increasing income, attending college without completing a four-year degree exacerbates financial distress by increasing debt without significantly increasing income); \textit{id.} at 96–97
\item The burdens of student loan debt have consequences for people’s financial futures [including] a lower rate of saving . . . retarding opportunity for other family members . . . defer[ing] entry into the job market, losing seniority and reducing the number of working years they have to save for retirement.
\item See also Warren, Baum & Sitaraman, supra note 65, at 130 (discussing the impact of educational debt on life decisions).
\item See Barry Adler et al., \textit{Regulating Consumer Bankruptcy: A Theoretical Inquiry}, 29 J. LEGAL STUD. 585, 591 (2000) (analogizing bankruptcy discharge to a form of social insurance, both in terms of its ability to mitigate poverty and in terms of ex-ante moral hazard concerns); \textit{Jackson}, supra note 196, at 230–32 (same).
\item In spite of formal language in the Bankruptcy Code limiting discharge, some student borrowers have received relief from their loans in bankruptcy. See Rafael I. Pardo & Michelle R. Lacey, \textit{The Real Student-Loan Scandal: Undue Hardship Discharge Litigation}, 83 AM. BANKR. L.J. 179, 188 (2009) [hereinafter \textit{Pardo & Lacey, Real Student-Loan Scandal}].
\item According to one scholar, relief is often offered to those who seek it, but too few bankruptcy lawyers attempt to discharge student loans in bankruptcy. See Jason Iuliano, \textit{An Empirical Assessment of Student Loan Discharges and the Bankruptcy Undue Hardship Standard}, 86 AM. BANKR. L.J. 495 (2012).
these loans were dischargeable in bankruptcy, the same as credit card debt or any other unsecured loan.213

In the early 1970s, salacious stories began to circulate in Congress about young graduates discharging their debts en masse in bankruptcy shortly after completing their educations.214 According to the rumors, young professionals with few hard assets to lose and a lifetime of high future incomes around the corner were opportunistically trying to discharge their student loan obligations.215

The rumors of strategic default were largely unfounded. According to research by Professor Rafael Pardo, strategic default by high-income student debtors was extremely rare.216 Professor Pardo supports this assertion with both a contemporary study by the U.S. General Accounting Office (GAO)217 and comments by members of the House of Representatives.218 Similarly, empirical studies of student debtors in Canada have found that bankruptcy

---

213. See id.


215. See Jackson, supra note 196, at 250–51

As a general rule, college and graduate students have few current assets but large future income streams. Using bankruptcy is relatively painless to them, as they have few assets to lose, and obtaining a discharge offers a substantial benefit, as it frees up the future income stream from the substantial obligation of repaying a student loan.

216. Pardo & Lacey, Undue Hardship, supra note 212, at 420

Despite evidence presented to the Commission [on the Bankruptcy Laws of the United States] that less than one percent of federally insured loans were discharged in bankruptcy, its recommendation essentially sought to preempt “potential abuses,” defaults that industry representatives of the student loan system anticipated would occur. The Commission thus reacted viscerally to anecdotal evidence.


218. Id. at 422–23 (quoting Representative James O’Hara’s critique of nondischargeability of student loans as treating student debtors like criminals or frauds).
abuse by high-income professionals is rare, and a recent study has found extremely low student loan default rates among former U.S. law students.\textsuperscript{219}

Nevertheless, tales of student opportunism may have persuaded Congress to tighten the rules for student loans. In 1976, a time delay was imposed so that debtors could not discharge student loans that were incurred within five years before the bankruptcy filing, unless denial would constitute an “undue hardship” for the debtor or his or her family.\textsuperscript{220} Over the next several decades, Congress further restricted dischargability of student loans\textsuperscript{221} and some bankruptcy courts interpreted this as a cue to raise the bar in their interpretation of “undue hardship.”\textsuperscript{222} In 1990, the time-bar was extended from five years to seven, and in 1998 it became indefinite (i.e., a debtor would be required to demonstrate “undue hardship,” no matter how long ago the student loan debts were incurred).\textsuperscript{223} In 2005, restrictions on dischargability were extended to fully private loans, not just guaranteed or direct loans.\textsuperscript{224}

Although various judicial tests have been advanced to clarify the meaning of “undue hardship,” Professor Pardo’s research suggests that hardship remains in the eye of the beholder: in practice, discharge depends more on the particular bankruptcy judge than on the objectively measurable financial condition of the student debtor.\textsuperscript{225}

\begin{flushleft}
\textsuperscript{219} Stephanie Ben-Ishai, *Government Student Loans, Government Debts and Bankruptcy: A Comparative Study*, 44 CAN. BUS. L.J. 211, 237–38 (2006-2007) (reviewing the Canadian literature); Simkovic & McIntyre, supra note 186 (showing very low student loan default rates among former law students who entered repayment from 1990 to 2010).


\textsuperscript{221} Id. at 427; John A.E. Pottow, *The Nondischargeability of Student Loans in Personal Bankruptcy Proceedings: The Search for a Theory*, 44 CAN. BUS. L.J. 245, 249–50 (2006); Braucher, supra note 71, at 473.

\textsuperscript{222} Pardo & Lacey, *Undue Hardship*, supra note 212, at 428.

\textsuperscript{223} Braucher, supra note 71, at 473–74.

\textsuperscript{224} Id. at 473–74.

\textsuperscript{225} Pardo & Lacey, *Real Student-Loan Scandal*, supra note 211, at 185;
\end{flushleft}
Many scholars have questioned whether these special restrictions on discharge of student loans were needed to prevent abuse of the bankruptcy system by strategic defaulters. Most scholars conclude that special protections for student loans are unnecessary, empirically or theoretically unjustifiable, or that other factors weigh heavily in favor of a shift toward greater dischargeability.

For example, Thomas Jackson, Rafael Pardo, Katherine Porter, John Pottow, and Abbye Atkinson have all pointed out that student loan abuse would be curtailed through other provisions of the Bankruptcy Code. In 1984, the Bankruptcy Code was amended so that a bankruptcy court could dismiss a case without granting a discharge if the petitioner's debts were “primarily consumer debts” and if granting a discharge “would be a substantial abuse” of the bankruptcy system. These anti-abuse provisions were strengthened in 2005. Perhaps Congress was concerned that some student loans would not meet the threshold test as “consumer debts” because they were incurred with a “profit motive”—that is, the student borrower expected the degree to boost future income by more than direct educational costs and opportunity costs.

---

Pardo & Lacey, Undue Hardship, supra note 212, at 411, 478–509; cf. Iuliano, supra note 211.

226. See JACKSON, supra note 196, at 251.


228. See Porter, supra note 200, at 98.

229. See Pottow, supra note 221, at 251–55.

230. See Atkinson, supra note 9, at 34–37.

231. 11 U.S.C. § 707(b) (2006); see also id. § 1325(a)(3) (requiring that a Chapter 13 plan be “proposed in good faith”).

232. The Bankruptcy Abuse and Consumer Protection Act of 2005 (BAPCPA) changed the standard for dismissal of a Chapter 7 petition from “substantial abuse” to “abuse” and eliminated a presumption in favor of granting the debtor relief. Id. § 707(b)(1); 6-707 COLLIER ON BANKRUPTCY 707.4 (16th ed. 2011).

233. See In re Dickerson, 193 B.R. 67, 70 (Bankr. M.D. Fla. 1996) (finding that student loans were not consumer debt); In re Gentri, 185 B.R. 368, 373 (Bankr. M.D. Fla. 1995) (finding that medical school loans were not consumer debts); In re Hill, No. 94-01881, 1994 WL 738663, at *1 (Bankr. D. Idaho Dec. 22, 1994) (finding that student loan debt was not consumer debt); In re Groves, 160 B.R. 121, 123 (Bankr. E.D. Mo. 1993) (finding that student loans were not
However, if the concern is that student loans might be profitable investments, then why should a legitimate investment in human capital subject the student investor to unique risks that do not apply to investors in any other form of capital? Conversely, why should a “substantial abuse” standard not apply equally to other unsecured business debts? The Bankruptcy Code already limits discharge of debts obtained by fraud or false pretenses, regardless of whether those debts are business or consumer debts. So why is greater protection required for student loans?

John Pottow explores several possible theoretical justifications for the special limitations on the discharge of student loans, and finds all of them wanting. Professor Pottow considers the possibility that student debtors are particularly dishonest and student loans presumptively fraudulent and the similar possibility that the inalienability of an education and higher future wages incentivizes opportunistic behavior. He finds both theories consistent with the treatment of student debtors under U.S. law, but empirically doubtful in light of research by Professor Pardo and the GAO. Professor Pottow

---

234. See Pottow, supra note 221, at 254 (“In fact, in business, far from being disparaged as fomenting ‘opportunism,’ the bankruptcy discharge is styled as fostering ‘entrepreneurialism.’”).


236. See Pottow, supra note 221, at 276 (“[T]he most attractive [theories for restricting student loan discharge] seem to be the ones least reflected in many of the current bankruptcy laws, just as the ones most recognizable in today’s statutes seem grounded in confusion and myth.”).

237. Id. at 251–55.

238. Id. (“The theory that comes closest to persuasion as to why student loans should have restricted dischargeability in bankruptcy is that of the opportunistic debtor, ‘softly’ defrauding the system if she walks away from publicly subsidized debt that enables a high-income career.”).

239. Id. at 255 (“[T]here seems to be a documented lack of empirical evidence supporting routine abuse by rich-career students using bankruptcy just out of...
suggests that much of the opposition to dischargability of student loans stems from a kind of class envy—a vision of student debtors as individuals who became rich at the expense of the public. 240 Pottow therefore argues for income-contingent repayment to reduce the burdens on debtors who have lower incomes. 241

Apparently in agreement with Professor Pottow, several scholars argue for greater student loan forgiveness toward those for whom education does not produce a private financial benefit. Abyye Atkinson presents evidence that African Americans receive less financial benefit from education than whites, and argues that the nondischargeability of student loans therefore disproportionately harms college-educated African Americans. 242

Kathryn Porter presents evidence that those who receive some college education but do not complete a four-year degree—either because they pursue a two-year associate’s degree or vocational program, or because they drop out prior to completing a four-year degree—will often be in a worse financial position than those who never attended college at all. 243 Porter argues for greater forgiveness for those who fail to complete their educations because of family or financial misfortune. 244

240. Id.
241. Id. at 276–78.
242. See Atkinson, supra note 9, at 2–5; id. at 5–6

Congress has largely placed the burden and risk of paying for college firmly on the shoulders of the student . . . . [T]hese educational loan policies may reveal a judgment, however inadvertent, about who, as a practical matter, should and who should not be going to college. More troubling is that this judgment seems to track racial divisions.

243. See Porter, supra note 200, at 85–100 (arguing that educational debt is risky because, although completing a four-year college degree reduces the risk of bankruptcy by increasing income, attending college without completing a four-year degree exacerbates financial distress by increasing debt without significantly increasing income); id. at 96–97

The burdens of student loan debt have consequences for people’s financial futures [including] a lower rate of saving, . . . retarding opportunity for other family members, . . . defer[ring] entry into the job market, losing seniority and reducing the number of working years they have to save for retirement.

244. Id. at 98–100.
Jean Braucher questions the nondischargeability of student loans for those who attend low-quality institutions that do little to enhance their students’ career prospects, particularly given the governments’ limited efforts to police educational quality and protect students from deceptive sales and marketing practices.245

Professor and now Senator Elizabeth Warren argues that college-educated individuals who forgo a higher paying job in the private sector to pursue public service after graduation should receive more generous loan forgiveness.246 Professor Warren bases her argument in part on an assumption that individuals who work in public service are undercompensated relative to the value they contribute to society.247

H. Income-Based Repayment Plans

The scholarly arguments for greater student loan forgiveness have been partially successful. Although the Bankruptcy Code continues to restrict student loan discharge, debt forgiveness has been advanced through recent changes to federal student loan programs known as income based repayment plans (IBR), introduced in 2007 through the College Cost Reduction and Access Act.248 Under IBR, federal student loan payments are capped at a percentage of the student debtor’s income, typically around 10%.249 The payment decreases as the number of people in the debtor’s household increases.250 After the student debtor

246. Warren, Baum & Sitaraman, supra note 65, at 131–36, 142.
247. Id. at 142 (“By tying debt forgiveness to public service, Americans would have the chance to say that everyone who does this kind of work deserves a substantial reward from the rest of us. No longer would public service opportunities be limited to a few poorly funded programs.”).
248. For an overview of the new Income Based Repayment Plan, see Philip G. Schrag & Charles W. Pruett, Coordinating Loan Repayment Assistance Programs with New Federal Legislation, 60 J. LEGAL EDUC. 583, 590–97 (2010).
249. The annual payment is 15% multiplied by the amount by which the debtor’s adjusted gross income exceeds 150% of the poverty level for a household the size of the debtor’s family. This will usually work out to around 10% of the debtor’s adjusted gross income. Id. at 590–91.
250. Id. at 594 (“A larger family entitles the borrower to a larger deduction, and therefore permits a smaller monthly payment.”).
makes all payments for a number of years, any remaining debt is forgiven.

Consistent with Professor Warren’s views, IBR is more generous to student debtors who work in the public sector. Public sector workers need only work and make payments for ten years prior to forgiveness of the balance of their federal student loans, whereas private sector workers must work and make payments for twenty to twenty-five years. If the debtor’s income rises so that payments would be lower under a traditional (non-IBR) ten-year repayment plan, debtors may instead make the lower traditional ten-year payment.

IBR is less attractive than discharge in bankruptcy. IBR debt forgiveness may result in taxable income to student debtors who work in the private sector, whereas a bankruptcy discharge is not treated as income. Whereas a Chapter 7 bankruptcy discharge would provide immediate relief, and a Chapter 13 bankruptcy discharge would provide relief after a three-to-five-year period of income-based repayment, IBR requires ten to twenty years of payments.

251. For Warren’s views generally on student loans and public sector employment, see supra notes 246–50 and accompanying text.

252. Schrag & Pruett, supra note 248, at 591–92; see also Alison Damast, Obama’s New ‘Pay as You Earn’ Plan a Windfall for MBAs, BLOOMBERG BUSINESSWEEK, Nov. 2, 2012 (describing a new federal plan that provides student loan forgiveness after 20 years); Philip G. Schrag, Failing Law Schools: Brian Tamanaha’s Misguided Missile, 26 GEO. J. LEGAL ETHICS (forthcoming 2013) (same).


254. Id. at 593; see also 26 U.S.C. § 108(f) (2006) (describing the income from discharge of indebtedness with respect to student loans). Debt forgiveness is treated as taxable income to the extent the student is balance-sheet solvent or rendered balance-sheet solvent by the forgiveness, that is, the debtors’ assets exceed the debtors’ liabilities at the time of the forgiveness. See id. § 108(d)(3) (defining “insolvent”); id. § 108(a)(1)(B) (providing for exclusion of discharge income if “the discharge occurs when taxpayer is insolvent”); id. § 108(a)(3) (defining the limitation on the insolvency exclusion).

255. Id. § 108(a)(1)(A).

256. Post-petition wages are not property of the estate under 11 U.S.C. § 541(a)(6) for a Chapter 7 bankruptcy. See, e.g., In re Hellums, 772 F.2d 379, 380 (2nd Cir. 1985) (per curiam). A Chapter 7 discharge will therefore leave the debtor’s future wage income unencumbered.

257. A five-year repayment plan is required in Chapter 13 for above-median income debtors. The payments are based on the debtor’s income, less minimal
twenty-five years of repayment. The lengthy repayment period under IBR may mitigate moral hazard. IBR is formulaic, and may therefore be less expensive to administer and more consistently applied than the “undue hardship” standard in bankruptcy.

IBR reduces the downside risk of education, and may provide welcome relief to student debtors who are in a difficult financial position. However, IBR only provides relief ex-post. It does not generate information or establish incentives that lead to a more efficient allocation of educational resources ex-ante.

Risk-based pricing is compatible with IBR and could incorporate risk of loss due to borrowers entering IBR. To the extent that IBR is intended as a back-door wage subsidy for public service workers, losses from IBR for students entering public service could be excluded from risk-based pricing.


259. See Pottow, supra note 221, at 268 (describing the application of “the U.S. ‘undue hardship’ test” as “an unpredictable and expensive way” to “back-end income-contingency into the American system”).
VI. Ethical Considerations and the Limits of Risk-Based Pricing

A. Factors That Predict Student Loan Defaults

The most important predictor of default is probably the student borrower’s employment prospects and whether post-graduation income is adequate to service educational debts. Studies have also found that students are less likely to default if they are employed in a field that is related to their major. Some studies report that defaults are lower for STEM majors. Students who drop out are much more likely to default, and attrition can be predicted from poor academic performance both before college and during college.

260. See Thomas A. Flint, Predicting Student Loan Defaults, 68 J. HIGHER EDUC. 322, 344 (1997) (noting that in contrast to other factors, “the borrowers’ own disposable incomes do significantly influence default”); Jacob P. K. Gross et al., What Matters in Student Loan Default: A Review of the Research Literature, 39 J. STUDENT FIN. AID, no. 1, 2009, at 23 (“Most students who default do so because their personal income is inadequate to keep up with their payments.”); Laura G. Knapp & Terry G. Seaks, An Analysis of the Probability of Default on Federally Guaranteed Student Loans, 74 REV. ECON. & STAT. 404, 410 (1992) (noting that graduation correlates strongly with lower default, perhaps because it increases job and wage prospects); Kirk Montverde, Managing Student Loan Default Risk: Evidence from a Privately Guaranteed Portfolio, 41 RES. HIGHER EDUC. 331, 336, 350–52 (2000) (emphasizing the importance of the individual student’s ability to attain employment); J. Fredericks Volkwein et al., Factors Associated with Student Loan Default Among Different Racial and Ethnic Groups, 69 J. HIGHER EDUC. 206, 223, 228 (1998) (“[E]ven though student borrowers with advanced degrees emerge from college with higher levels of debt, their investment generally enables them to enter careers that . . . make loan repayment more likely.”).

261. Flint, supra note 260, at 346.

262. See id. at 330; Volkwein et al., supra note 260, at 222; see also J. Fredericks Volkwein & Bruce Szelest, Individual and Campus Characteristics Associated with Student Loan Default, 36 RES. HIGHER EDUC. 41, 41–72 (1995) (suggesting that while type of major is “not itself significantly influential in default,” a science or technology major may have “an indirect influence”).

263. See Flint, supra note 260, at 330 (higher grade point average and graduation both decrease the probability of nonpayment of loans); Knapp & Seaks, supra note 260, at 408 (noting that “[t]he single variable with the greatest statistical and economic significance is the occurrence of the student’s graduation”); Montverde, supra note 260, at 336 (noting that studies had found “degree completion” and “college GPA” as “significantly predictive of default risk”); Volkwein et al., supra note 260, at 222 (“[D]egree completion has a dramatic influence on lowering the rate of loan default.”).
Other individual characteristics that have been shown to predict default include race, age, parental education, and college major.

264. Student borrowers who are members of racial minority groups are more likely to default than white students, and African Americans are the most likely to default, and this relation holds true even after controlling for post-graduation income. Gross et al., supra note 260, at 21–22; see also Matt Steiner & Natali Teszler, Tex. Guaranteed & Tex. A&M Univ., The Characteristics Associated with Student Loan Default at Texas A&M University 48 (2003), http://www.tgslc.org/pdf/tamu_default_study.pdf (finding that white borrowers at Texas A&M in College Station “had” default rates below the average, while minorities had above-average rates, with “black borrowers hav[ing] the highest default rate”); J. Frederick Volkwein & Alberto Cabrera, Who Defaults on Student Loans? The Effects of Race, Class, and Gender on Borrower Behavior, in CONDEMNING STUDENTS TO DEBT: COLLEGE LOANS AND PUBLIC POLICY 105, 109 (Richard Fossey & Mark Bateman eds. 1998) [hereinafter Volkwein & Cabrera] (“African and Native American borrowers from all institution types have high default rates.”); Dana E. Christman, Multiple Realities: Characteristics of Loan Defaulters at a Two-year Public Institution, 27 COMMUNITY COLL. REV. 16, 23–25 (2000) (noting a study in which “default rates were found to be higher for minority students” and data to support the notion that “Non-Whites” are more likely to default); Laura L. Greene, An Economic Analysis of Student Loan Default, 11 EDUC. EVALUATION & POL’Y ANALYSIS 61, 61–68 (1989) (noting that African American status correlates with higher default and higher default amount); Elizabeth Herr & Larry Burt, Predicting Student Loan Default For The University Of Texas at Austin, 35 J. STUDENT FIN. AID, no. 2, 2005, at 37 (2005) (noting the results of the study “impl[y] that minority students, particularly Blacks and Hispanics, are at a greater risk of default”); Knapp & Seaks, supra note 260, at 408 (noting that African Americans are more likely to default); Volkwein & Szelest, supra note 262, at 51–52 (showing data that indicates higher default rates for African Americans and Native Americans); Wellford W. Wilms, Richard W. Moore & Roger E. Bolus, Whose Fault Is Default? A Study of The Impact of Student Characteristics and Institutional Practices on Guaranteed Student Loan Default Rates in California, 9 EDUC. EVALUATION & POL’Y ANALYSIS 41, 46 (1987) (“Blacks have the highest propensity to default . . . .”). This may be due to higher dropout rates, higher divorce rates, lower wealth levels, and larger numbers of dependents. Volkwein, et al., supra note 260, at 224–25. A risk-based approach that focused on expected income as the sole proxy for ability to pay would probably under predict defaults by minority borrowers and help maintain educational access for students from diverse backgrounds.

265. Older students are more likely to default. See, e.g., Jennie H. Woo, CLEARING ACCOUNTS: THE CAUSES OF STUDENT LOAN DEFAULT 6 (2002), http://cdm16254.contentdm.oclc.org/cdm/singleitem/collection/p178601cc2/id/29/08/rec/3 (“Older students default more often than younger ones.”); Christman, supra note 264, at 23 (noting that “being over 25 years old [was a] characteristic associated with high default rates”); Flint, supra note 260, at 347 (noting that it is uncertain why “older borrowers constitute a greater risk for default”). But see Steiner & Teszler, supra note 264, at 48 (finding that for
levels, family income levels, family structure, debt burdens, and (for law students) credit scores.

Characteristics of educational institutions may not provide much additional predictive accuracy. Although two-year community colleges have higher default rates than traditional nonprofit four-year institutions—and wealthier institutions tend

students at Texas A&M in College Station that “[b]orrowers between the ages of 23 and 26 have the lowest default rate (3.2 percent), with both younger borrowers and older borrowers representing increased levels of default risk”); Herr & Burt, supra note 264, at 39 (noting that while students “over 40 have higher loan default rates than borrowers in their late twenties and thirties,” so do younger borrowers, i.e., “between the ages of 20–24”).

266. Students whose parents are less educated are more likely to default. See Sandra Barone, Matt Steiner & Natali Teszler, Tex. Guaranteed Student Loan Corp., Multivariate Analysis of Student Loan Defaulters at Texas A&M University-Kingsville 23–24 (2005), available at http://www.tgslc.org/pdf/tamu_k_multivariate_analysis.pdf (indicating that parental education level, if sufficiently high, “may reflect a borrower’s previous exposure to responsibilities such as repaying a student loan”); Steiner & Teszler, supra note 264, at 49, 51 (finding that for students of Texas A&M in College Station, parental education attainment generally relates inversely to rate of default); Volkwein, et al., supra note 260, at 215 (noting that one of the features “associated with low levels of loan default include . . . a college-educated parent”); Volkwein & Szelest, supra note 262, at 51–52 (showing tables indicating higher default rates for those whose parents have lower education levels); Herr & Burt, supra note 264, at 35 (showing higher rates of default for lower mother and father educational attainment levels).

267. Students whose parents have lower incomes are more likely to default. See Steiner & Teszler, supra note 264, at 57 (finding that for students at Texas A&M in College Station, “[i]n general, default rates decrease as income increases”); Woo, supra note 265, at 5 (“[A]mple family resources, either higher incomes or assets, significantly lowered the probability of default.”); Knapp & Seaks, supra note 260, at 406 (indicating that parental income level is a significant factor and correlates negatively with default); Volkwein et al., supra note 260, at 221 (highlighting “parent income below $17,000” as one of three factors resulting in “significant increases in the probability of loan default”); Volkwein & Szelest, supra note 264, at 51–52 (showing higher rates of default for lower parental income levels); Wilms et al., supra note 264, at 42 (indicating that previous studies had shown that “[l]ow family income [is] also associated with high defaults”); Herr & Burt, supra note 264, at 37 (noting that, in line with the results of other studies, “students whose parents have higher incomes are less likely to default”).

268. Students who are divorced or separated or who have dependents are more likely to default. See Montverde, supra note 260, at 336; Volkwein & Szelest, supra note 262, at 57.

269. Montverde, supra note 260, at 340–44.
to have lower default rates—studies suggest that higher default institutions’ loan performance may be due to these institutions disproportionately serving students whose individual characteristics make them more likely to default—for example, students who are from lower income or less wealthy families.270

B. Preserving Equal Opportunity, Social Mobility, and Individual Choice

Although risk-based pricing involves technical analysis of data, it also implicates important ethical considerations. As discussed above, risk-based pricing reduces the transfer of value from low-risk borrowers to high-risk borrowers by forcing all borrowers to internalize their own risks.

In some situations, differences in relative risk levels may be driven by choices and behaviors that can be changed, and that we might affirmatively wish to encourage borrowers to change. In such situations uniform pricing creates moral hazard—that is,

---

270. Gross et al., supra note 260, at 21; see also Flint, supra note 260, at 348 (noting that “[l]arge numbers of low-income and minority students enroll in proprietary schools,” which also tend to have higher rates of default); Knapp & Seaks, supra note 260, at 406–07, 410 (finding the variables of two-year versus four-year institution not statistically significant and concluding that “individual characteristics, and not institutional characteristics, [that] are key determinates of default”); Montverde, supra note 260, at 351–52 (indicating that personal characteristics like the “borrower-based credit effect” overcome institutional factors in determining default probability, and underlying factors that affect individual capacity to repay such as “location” or the “prevailing . . . labor market” are what “may actually lie behind the apparent school effect”); Volkwein & Cabrera, supra note 264, at 109–13 (finding, inter alia, that incidences of default at certain institutions are primarily attributed to certain minority groups and “organizational characteristics of institutions” do not have a significant effect on default); Volkwein et al., supra note 260, at 226, 231–33 (indicating that in controlling for student characteristics, institutional characteristics had no significant impact on probability of default and certain institutions tend to attract students with greater default risk that is beyond the control of the institution, dependent instead on personal characteristics); cf. Gov’t Accountability Off., Postsecondary Education: Student Outcomes Vary at For-Profit, Nonprofit, and Public Schools 5–8 (2011) (finding worse outcomes at proprietary institutions after controlling for student characteristics).
uniform pricing encourages high-risk behavior. Risk-based pricing could improve efficiency by forcing borrowers to internalize risk and thereby cause them to make more responsible choices. The most obvious example would be choice of courses and major, which is almost entirely under the student's control. Indeed, students from disadvantaged backgrounds already disproportionately choose fields of study linked to higher post-graduation wages. Risk-based pricing would benefit these students by reducing the total cost of their educations.

In other situations, relative risk levels may be driven by factors that are beyond the borrower's control. In such cases, we might question the propriety of compounding misfortune by charging the unfortunate a higher interest rate than the fortunate. In such situations, risk-based pricing is unlikely to improve efficiency because borrowers cannot reduce their risk levels by making different decisions.

The most obvious examples of factors that are outside the realm of choice and may predict default risk include race and parents' socio-economic status.


272. See supra note 133 and accompanying text; see also WILLIAM G. BOWEN & DEREK BOK, The Shape of the River: Long Term Consequences of Considering Race in College and University Admissions 70–72 (1998) (finding that black and white students at selective colleges were equally likely to major in engineering, natural science, and economics); id. at 99–103 (finding that black graduates of selective colleges were more likely than white graduates of selective colleges to attend law schools and medical schools, and also more likely to attend top programs in law, medicine, or business); cf. ARCIDIACONO ET AL., supra note 17, at 3 (“Although blacks and whites initially have similar interests regarding whether to major in the more strictly graded fields [of natural science, engineering, and economics], the patterns of switching result in 68% of blacks choosing humanities and social science majors compared to less than 55% of whites.”).

273. Black male college graduates earn less than white male college graduates, even after controlling for socioeconomic status, SAT scores, high school grades, college selectivity, college major, college class rank, graduate school, and sector of employment. BOWEN & BOK, supra note 272, at 144–48.

274. Students from higher socioeconomic status backgrounds earn more than those from low socioeconomic status backgrounds. BOWEN & BOK, supra note 272, at 136. The relation is partly due to higher socioeconomic students being more likely to attend graduate school. Id. However, higher SES students
strong predictors of the likelihood of default, but students do not get to choose their parents.\textsuperscript{275} Even credit scores of students will sometimes reflect the credit histories of their parents rather than choices made by the individual student.

\textbf{C. Risk-Based Pricing and Choice of Major}

As discussed above, risk-based pricing reduces the transfer of value from low-risk borrowers to high-risk borrowers by forcing all borrowers to internalize their own risks. Choice of major or field of graduate study probably represents the clearest example of student loan risk driven by a borrower’s personal decision. The data suggests that there are certain majors that are much lower risk than others, as measured by post-graduation wages and debt to income ratios—specifically, engineering, certain science and technology majors, and business majors are relatively low-risk.\textsuperscript{276} By contrast, humanities and education majors are relatively high risk.\textsuperscript{277} Many of the differences in wages across majors and related occupations persist even after controlling for differences in student ability. Risk-based pricing of student loans would encourage more college students to choose majors that would better prepare them for post-graduation employment opportunities, could reduce unemployment rates, and reduce default rates on student loans.

Some may be concerned that risk-based pricing would channel too many students who are incapable of succeeding in STEM or business into these majors and produce a surplus of low-quality scientists or engineers. The data does not support this view. Although there are differences in average standardized test scores across majors, there is substantial overlap in the distribution of abilities across majors, and many students in

\textsuperscript{275} See supra note 239 and accompanying text.

\textsuperscript{276} See supra note 108 and accompanying text; see also supra note 232 and accompanying text.

\textsuperscript{277} Supra note 108 and accompanying text.
majors with low value in the labor market should be capable of succeeding in fields that are more highly valued in the labor market. Indeed, business majors have among the lowest standardized test scores, but have above-average labor market outcomes. Figures 7.1 and 7.2 below highlight the overlapping distribution of abilities of students in different fields. Specifically, the charts show what percent of GRE test-takers intent on graduate study in various fields scored higher on the quantitative portion of the GRE than the median test taker intent on studying business (7.1) or engineering (7.2).

278. See supra notes 13–18 and accompanying text; see also Educ. Testing Serv., GRE Guide to the Use of Scores 2010–11, at 17–19 (2010), http://www.ets.org/s/gre/pdf/2010-11_gre_guide.pdf (providing in Table 4 GRE scores by intended major of those taking the test); Arcidiacono, supra note 142, at 344 (“Even after controlling for selection, large earnings premiums exist for certain majors.”); Stephen D.H. Hsu & James Schombert, Univ. of Or. Dept. of Physics, Data Mining the University: College GPA Predictions from SAT Scores (Apr. 15, 2010), http://arxiv.org/pdf/1004.2731v1.pdf (studying a sample of University of Oregon graduates and finding that SAT scores predict academic success, but even students with relatively low SAT scores are capable of succeeding in most college majors—including economics, chemistry, biology, and computer science).
Figure 7.1: Many Students Who Currently Choose Lower Value Fields Have the Ability to Succeed in Higher Value Fields, Such as Business

Percent of GRE test takers who have quantitative GRE scores that are above the average scores of students who intend to study business in graduate school, by intended graduate major, Aug. 2011 - Apr. 2012

Percent of students with high quantitative GRE scores

Note: Most graduate business students take the GMAT rather than the GRE. Those taking the GRE may intend to pursue a Ph.D. in business rather than an MBA.

Figure 7.2: Many Students Who Currently Choose Lower Value Fields Have the Ability to Succeed in Higher Value Fields, Such as Engineering

Risk-based pricing could make labor market data more salient to college students by making the long-term consequences of educational choices apparent the moment the student needs to borrow. Prior to matriculation, when students have not yet begun to specialize, students could be offered blended “institutional rates” that reflect the distribution of majors among graduates of their college programs. At the start of every subsequent semester—when students are about to borrow additional money
to pay tuition—students could be presented with a fixed interest rate for their new loans that reflects the individual students' course selections and developments in the labor market.

Students could also be given an explanation of the implications of the interest rate for the long-term cost of their education, and a list of actions they could take that would reduce their interest rate, such as changing majors and/or courses. This could be especially helpful to students from low-income backgrounds, who may be the least informed about occupational

---

279. It would be unfair and impractical to change the interest rate for outstanding loans based on new information that was not available to students at the time they borrowed. It may also be sensible to permit students to lock in a rate for a certain number of semesters or years if providing predictability is more important than encouraging mid-course adjustments.

280. It makes more sense to focus on observable behavior—such as the courses students actually complete—rather than declarations of intent (i.e., a declared major). For example, a student might receive the pre-med interest rate after successfully completing a gatekeeper class such as organic chemistry. Declarations of intent can more easily be gamed by students seeking a low interest rate for the early years of a high-risk course of study.

281. Whether risk-based pricing of student loans or economically equivalent grants would be more effective to change student behavior is an empirical question. The empirical literature on whether borrowers react rationally to interest rates is somewhat mixed. See, e.g., Sumit Agarwal, Souphala Chomsisengphet, Chunlin Liu & Nicholas S. Souleles, Do Consumers Choose the Right Credit Contracts?, 15–17 (Working Paper, 2007), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=843826 (finding that a significant proportion of borrowers make suboptimal choices when trading off between upfront fees and interest rates, although mistakes become less common as the amount of money at stake increases); Sumit Agarwal, Paige Marta Skiba & Jeremy Tobacman, Payday Loans and Credit Cards: New Liquidity and Credit Scoring Puzzles?, in 99 AM. ECON. REV., PAPERS AND PROCEEDINGS OF THE 121ST MEETING OF THE AM. ECON. ASS’N 412, 416 (2009) (concluding that some consumers will borrow using more expensive payday loans even when less expensive credit card debt is available to them); Block-Lieb & Janger, supra note 193, at 1535–48 (discussing limitations on consumer borrowers comprehension of interest rates); David B. Gross & Nicholas S. Souleles, Do Liquidity Constraints and Interest Rates Matter for Consumer Behavior? Evidence from Credit Card Data, 117 Q. J. ECON. 149, 150–52, 182 (2002) (finding that borrowers react strongly to credit card interest rates by adjusting their borrowing up when rates decrease and down when rates increase, and arguing that previous studies that have found a limited impact of interest rates suffered from inadequate data on household-specific interest rates). Much of the literature has focused on subprime consumer debtors, and it is unclear to what extent it can be generalized to college students.
wages and the most likely to borrow. And by studying toward higher wage and lower risk occupations, students from low-income backgrounds might be able to maximize the economic value of higher education and their own opportunities for upward social mobility.

We need not assume that every student chooses a course of study based purely on financial considerations. All that is required for risk-based pricing to change behavior is that some proportion of students are motivated at least in part by financial considerations, and that monetary incentives coupled with additional information could change those students’ behavior.

Over the long run, if risk-based pricing of student loans succeeds in channeling enough students toward high-demand occupations, the gap between wages in different occupations and for different college majors could shrink, reducing inequality.

D. The Promise and Perils of “Meritocratic” Risk-Based Pricing

The distinction between risks driven by student choice and risks driven by misfortune is not always clear-cut. Factors like class ranking\(^{282}\) or standardized test scores\(^{283}\) may be driven in part by choice—how much time to devote to studying and how much to leisure—and in part by innate ability or advantages related to being from a prosperous family.

However, the use of class ranking\(^{284}\) or test scores could help channel students toward the areas in which they have the greatest competitive advantage and therefore the greatest opportunity for success. For example, notwithstanding the strong average career prospects for engineers and doctors, it might be

\(^{282}\) Higher grades predict higher earnings, even after controlling for race, socioeconomic status, gender, SAT scores, college selectivity, college major, college class rank, graduate school, and sector of employment. Bowen & Bok, supra note 272, at 140–42.

\(^{283}\) SAT scores somewhat predict earnings, particularly between low scores and moderate scores. Id. at 133–35.

\(^{284}\) GPA is a poor measure because of differences in grading distributions across institutions, majors, courses, and professors. See supra notes 113–23 and accompanying text. Standardized percentile rankings within each course and aggregated by major would be more meaningful.
preferable for great writers with limited spatial abilities to pursue careers in law or journalism. Risk-based pricing that incorporates some measure of field-specific ability would do a better job of sorting students into areas that are the best fit for the individual student’s talents.

The decision to use factors such as rankings and test scores as predictors may involve a tradeoff between equality and efficiency. On average, minorities and students from less prosperous backgrounds tend to have lower test scores and grades. Nevertheless, grades and test scores remain good predictors of academic and financial success, even after controlling for race and socioeconomic status. One possible compromise would be to use class ranking or standardized test scores that have been adjusted to remove differences that might be explained by race or parental socioeconomic status. Such an approach, however, would entail subjective and potentially controversial judgments.

285. See, e.g., Richard D. Kahlenberg, Rewarding Strivers: Helping Low Income Students Succeed in College 10–13 (2010) (noting significantly lower predicted SAT scores for those who are disadvantaged and black as opposed to advantaged and white, respectively).

286. Many studies that have questioned the predictive value of the SAT have used statistically questionable techniques, such as over-controlling (by using grades or other standardized test scores, which are meant to measure academic ability and correlate with SAT scores), or truncating the range of scores by only examining students who already have very similar SAT scores because they attend the same caliber of institution. Some studies have also failed to correct for differences in grading distributions. See, e.g., Christopher M. Berry & Paul R. Sackett, Individual Differences in Course Choice Results in Underestimation of the Validity of College Admissions Systems, 20 Psychol. Sci. 822, 822 (2009) (“[T]he validity of SAT scores and high school [GPAs] as predictors of academic performance has been underestimated because of previous studies’ reliance on flawed performance indicators . . . .”); Nathan R. Kuncel & Sarah A. Hezlett, Fact and Fiction in Cognitive Ability Testing for Admissions and Hiring Decisions, 19 Current Directions in Psychol. Sci. 339, 340–44 (2010) (reaffirming the efficacy of standardized testing in predicting future academic performance despite various studies that questioned the validity of tests).

287. See, e.g., Kahlenberg, supra note 285, at 167–75, 185–90 (describing how a system of adjustment for socioeconomic factors at more extreme ends of disadvantage may be implemented with respect to the SAT and ACT).

288. For example, in the late 1990s, Educational Testing Services attempted to develop an alternative SAT “strivers” score that would flag students who
E. Debt-to-Income Ratios and Paternalistic Borrowing Limits

Student debt levels probably primarily reflect external circumstances—parental wealth, socioeconomic status, and extent of parental support—but may also somewhat reflect factors within the student’s control—whether to attend a more expensive college or accept a scholarship at a less prestigious institution, whether to live at home, whether to work during school, or whether to overload on credits to graduate early. Because some debt-reducing choices might adversely affect students’ academic performance and career prospects, risk-adjusting student loans based on debt levels may disproportionately harm high-ability students of limited means—precisely the upwardly mobile clientele that federal student loans are meant to serve.

The focus should not be exclusively on the cost of education—the focus should be on whether education provides value that exceeds its cost. Nonprofit universities with higher tuition prices generally spend more on instruction per student, and, after controlling for student characteristics, their graduates earn more money. Attempts to cap tuition, including arbitrary limits on access to student loans, could performed better than expected based on factors such as parental socioeconomic status. The original strivers project was discontinued amid heated controversy. Claire Barliant, Striving to Stay Alive, SALON (Oct. 18, 1999, 12:00 PM), http://www.salon.com/1999/10/18/strivers/ (last visited on Feb. 3, 2013) (on file with the Washington and Lee Law Review).

289. Stacy Berg Gale & Alan B. Kruger, Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables, 117 Q. J. ECON. 1491, 1524 (2002) (“We do find that students who attend colleges with higher average tuition costs tend to earn higher income years later, after adjusting for student characteristics. . . . [T]uition matters because higher cost schools devote more resources to student instruction.”).

290. William Bennett, former Secretary of Education under President Ronald Reagan, claimed that student loans and other government aid increase the cost of education, and have used these claims to justify budget cuts that target government support for education. William J. Bennett, Op-Ed, Our Greedy Colleges, N.Y. TIMES (Feb. 18, 1987), http://www.nytimes.com/1987/02/18/opinion/our-greedy-colleges.html (last visited Feb. 3, 2013) (on file with the Washington and Lee Law Review). Most empirical investigations of the “Bennett Hypothesis” have focused on grant aid rather than student loans, and have found mixed results. None of these studies have established that funding captured by universities does not ultimately benefit students through increased educational quality, student support services, or access for low-income students.
RISK-BASED STUDENT LOANS

See, e.g., Michael McPherson & Morton O. Schapiro, Keeping College Affordable: Government and Educational Opportunity 72–73 (1991) (rejecting the Bennett Hypothesis and finding that colleges tend to provide additional grant aid in response to federal grant aid); Lesley J. Turner, Columbia Univ., The Incidence of Student Financial Aid: Evidence from the Pell Grant Program 26 (2012), http://www.columbia.edu/~ljt2110/LTurner_JMP.pdf (“Across all sectors, every dollar of Pell Grant aid reduces students’ effective prices by 84 cents, with institutions appropriating the remaining 16 cents through price discrimination.”); Bridget T. Long, How do Financial Aid Policies Affect Colleges? The Institutional Impact of the Georgia HOPE Scholarship, 39 J. Hum. Resources 1045, 1062–63 (2004) (finding that private universities captured at most 30% of new grant aid in Georgia’s HOPE program, which was viewed as not rising to “the level of college exploitation insinuated by Bennett”); Michael S. McPherson et al., Recent Trends in U.S. Higher Education Costs and Prices: The Role of Government Funding, in 79 Am. Econ. Rev. 253, 255 (1989) (finding the Bennett Hypothesis implausible because tuition increased the fastest at well-endowed, elite private institutions that were the least dependent on government aid as a source of revenues); Larry D. Singell, Jr. & Joe A. Stone, For Whom the Pell Tolls: The Response of University Tuition to Federal Grants-in-aid, 26 Econ. Educ. Rev. 285, 291–94 (2007) (rejecting the Bennett Hypothesis for in-state tuition at public universities, but finding that private universities adjust tuition to capture most grant aid and public universities adjust out-of-state tuition); cf. Michael J. Rizzo & Ronald G. Ehrenberg, Resident and Nonresident Tuition and Enrollment at Flagship State Universities, in College Choices: The Economics of Where to Go, When to Go and How to Pay for It 303, 338–39 (Caroline M. Hoxyby ed., 2004) (finding that flagship public universities increase in-state tuition to absorb grant aid, “[c]onsistent with the Bennett Hypothesis,” but do not increase out-of-state tuition to absorb grant aid).

The empirical evidence for the claim that federal student loans increase education costs is weak. Alisa F. Cunningham et al., Nat’l Ctr. for Educ. Statistics, 2 Study of College Costs and Prices, 1988–89 to 1997–98, at 80–81, 10 (U.S. Dep’t Educ. 2001), http://nces.ed.gov/pubs2002/2002158.pdf; see also Glater, supra note 35, at 66 (“Empirical studies of changes in tuition do not support the assertion that colleges raise prices in response to greater perceived availability of funds to students.”); Warren, Baum & Sitaraman, supra note 65, at 141 n.71 (“Some observers suggest that the increased availability of student loans fuels increases in college prices. However, most empirical analyses fail to find such an effect.”). One of the few studies that may provide limited support for the Bennett Hypothesis focused exclusively on for-profit educational institutions, and the results therefore cannot be generalized to nonprofit higher education. The results may also be explained by grant aid rather than loans, or by unobserved differences in institutional costs or quality. Stephanie Riegg Cellini & Claudia Goldin, Does Federal Student Aid Raise Tuition? New Evidence on For-Profit Colleges 1, 13–25 (Nat’l Bureau of Econ. Research, Working Paper No. 17827, 2012), http://scholar.harvard.edu/goldin/files/does_federal_student_aid_raise_tuition_new_evidence_on_for-profit_colleges.pdf.

There are more plausible explanations for rising costs of higher education, such as an economy-wide increase in demand for educated labor and therefore an increase in costs for all service industries that rely on highly educated labor,
degrade the quality, availability and value of education.

Borrowing limits are a heavy-handed approach to managing risk, and impinge on the freedom of students and their access to university education. There are less intrusive approaches available, such as risk-adjusting interest rates to account for default risk.

In some extreme and rare situations, higher interest rates may be inadequate to offset losses because default rates are already very high and there are too few nondefaulting borrowers in the same risk pool. In such rare situations, loan limits could include medicine, dentistry, and legal services, as well as higher education. Robert B. Archibald & David H. Feldman, Why Do Higher Education Costs Rise More Rapidly than Prices in General?, CHANGE, May–June 2008, at 30–31. For a discussion of factors contributing to the costs of higher education, see generally Robert B. Archibald & David H. Feldman, Why Does College Cost So Much? (2011).

Another explanation for perceived increase in costs is a shift in costs from taxpayers to students and their families, as per-student real public support for higher education has generally declined since the early 1980s. Evidence suggests that the decline in government support for education is linked to the growth of anti-taxation political movements. See Robert B. Archibald & David H. Feldman, State Higher Education Spending and the Tax Revolt, 77 J. HIGHER EDUC. 618 (2006). Tuition sticker prices also increase as universities charge wealthy students more to fund need-based aid—and lower net cost—for poorer students. See, e.g., Archibald & Feldman, supra at 150–53.


293. Glater, supra note 35, at 72–73 (arguing for higher federal student loan limits to increase access and reduce the need for high-cost private loans).

294. See Block-Lieb & Janger, supra note 193, at 1513–18 (explaining how the growing use of risk-based pricing by lenders enables them to profitably make more credit available to risky borrowers and mitigates the need for credit rationing).

be used as a last resort. However, loan limits can only be fully effective as a paternalistic measure if they apply to all debt students might turn to if access to federal student loans is restricted—including private student loans, credit cards, home equity loans and other sources of credit.296

To the extent that loan limits are used at all, the following risk-based principles should be applied: Education programs and majors that are linked to higher post-graduation incomes and higher employment rates should have higher loan limits than programs and majors that are linked to lower post-graduation incomes and lower employment rates. All else being equal, educational programs that require fewer years to complete should have higher borrowing limits per year, and higher total borrowing limits. An educational program that can produce an equally productive—and equally well-paid—skilled worker in fewer years is worth more than an educational program that takes longer to produce the same economic result. Expanding loan limits for productive programs would enable students to pay a premium for efficiency, and encourage universities to become more efficient.

F. Risk-Based Pricing and Institutional Autonomy

Risk-based pricing could also be used to change the behavior of educational institutions. As discussed above, educational institutions currently have financial incentives to channel students away from classes and majors that are expensive to teach—because the instructors have skills that are valuable in the labor market outside the universities—and toward classes that are less expensive, because they are less valuable in the

If risk-based pricing emphasized the market value of different majors, and students responded accordingly by shifting toward majors linked to higher-income employment and better job prospects, universities would feel pressure to shift more resources toward teaching marketable skills. The government could dissuade colleges from using grades to channel students away from expensive majors by requiring any educational institution that accepts federal student loans to disclose percentile class rankings on any documents that disclose grades or GPA. If percentile ranking data became widely available and well understood, graduate schools, employers, and students themselves would likely turn away from letter grades and toward more meaningful and standardized percentile rankings. Because every class would be subject to the same percentile distribution scheme, grading distributions could not so readily be used to alter enrollments to the benefit of universities and the detriment of students, employers, and taxpayers.

Risk-based pricing could also make the allocation of educational resources more salient to prospective students. For example, entering freshmen, who will not yet have declared a major or taken any classes suggesting a specialty, could be offered an interest rate that reflects a weighted average based on the majors of upper level students or graduating seniors at their college—a reasonable proxy for the allocation of educational resources at that particular institution. These “institutional rates” could be made publicly available to help prospective students choose between the institutions to which they have been admitted, and to help shift students toward the institutions that are most responsive to the needs of the labor market.

Shifting resources toward courses of study that are in demand and have a high value in the labor market will entail real and substantial costs for universities. To offset these costs, it may be necessary to increase student loan limits and to accept

---

297. See supra notes 108–20 and accompanying text.
298. Or possibly a variant such as the weighted measures developed by Valen E. Johnson. See supra notes 113–23 and accompanying text.
tuition increases, with the understanding that although the cost of higher education may increase, the value of higher education will increase by even more.

Some may object to risk-based pricing on the grounds that it is likely to shrink enrollment in, and resources dedicated to, the humanities in order to increase enrollment and resources dedicated to STEM and business fields. We might also be concerned about the use of risk-based pricing as a cover for politically motivated interference with university research. The ability to target particular universities or particular departments within universities might raise such concerns. The more tightly risk-based pricing is tied to objectively verifiable loan loss data or data on occupational wages and employment—as opposed to forecasts or any other subjective criteria—the lower the risk of veiled attacks on academic freedom and impartial research.

Some have argued that the humanities pay off economically in the long run, even if humanities graduates do not do as well in the short run. Better data would be needed to evaluate these claims—most studies of income by major rely on a few years of post-graduation data. Risk-based pricing can and should consider not just employment and income at graduation or the first few years thereafter, but longer-term economic outcomes. Such long-term data could be gathered by a cooperative effort between government and universities to match schooling records with student loan performance data and federal income tax and Social Security Administration records. It is perfectly plausible that at least some humanities graduates go on to have successful, stable, and lucrative careers, for example, as lawyers or other professionals. Risk-based pricing is not inherently in favor or

---


300. Student and taxpayer privacy concerns could be addressed by releasing only aggregate data, or obscuring information that could be used to identify individuals.

301. See Simkovic & McIntyre, supra note 186 (finding that average lifetime earnings of law degree holders, discounted to present value at the start of law school, are approximately a million dollars greater than earnings of similar workers whose highest degree is a Bachelor’s).
against any particular discipline—it is in favor of allocating resources according to the needs of a dynamic labor market.

Figure 8.1: Some Majors May Provide Better Opportunities to Boost Earnings with Additional Work Experience or Graduate Education

Median earnings by college major, age, and education attainment 2009-2010
Real 2011 USD thousands

- Graduate Degree Holder (age 30-54)
- Experienced College Graduate (age 30-54)
- Recent College Graduate (age 22-26)

Source: U.S. Census Bureau, American Community Survey 2009 & 2010; Anthony P. Carnevale, Ban Cheah, & Jeff Strohl, Georgetown University Center on Education and the Workforce, Hard Times, College Majors, Unemployment, and Earnings: Not All College Degrees Are Created Equal (Jan. 2012).
Figure 8.2.1: Workers with Undergraduate Degrees in Some Fields with Low Starting Salaries Are Likely to Attend Law School or Medical School

<table>
<thead>
<tr>
<th>Undergraduate Major</th>
<th>Propensity for Pursuing Professional Degrees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Political science</td>
<td>24.7%</td>
</tr>
<tr>
<td>Biology</td>
<td>20.7%</td>
</tr>
<tr>
<td>Chemistry</td>
<td>13.0%</td>
</tr>
<tr>
<td>History</td>
<td>11.5%</td>
</tr>
<tr>
<td>Philosophy &amp; theology</td>
<td>9.8%</td>
</tr>
<tr>
<td>Economics</td>
<td>9.0%</td>
</tr>
<tr>
<td>Psychology</td>
<td>7.2%</td>
</tr>
<tr>
<td>English</td>
<td>7.0%</td>
</tr>
<tr>
<td>Foreign languages</td>
<td>6.1%</td>
</tr>
<tr>
<td>Sociology</td>
<td>4.1%</td>
</tr>
<tr>
<td>Chemical engineering</td>
<td>3.2%</td>
</tr>
<tr>
<td>Secondary Education</td>
<td>2.6%</td>
</tr>
<tr>
<td>Maths</td>
<td>2.6%</td>
</tr>
<tr>
<td>Accounting</td>
<td>2.5%</td>
</tr>
<tr>
<td>Fine arts</td>
<td>2.5%</td>
</tr>
<tr>
<td>Physics</td>
<td>2.3%</td>
</tr>
<tr>
<td>Business administration</td>
<td>2.2%</td>
</tr>
<tr>
<td>Mechanical engineering</td>
<td>1.9%</td>
</tr>
<tr>
<td>Social work</td>
<td>1.8%</td>
</tr>
<tr>
<td>Nursing</td>
<td>1.6%</td>
</tr>
<tr>
<td>Electrical engineering</td>
<td>1.6%</td>
</tr>
<tr>
<td>Physical Education</td>
<td>1.4%</td>
</tr>
<tr>
<td>Marketing</td>
<td>1.2%</td>
</tr>
<tr>
<td>Civil engineering</td>
<td>0.7%</td>
</tr>
<tr>
<td>Elementary education</td>
<td>0.7%</td>
</tr>
<tr>
<td>Computer science</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

Figure 8.2.2: Workers with Undergraduate Degrees in Some Fields with Low Starting Salaries Are Likely to Attend Law School or Medical School

Propensity for Pursuing Professional Degrees by Undergraduate Major, 2009
Percent of college graduates aged 18 and over with professional degree


Note: Preprofessional majors, not shown, have the highest rates of professional school attendance, at 23.4% law degrees and 26.8% medical degrees.
Figure 8.3: The Most Valuable Graduate Degree Fields Are Medicine, Computers, Engineering, Law, and Business

Annualized median earnings by advanced degree field, 2009
Population age 18 and over where highest degree is an advanced degree
2011 USD thousands

Source: US Census Bureau, Survey of Income and Program Participation, 2008 Panel, Table 4H.
Note: Annualized earnings calculated by multiplying monthly earnings by 12.
Figure 8.4.1: Among Those with a Law Degree, Workers with High-Value Undergraduate Degrees Earn the Most

Earnings of workers with a law degree, by undergraduate major

Earnings as a percent of earnings for economics majors, workers aged 35 to 55


Note: * No statistically significant difference compared to economics majors.
Figure 8.4.2: MBAs with High-Value Undergraduate Majors Generally Earn More Than MBAs with Low-Value Undergraduate Majors

<table>
<thead>
<tr>
<th>Undergraduate Major</th>
<th>Earnings as a Percent of Earnings for Economics Majors, Workers Aged 35 to 55</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical engineering*</td>
<td>108.2%</td>
</tr>
<tr>
<td>Economics</td>
<td>100.0%</td>
</tr>
<tr>
<td>Electrical engineering*</td>
<td>99.0%</td>
</tr>
<tr>
<td>Aerospace engineering*</td>
<td>96.6%</td>
</tr>
<tr>
<td>Maths*</td>
<td>94.6%</td>
</tr>
<tr>
<td>Political science*</td>
<td>92.1%</td>
</tr>
<tr>
<td>English*</td>
<td>91.2%</td>
</tr>
<tr>
<td>Industrial engineering</td>
<td>89.9%</td>
</tr>
<tr>
<td>Civil engineering*</td>
<td>89.5%</td>
</tr>
<tr>
<td>Computer science*</td>
<td>88.2%</td>
</tr>
<tr>
<td>Finance</td>
<td>85.5%</td>
</tr>
<tr>
<td>Chemistry</td>
<td>83.2%</td>
</tr>
<tr>
<td>Mechanical engineering</td>
<td>82.7%</td>
</tr>
<tr>
<td>Marketing</td>
<td>82.0%</td>
</tr>
<tr>
<td>Accounting</td>
<td>81.1%</td>
</tr>
<tr>
<td>Business...</td>
<td>80.2%</td>
</tr>
<tr>
<td>Other business</td>
<td>80.2%</td>
</tr>
<tr>
<td>Physics</td>
<td>79.3%</td>
</tr>
<tr>
<td>Biology</td>
<td>77.7%</td>
</tr>
<tr>
<td>Psychology</td>
<td>77.4%</td>
</tr>
<tr>
<td>Sociology</td>
<td>72.4%</td>
</tr>
<tr>
<td>History</td>
<td>70.3%</td>
</tr>
<tr>
<td>Foreign languages</td>
<td>68.8%</td>
</tr>
</tbody>
</table>


Note: * No statistically significant difference compared to economics majors.
A rather more parochial argument is that, notwithstanding relatively poor job prospects of humanities graduates, the humanities have spiritual or moral value that makes them inherently superior to STEM or business or social science majors and therefore deserving of subsidies at every other field's expense. There are few empirical studies to support this view—much of the literature on “over-education” suggests that education that does not enhance employment prospects produces cynicism and dissatisfaction among graduates. Many of those who subscribe to the view that the humanities should be privileged simply hold it as an article of faith.

Assuming arguendo that the humanities are spiritually sacred but economically marginal, then who should make the economic sacrifice to ensure that the humanities are taught? Should students of limited means be forced to mortgage their futures to pay for a humanities education of limited monetary value? Should poor students in more challenging and less spiritually rewarding disciplines be forced to overpay for their loans so that other students—generally from wealthier families—can enjoy the humanities? This is the way the economic burden is allocated under the current system, and even the most ardent supporter of the humanities would be hard-pressed to defend it.

If four years of postsecondary cultural edification really is a fundamental human right, shouldn't higher education be funded through taxation and provided free at the point of service to every citizen? Perhaps, but where should such an expenditure rank in voters’ list of priorities? In a world of limited resources, choices must be made and priorities established.

Most European and Asian governments that fund higher education through taxation have not treated the humanities as sacred—they have generally prioritized STEM instruction and labor market needs to a greater extent than have U.S. students

302. See, e.g., Val Burris, The Social and Political Consequences of Overeducation, 48 AM. SOC. REV. 454, 459–61 (1983) (finding that overeducation results in a statistically significant but small reduction in job satisfaction); id. at 463–64 (finding that overeducated workers are less likely to believe that success is the result of hard work rather than luck, have a less positive view of labor unions, and are more likely to be status-conscious).
More direct government funding of postsecondary education in the United States would probably accelerate a shift away from the humanities, away from education as consumption, and toward education as investment. Risk-based student loan pricing could be understood as a compromise—an attempt to promote educational alignment with labor market needs while respecting U.S. cultural and political preferences for decentralized decision making.

For better or worse, the United States has opted not to make higher education a basic human right paid for at public expense, but rather an economic investment primarily paid for by students and their families. The federal student loan program was established with clear policy goals: to provide skilled labor to meet the needs of a growing economy and to provide upward mobility by enhancing the earnings and employment prospects of young adults. It is inappropriate for the federal student loan program to subsidize programs that cannot meet these goals at the expense of programs that can.

Other options remain for funding the humanities. Colleges can still try to convince students that the spiritual splendor of a humanities degree really is worth the financial sacrifice. With risk-adjusted pricing, students will be more likely to be fully informed of the financial risks of their decisions, and must be willing to internalize them, including the upfront cost of tuition, the foregone future income, and the long-term risk-adjusted student loan interest. With better information and proper incentives, students remain free to make whatever decision they think best.

If colleges are truly dedicated to the proposition that everyone should have access to the humanities regardless of resources, colleges could make difficult budgetary decisions to reduce net tuition charges for humanities students of limited means. Or, colleges could admit more students from wealthy backgrounds who may prefer a leisurely and spiritually

303. See, e.g., CAROLINE KEARNEY, EFFORTS TO INCREASE STUDENT’S INTEREST IN PURSuing SCIENCE, TECHNOLOGY, ENGINEERING, AND MATHEMATICS STUDIES AND CAREERS 7–10 (2011) (discussing national strategies of several European countries that include promotion of STEM learning).
rewarding college experience, and can pay for it without borrowing.

The concept of higher education as primarily a spiritual experience comes from a time when higher education was a luxury available only to the wealthy. It may not suit a world of democratized access in which loans have become the funding mechanism of choice.

VII. Conclusion

Federal student loan programs were established to provide skilled labor to employers, to facilitate higher wages and economic advancement for students, and to promote human capital development and economic growth within the United States. These programs have successfully increased college degree attainment rates, boosted the incomes of millions of graduates, and benefited employers as well as the federal budget.

Unfortunately, limited labor market transparency and skewed incentives have contributed to a mismatch between the allocation of educational resources and the demands of the labor market. Many students do not know which majors and programs are the best investment until they have nearly completed their studies, and many universities find it more convenient to channel students toward whatever can engage them at the lowest cost rather than whatever is most valuable to students and employers. Perversely, uniform pricing of student loans subsidizes the subject areas that are least economically valuable, while penalizing those that are most valuable.

Risk-based pricing of student loan interest rates would help clarify the links between educational choices and employment opportunities. It would force students to internalize the risks created by their own decisions, and would pressure universities to become more responsive to employers’ needs. Ideally, it would channel educational resources to where they are valued most, reduce structural unemployment, reduce student loan loss rates, and boost wages and tax revenues. In the long run, it may even

304. See supra note 30 and accompanying text.
reduce wage inequality by channeling more workers toward the highest wage areas, fewer toward the lowest wage areas, and thereby causing wages to compress.

Risk-based student loan pricing preserves a considerable degree of autonomy for students and educational institutions. Students remain free to study whatever they wish—they need only internalize the risk. Educational institutions remain free to allocate resources as they see fit, subject only to the constraints of a better-informed and more market-savvy population of students.

Risk-based pricing does not require an assumption that the existing allocation of wages and employment opportunities is fair or even fully efficient. It simply rests on an assumption that market prices contain useful information about the extant allocation of resources, that is unfair to channel indebted students into lifelong financial sacrifices they may not fully understand (and in so doing, exacerbate those sacrifices by driving down wages in already low wage fields), and that possible inefficiencies in the wage structure are best corrected through the normal workings of a mixed economy—regulation and taxes designed to curb socially harmful activity, and wage subsidies and public sector employment designed to provide public goods. Risk-based pricing does not create the allocation of resources in the U.S. economy—it merely reflects political and economic reality.

Within the umbrella framework of risk-based pricing, there remain a variety of possible solutions and tradeoffs. Should risk-based pricing attempt to channel students toward their areas of competitive advantage by incorporating some measure of field-specific ability? If so, what “meritocratic” measures are most predictive and least likely to undermine the goals of equal opportunity? Should risk-based pricing take into account debt levels to reward institutions that are cost efficient and students who are price conscious? Or would focus on debt levels

---

305. The suggestion is not that the government should directly set wages in the private sector. Rather, if there is a specific industry or activity in which wages are too high because of negative externalities, then the appropriate solution would be to regulate or tax the harmful industry or activity, which would increase noncompensation costs and indirectly reduce wages or employment in fields associated with negative externalities.
disproportionately harm the poorest students who have the greatest need to borrow? To what extent should risk-based pricing reflect historic data, and to what extent should it deviate based on forecasts? Which data sources are the most reliable?

This Article introduces the basic concept of risk-based pricing in the student loan context, and outlines some of the key technical and ethical considerations. It is the beginning of a long conversation, and hopefully, a path toward ameliorating some of our nation’s greatest economic and social challenges.
THE WASHINGTON AND LEE LAW ALUMNI ASSOCIATION
STUDENT NOTES
COLLOQUIUM
Is Hedge Fund Adviser Registration Necessary to Accomplish the Goals of the Dodd–Frank Act’s Title IV?†

Luther R. Ashworth II∗

Table of Contents

I. Introduction ................................................................. 652

II. Hedge Funds in General .................................................. 655
    A. Defining Hedge Fund ............................................. 655
    B. Legal Framework ............................................... 660

III. (Attempted) Regulation Before the Dodd–Frank Act .......... 661
    A. Traditional Securities Regulation: Historical
        Exemptions for Hedge Funds ............................. 662
    B. 1990s: Major Hedge Fund’s Collapse Spurs
        Regulation Debate ............................................ 669
    C. 2004: Push for Direct Hedge Fund Regulation ........... 675

IV. Direct Hedge Fund Regulation Under Title IV
    (PFIARA) of the Dodd–Frank Act ............................. 679
    A. Two Failed Hedge Funds Kick-Off Financial
        Crisis ............................................................... 679
    B. Dodd–Frank Title IV (PFIARA) ............................ 683
       1. What Is the Point of the PFIARA? ..................... 684

† This Note received the 2012 Roy L. Steinheimer Award for outstanding student Note.
∗ J.D., Washington and Lee University School of Law, expected May, 2013; B.S., Virginia Tech, 2010. I would like to thank Professor Christopher Bruner for his invaluable advice and guidance. Many thanks to Professor Lyman Johnson and George Mason Law Professor J.W. Verret for commenting on my Note at our Notes Colloquium. I would also like to thank my colleagues on the Washington and Lee Law Review—a special thanks goes out to Brandt Stitzer for taking the time to provide fresh perspective and insight throughout the Note-writing process. Lastly, and most importantly, I want to thank my family (Munford, Kelly, Reid, and Berkley) and Jen for their unwavering love and support.
I. Introduction

Hedge funds have avoided direct regulation under federal securities laws for most of their existence. The hedge fund industry has gained a reputation for being secretive and opaque mainly because information available on hedge funds is scarce. Since the collapse of a massive hedge fund in the late 1990s, however, hedge funds have been targeted for increased regulation. Over the decades, the federal government has provided a variety of policy reasons in favor of regulating hedge funds. The ebb and flow of hedge fund scrutiny, and the policy

1. See infra Part II.A (discussing how hedge funds have historically avoided traditional federal securities laws).


3. See infra Part II.B (discussing the implications that the collapse of the massive hedge fund Long-Term Capital Management (LTCM) had on hedge fund regulation).
rationales behind hedge fund regulation, have predictably correlated with major financial events.4

In 2007, two Bear Stearns hedge funds that had largely invested in mortgage-backed investments5 collapsed.6 This signaled a deteriorating U.S. mortgage market that would eventually lead to the financial meltdown known as the subprime mortgage crisis of 2008 (Financial Crisis).7 Soon after, the Financial Crisis caused the United States to fall into a deep recession.8 Hedge funds’ involvement in the Financial Crisis, like the hedge fund industry itself, is not fully understood.9 Nevertheless, the Financial Crisis finally tipped the scales back toward hedge fund regulation.

In 2010, Congress passed the Dodd–Frank Wall Street Reform and Consumer Protection Act (Dodd–Frank Act)10 to provide stability to the damaged U.S. financial system.11 Although several areas of the Dodd–Frank Act affect hedge funds, this Note focuses on Title IV of the Act, entitled the

4. See SCOTT J. LEDERMAN, HEDGE FUND REGULATION § 3:1 (2011) (noting that “the laws and rules to which hedge funds and their managers must adhere are found in a variety of statutes and regulations, reflective, to a large extent, of the gradual emergence of the hedge fund”).

5. See infra notes 187–95 and accompanying text (explaining the basics of mortgage-backed investments and their involvement in the financial crisis of 2008).


7. See Matthew Beville, Dino Falaschetti & Michael J. Orlando, An Information Market Proposal for Regulating Systemic Risk, 12 U. PA. J. BUS. L. 849, 891 (2010) (“Though it was unclear at the time, the collapse of Bear Stearns’s funds was the first sign that the mortgage market was collapsing and that a large number of financial firms were overexposed to asset-backed securities and related derivatives.”).

8. See BAINBRIDGE, supra note 6, at 5 (noting that the “economy sank into the deepest recession in decades” because of the Financial Crisis).

9. See George et al., supra note 2, at 359–60 (noting that some argue that “hedge funds were among the contributors to the fiscal crisis” of 2008, while others argue “that the hedge fund industry did not play a direct precipitating role in the events leading to the financial meltdown”).


11. Id.
Private Fund Investment Advisers Registration Act of 2010 (PFIARA). The PFIARA directly regulates the hedge fund industry by requiring certain hedge fund advisers to register with the Securities and Exchange Commission (SEC) under the Advisers Act of 1940 (Advisers Act). This Note evaluates whether hedge fund adviser registration is necessary in light of the Financial Crisis and the goals of the PFIARA (and the Dodd–Frank Act generally), and if so, what form that regulation should take.

Part II provides an introduction to hedge funds by focusing on their history, general structures, and legal frameworks. Part III discusses regulatory aspects of hedge funds prior to the Financial Crisis and the enactment of the Dodd–Frank Act. Specifically, Part III focuses on how hedge funds avoided regulation over the years and looks to specific events that spurred interest in hedge fund regulation. Part IV explains the basics of the Financial Crisis and hedge funds’ involvement in it. Part IV then details the implications of the PFIARA’s enactment. Part V analyzes whether hedge fund adviser registration under the Advisers Act is necessary in light of the PFIARA’s goals. Next, Part V provides recommendations for hedge fund regulation going forward. Finally, Part VI offers conclusions.

Ultimately, this Note proposes that hedge fund adviser registration under the Advisers Act is unnecessary to advance the PFIARA’s goals because (i) there is already an adequate hedge fund anti-fraud rule in place; (ii) hedge funds have increased transparency to investors over the years; and (iii) hedge funds have a sophisticated investor class that does not need the same protections provided to ordinary investors. Because hedge fund adviser registration is unnecessary to the PFIARA’s goals, it is a waste of the hedge fund industry’s and the SEC’s resources.

The collection of systemic-risk-related data from hedge funds, however, is necessary in light of the Financial Crisis, but
adviser registration is not needed to achieve this task. This Note asserts that once a threshold (based on hedge fund size) is determined for an aggregate group of hedge funds most pertinent to systemic risk assessment, the Financial Stability Oversight Council (FSOC) should collect data directly through Form PF. Collecting data from smaller hedge funds that do not meet the determined threshold will produce an over-inclusive regime. At the same time, this Note argues that once a proper threshold is established, no hedge funds with assets under management (AUM) exceeding the determined threshold should be exempt from providing information related to systemic risk. Because size is critical to assessing systemic risk concerns, exempting any hedge funds with AUM exceeding the determined threshold would produce an under-inclusive element to the regime as well. This Note explores detailed recommendations to alleviate these issues.

II. Hedge Funds in General

Part II of this Note addresses historical and legal aspects regarding hedge funds before the Dodd–Frank Act. Part II.A gives a history of hedge funds, while also discussing the general structures and investment strategies of more popular hedge funds. Part II.B examines hedge funds’ general legal frameworks.

A. Defining Hedge Fund

Hedge funds are hard to define because of their diverse, complex, and secretive trading strategies. Typically, “hedge funds” refer to private funds that pool the assets of wealthy and institutional investors “to invest and trade in equity securities, fixed-income securities, derivatives, futures and other financial
instruments.” Most hedge funds are professionally managed and carry high levels of debt to increase certain investment positions with the intention of amplifying gains. Hedge funds are sometimes referred to as alternative investments because usually they are not publicly traded, not very liquid, and difficult to value. Also, despite the fact that hedge funds vary broadly in investment strategy, they all strive to achieve absolute return, which means that their “strategies are designed to generate positive return regardless of overall market performance.”

Most agree that Alfred Winslow Jones, a Columbia University sociologist, pioneered the modern day hedge fund in 1949. Jones created a fund that would “hedge” against market risks by offsetting declining values of long-stock positions with appreciating values of short-stock positions and vice versa. Also, Jones borrowed money for a portion of the fund’s investments to increase leverage (debt-to-equity) in the hopes of magnifying returns. Because Jones’s fund outperformed the leading mutual

17. See Verret, supra note 2, at 803 (“High leverage, management expertise, performance fees, and absolute return strategies are the hallmarks of the industry.”).
18. See Lederman, supra note 4, § 1:3 (noting that alternative investments differ from traditional investments because they are “generally not traded on a public market and therefore tend to be less liquid and more difficult to value”).
19. Id.
21. See Lederman, supra note 4, § 1:1 (discussing that “Jones reasoned that complementing a long portfolio with short positions would provide a ‘hedge’ against the influence of market movements on his portfolio as market-generated declines in the value of the long portfolio would be offset by similarly generated gains in the short portfolio”). This is because a long position appreciates when the value of the held security rises, while a short position appreciates when the underlying security’s value decreases. Id. Note that a long position generally refers to a speculative position in an asset that is purchased and held with the hopes that the asset’s value will rise over time. Id. Jones typically obtained his short positions by short selling stocks; this is where Jones would sell a stock that he did not own (borrowed the sold stock) and then would replace the sold stock once the price fell to make a profit. Id. In sum, Jones was taking long positions in stocks that he thought were undervalued and short positions in stocks that he thought were overvalued to reduce “the prospect of losses by taking a counterbalancing transaction.” Edwards, supra note 20, at 190.
22. See Lederman, supra note 4, § 1:1 (detailing how borrowing, or
funds of that era, the hedge fund concept became increasingly popular.\textsuperscript{23}

In 1970, an estimated 150 funds were managing over $1 billion in assets.\textsuperscript{24} By 2007, many factors—including the creation of new financial markets and instruments, an influx of capital, and an increase in the number of hedge funds—led to an estimated $1.93 trillion in total AUM for hedge funds.\textsuperscript{25} Nevertheless, the Financial Crisis negatively impacted the hedge fund industry both in performance\textsuperscript{26} and reputation.\textsuperscript{27} Although the hedge fund industry has slowly recovered, total AUM “remain below their peak level in 2007.”\textsuperscript{28} Still, hedge funds play a key role in the U.S. economy, and it is estimated that there are almost “ten thousand hedge funds currently operating and managing $1.5 trillion in assets.”\textsuperscript{29}

Hedge funds’ general structures are diverse, but the majority of funds share some common characteristics. Most hedge funds require a significant initial minimum investment from their investors\textsuperscript{30} and restrict their investors from withdrawing capital.

\begin{itemize}
  \item leverage, “can magnify returns” by increasing the amount in a certain position, but it can also magnify losses if the leveraged position moves contrarily of where the investor intended.
  \item See id. (explaining that as Jones’s fund outperformed the leading mutual funds of his day, the hedge fund concept “generated interest with the investment community”).
  \item Id.
  \item See id. (discussing how new financial markets and instruments in the 1980s, combined with an increase in capital in the 1990s, spurred hedge fund growth).
  \item See id. (examining how the Financial Crisis, among other reasons, led to “negative performance as well as record withdrawals by investors which combined to result in a decline in total assets managed in the hedge fund industry . . . at the end of 2008”).
  \item See Wulf A. Kaal, Hedge Fund Regulation Via Basel III, 44 VAND. J. TRANSNAT’L L. 389, 391 (2011) (discussing that in the aftermath of the Financial Crisis, “[h]edge funds have been blamed for their part in the crisis and have become a scapegoat for the problems affecting many aspects of the financial markets”).
  \item Lederman, supra note 4, § 1:1.
  \item Scott V. Wagner, Comment, Hedge Funds: The Final Frontier of Securities Regulation and a Last Hope for Economic Revival, 6 J.L. ECON. & POL’Y 1, 3 (2009).
  \item See Edwards, supra note 20, at 191 (noting that hedge funds usually have high minimum-investment requirements and giving an example of a large hedge fund in the late 1990s that required a $5 million minimum investment).
\end{itemize}
during a fixed “lock-up” period. The lock-up period is designed to encourage long-term investment, while ensuring the fund’s liquidity by limiting withdrawals. Typically, hedge funds compensate their managers in ways that highly reward gains. Hedge fund managers usually require a 1–2% fee for all AUM to cover operating costs, while also requiring a 15–25% performance fee of all profits made in a given year. From a managerial standpoint, this is a very attractive feature because mutual funds and other institutional funds usually pay flat fees. To combat excessive risk-taking, and to further align the interests of management and clients, most hedge funds force managers to invest a certain amount of personal capital into the fund as well. There is no doubt, however, that the possibility of enormous profits has attracted great talent to the hedge fund industry and has driven the development of creative investment strategies typical of hedge funds.

31. See id. (explaining that most hedge funds limit the ability of their investors to withdraw funds so that the managers can invest in more illiquid instruments over longer periods of time). The term “lock-up” refers to the minimum holding period that the investors will have to wait until they can withdraw funds. Lederman, supra note 4, § 2:3.3.

32. See id.

33. See Wagner, supra note 29, at 3 (“Hedge fund managers structure their funds to create internal incentives that maximize returns.”).

34. See Edwards, supra note 20, at 191 (noting that hedge fund administrative fees usually range from 1–1.5% of assets under management, while large incentive fees range from 15–20%). Many funds also impose a “hurdle rate,” which requires fund managers to exceed a minimum rate of return before the manager’s performance fee will actually kick in. Lederman, supra note 4, § 2:3.3. Furthermore, some hedge funds subject their managers to “high water marks,” which require fund managers to cover prior years’ losses before earning a performance fee. Id.

35. See Edwards, supra note 20, at 191 (noting that mutual funds and other institutional investors are usually prohibited from using incentive performance fees, so they often have to employ a flat fee rate).

36. See Lederman, supra note 4, § 2:2.4 (pointing out that a major distinction between hedge funds and mutual funds is the fact that most “[h]edge fund managers usually invest a significant portion of their own liquid net worth in their hedge funds alongside of the fund’s other investors”).

37. See Verret, supra note 2, at 828–29 (stating that mutual funds have had a hard time competing with hedge funds for professional talent because of hedge fund fee structures).
As noted, hedge funds are extremely diverse because of their flexible nature. Exploring the basics of three of the more popular hedge fund strategies, however, will shed light on this alternative investment. One type of hedge fund uses a hedge-equity strategy similar to Alfred Jones’s long- and short-position model, but in the modern world, its focus is more specified—such as a country-specific equity market focus or an industry equity market focus. Another type of hedge fund, called a global-opportunistic hedge fund, looks to exploit macroeconomic factors in different countries or regions. These hedge funds are more event-driven—fund managers may look at political or currency trends—and use their managers’ discretion or advanced computer systems to find developing trends. A third type of hedge fund, known as an arbitrage (or relative-value) hedge fund, looks to “exploit pricing inefficiencies between or among related instruments.” This very small sample of hedge fund strategies shows just how diverse and flexible the hedge fund industry is as a whole.

38. See Sue A. Mota, Hedge Funds: Their Advisers Do Not Have to Register with the SEC, but More Information and Other Alternatives Are Recommended, 67 La. L. Rev. 55, 57–58 (2006) (“While many trade in securities, bonds, and currencies, some also trade in derivatives and other assets, such as movies and even the rights to soccer players.”).

39. See Lederman, supra note 4, § 1:2.1 (noting the hedged equity strategy is similar to the Jones model, but has evolved over time to focus more on areas like country-specific and industry-specific equity markets). It is estimated that 30% of hedge funds use a strategy similar to the hedge equity strategy. Id.

40. See id. § 1:2.2 (explaining that the global-opportunistic strategy looks at macroeconomic data to speculate on factors such as political or currency trends).

41. See id. (“Global macro managers have historically been known for making high risk, significantly leveraged investments that are often directional in nature rather than being hedged.”); see also Verret, supra note 2, at 803 (“They may trade commodities or currency swaps based on macroeconomic data, or trade on expected results of a merger or acquisition between two companies.”).

42. Lederman, supra note 4, § 1:2.3.
B. Legal Framework

The vast majority of domestic hedge funds are set up as limited partnerships or limited liability companies. These structures allow them flexibility in terms of the relationship between their managers and their investors. These are the most popular legal structures for hedge funds because they offer flexibility in governance (and management), limited liability, and certain tax advantages. Most hedge funds form in Delaware, “where they are subject to Delaware fiduciary duties.” Delaware provides hedge funds with legal predictability because it has well-developed case law for both limited partnerships and limited liability companies.

Most of the legal implications facing hedge funds before the Dodd–Frank Act hinged on the fact that sophisticated, wealthy investors have traditionally made up the majority of hedge fund investors. Historically, the government has viewed this affluent class of investors as having the capabilities necessary to assess the risks associated with hedge funds. For this reason, before the Dodd–Frank Act hedge funds were largely unregulated.
Some argue that minimal regulation in the hedge fund industry has provided a true alternative investment for investors because fund managers can creatively adapt to changing financial markets without the fear of competitors immediately copying their investment strategies.\textsuperscript{51} This freedom encourages the development of diverse financial products because hedge funds can produce unique investment strategies that may have little (or no) correlation with traditional financial benchmarks.\textsuperscript{52} Furthermore, hedge funds encourage broad and efficient markets by increasing trade in both established and less-established markets.\textsuperscript{53} Price discovery of nontraditional assets is possible because larger markets (or markets in general) are created for nontraditional assets that normally would be extremely difficult to value.\textsuperscript{54} Thus, liquidity becomes possible for traditionally illiquid assets.\textsuperscript{55} Opponents of hedge fund regulation argue that “[t]he lack of regulation has been paramount to the hedge fund’s success.”\textsuperscript{56}

\textbf{III. (Attempted) Regulation Before the Dodd–Frank Act}

Predictably, hedge fund regulation correlates with the growth and popularity of the hedge fund industry—the interest in regulating the hedge fund industry has grown with the development of the industry itself.\textsuperscript{57} Because hedge funds have

\begin{itemize}
\item \textsuperscript{51} See Wagner, supra note 29, at 5 (noting that the limited regulation of hedge funds has allowed them to “adapt better to dynamic markets by encouraging research and development of new and creative financial models”).
\item \textsuperscript{52} See Verret, supra note 2, at 804 (pointing out that hedge funds typically have higher than average returns that “do not correlate with returns from the long Standard and Poor’s 500 . . . or other traditional benchmarks”).
\item \textsuperscript{53} See Wagner, supra note 29, at 6–7 (discussing that hedge funds provide large amounts of capital and help increase efficiency in both traditional and nontraditional markets).
\item \textsuperscript{54} See id. at 6 (“Funds provide a means for a large amount of cash to enter nontraditional investments and help force assets to their true valuations.”).
\item \textsuperscript{55} See id. at 7 (explaining that markets become “more efficient as assets move closer to true valuation”). Furthermore, “with more capital in the market, investors can more easily trade through the increased liquidity.” Id.
\item \textsuperscript{56} Id.
\item \textsuperscript{57} See supra note 4 and accompanying text (noting that laws and rules
\end{itemize}
evolved over a number of decades, hedge fund regulation is scattered throughout various statutes and regulations.\textsuperscript{58} Part III.A analyzes how hedge funds have historically avoided traditional securities regulation. Next, Part III.B discusses past events that spurred interest in hedge fund regulation. Then, Part III.C explores past attempts to directly regulate hedge funds prior to the Dodd–Frank Act.

\textit{A. Traditional Securities Regulation: Historical Exemptions for Hedge Funds}

Four pieces of legislation—the Securities Act of 1933,\textsuperscript{59} the Securities Exchange Act of 1934,\textsuperscript{60} the Investment Company Act of 1940,\textsuperscript{61} and the Investment Advisers Act of 1940—make up the core federal securities laws applicable to hedge fund regulation.\textsuperscript{63} Nevertheless, prior to the Dodd–Frank Act, hedge funds generally were able to avoid regulation under these laws through various exemptions.\textsuperscript{64}

\begin{itemize}
\item \textsuperscript{58} See Lederman, \textit{supra} note 4, \S 3:1 ("As a result of the evolutionary development of the hedge fund, the regulation of these financial vehicles cannot be found in one concise codification.").

\item \textsuperscript{59} See 15 U.S.C. \S\S 77a–77aa (2010) (requiring that any offer or sale of securities be registered with the SEC pursuant to the Securities Act, unless an exemption from registration exists under the law).

\item \textsuperscript{60} See id. \S\S 78a–78mm (creating various regulations on U.S. financial markets and establishing the Securities and Exchange Commission (SEC) as the federal agency primarily responsible for enforcement of U.S. federal securities law).

\item \textsuperscript{61} See id. \S\S 80a-1 to 80a-64 (directing the SEC to regulate investment companies and securities exchanges).

\item \textsuperscript{62} See id. \S\S 80b-1 to 80b-21 (establishing federal laws to regulate and monitor investment advisers based on shareholder complaints of fraud).


\item \textsuperscript{64} See Mota, \textit{supra} note 38, at 58 (noting that hedge funds “often escape regulation because they fall within the exemption provisions” of the traditional U.S. federal securities laws).
\end{itemize}
The Securities Act of 1933 (Securities Act) regulates primary market transactions in which a business entity (issuer) offers and sells its securities publicly to raise money. Importantly, the Securities Act requires issuers offering public securities to file a registration statement with the SEC. The registration statement serves as a disclosure mechanism that allows potential investors to gather information regarding “the issuer's business, properties, material legal proceedings, directors and officers, ownership, and financials.” To raise money, hedge funds offer interests in their funds that usually fall within the Securities Act’s definition of “security.”

Recall that hedge funds typically form as hybrid entities, such as limited partnerships or limited liability companies. The Securities Act’s definition of security does not specifically mention any hybrid entity interests. Nevertheless, hedge fund interests offered to investors usually qualify as securities because they are interpreted as “investment contracts.” An investment contract, as defined by the Supreme Court, involves “investment in a common venture premised on a reasonable expectation of profits to be derived from the entrepreneurial or managerial

65. See Stephen J. Choi & A.C. Pritchard, Securities Regulation: Cases and Analysis 38 (3d. ed. 2012) (explaining that the Securities Act “focuses on primary market transactions” and “requires issuers making a public offering to file mandatory disclosure documents containing information deemed important to investors”).

66. See 15 U.S.C. § 77e(c) (2010) (“It shall be unlawful for any person . . . to sell or offer to buy through the use or medium of any prospectus or otherwise any security, unless a registration statement has been filed as to such security . . . .”).

67. See id. § 77g (detailing specific information required in the registration statement).

68. See id. § 77b(a)(1) (defining the term security under the Securities Act as any “note, stock, treasury stock, security future, bond, debenture . . . investment contract,” and many other financial instruments).

69. See supra notes 43–45 and accompanying text (noting that the vast majority of domestic hedge funds are set up as hybrid entities, such as limited partnerships or limited liability companies).

70. See supra note 68 and accompanying text (providing the definition of security under the Securities Act).

71. See Hammer, supra note 15, at 111 (explaining that most hedge fund interests offered to investors are interpreted as investment contracts, which is the “catch-all” category of the Securities Act’s security definition).
Hedge funds offer their limited partnership and limited liability company interests to passive investors who receive nominal (or no) management authority—and because these investors expect profits from management (efforts of others), most hedge fund offerings qualify as investment contracts.

Still, the majority of hedge funds avoid registration under the Securities Act because they use the private securities offerings exemption (private offering exemption). Under § 4(2) of the Securities Act, any “transactions by an issuer not involving a public offering” are exempt from having to comply with the Securities Act’s disclosure and registration requirements. Furthermore, in 1983, the SEC promulgated Regulation D under the Securities Act to clarify and provide more predictability for when funds trying to use the private offering exemption are exempt from registration. Rule 506 of Regulation D is most applicable to hedge funds because it provides a safe harbor for the private offering exemption in certain situations. For example, Rule 506 exempts from registration hedge fund offerings that meet certain conditions, including only offering interests to

72. United Hous. Found., Inc. v. Forman, 421 U.S. 837, 852 (1975). Note that the quoted investment contract test in Forman is derived from the original investment contract test in SEC v. Howey. SEC v. Howey, 328 U.S. 293 (1946). The original Howey investment contract test states, “The test is whether the scheme involves an investment of money in a common enterprise with profits to come solely from the efforts of others.” Id. at 301.

73. See Mota, supra note 38, at 59 (detailing how most hedge fund securities offerings may avoid registration under the Securities Act using the private offering exemption found in § 4(2) of the Securities Act).

74. See 15 U.S.C. § 77d(2) (2010) (establishing that the registration requirements provided in the Act do not apply to “transactions by an issuer not involving any public offering”).


76. See Choi & Pritchard, supra note 65, at 550 (explaining that the SEC promulgated Regulation D of the Securities Act to “provide issuers greater certainty in private placements”).

77. See 17 C.F.R. § 230.506 (establishing a safe harbor for the § 4(2) exemption in the Securities Act).

78. See id. § 230.506 (detailing conditions that must be met in order for an offer or sale of securities to be exempt from registration requirements under Regulation D of the Securities Act).
purchasers that qualify as “accredited investors.” Most hedge funds comply with these limits and avoid registering their offerings under the Securities Act. In fact, the SEC has proposed rules to implement the newly enacted Jumpstart Our Business Startups Act that make Rule 506 even friendlier to hedge funds by eliminating the prohibition on general solicitation in private placements so long as the only purchasers are accredited investors.

The Securities Exchange Act of 1934 (Exchange Act) created the SEC and contains provisions that are relevant to hedge funds. Specifically, the Exchange Act regulates secondary market transactions in which “one investor resells securities of the issuer to another investor.” The Exchange Act requires periodic reporting to the SEC for certain companies that qualify under the statute (usually publicly-traded companies). Some of
the Exchange Act’s provisions could affect hedge funds issuing equity.87 For example, a hedge fund would have to make extensive disclosures, through periodic reporting, to the SEC if it qualified as a public company under § 12(g)88 of the Exchange Act.89 To trigger § 12(g) and its associated rules, a hedge fund would have to issue equity interests to over 2000 persons—or 500 persons who are not accredited investors—and the fund’s assets would need to exceed $10 million.90 Regardless, this threshold has been easy for hedge funds to avoid.91 Thus, the Exchange Act has not played a large part in the history of hedge fund regulation.

Another traditional U.S. securities law, the Investment Company Act of 1940 (Investment Company Act), has the potential to regulate hedge funds. Many financial funds, such as mutual funds, register as investment companies under the Investment Company Act.92 Registered investment companies are limited in the types of transactions they can use.93 For example, registered investment companies are restricted in their use of short sales and must obtain shareholder approval for investing in certain assets and borrowing substantial money.94 As discussed above, “[t]hese transactions are core elements of most hedge

87. See Mota, supra note 38, at 59–60 (discussing that the Exchange Act’s periodic reporting requirements could be relevant to hedge funds).
88. See 15 U.S.C. § 78l(g) (outlining when an issuer under the Exchange Act must register and providing certain exemptions).
89. See Mota, supra note 38, at 59–60.
91. See Mota, supra note 38, at 60 (“Most hedge funds . . . avoid registration under the 1934 Act . . . .”).
92. See Goldstein v. SEC, 451 F.3d 873, 875 (D.C. Cir. 2006) (noting that mutual funds must register with the SEC under the Investment Company Act).
93. See id. (“The Investment Company Act places significant restrictions on the types of transactions registered investment companies may undertake.”).
94. See 15 U.S.C. § 80a-12(a)(3) (2010) (stating that it is unlawful for a registered investment company “to effect a short sale of any security,” except in certain situations, contrary to the rules and regulations promulgated by the SEC under the Investment Company Act); see also id. § 80a-13(a)(2) (stating that in most situations a registered investment company, without majority shareholder approval, cannot “borrow money, issue senior securities, underwrite securities issued by other persons, purchase or sell real estate or commodities or make loans to other persons”).
funds’ trading strategies.”95 The Investment Company Act requires investment companies—defined as almost any issuer which is “in the business of investing, reinvesting, or trading in securities”96—to register with the SEC and disclose investment activities, investment policies, and other information.97 Although hedge funds fall within the definition of an investment company, the Investment Company Act has two exemptions available to most hedge funds.98 First, any investment company that is not owned by more than 100 investors and does not plan to make a public offering of its securities is exempt.99 Hedge funds generally do not make public offerings, so this exemption is favorable to hedge funds with less than 100 investors.100 Second, the Investment Company Act exempts investment companies exclusively owned by “qualified purchasers.”101 “This allows hedge funds owned solely by “qualified purchasers” to circumvent

95. Goldstein, 451 F.3d at 875.
96. See 15 U.S.C. § 80a-3(a)(1) (stating that an investment company under the Investment Company Act is a broad term that means, but is not limited to, “any issuer which is or holds itself out as being engaged primarily, or proposes to engage primarily, in the business of investing, reinvesting, or trading in securities”).
97. See id. § 80a-8(b) (stating that an investment company must disclose investment activities, investment policies, and other information when an investment company is required to register with the SEC under the Investment Company Act).
99. See 15 U.S.C. § 80a-3(c)(1) (explaining that “[a]ny issuer whose outstanding securities . . . are beneficially owned by not more than one hundred persons and which is not making and does not presently propose to make a public offering of its securities” is not recognized as an investment company under the Investment Company Act).
100. See Goldstein, 451 F.3d at 876 (explaining most hedge funds “are exempt . . . because they have one hundred or fewer beneficial owners and do not offer their securities to the public”).
101. See 15 U.S.C. § 80a-3(c)(7) (2010) (noting an exemption for registration under the Investment Company Act is allowed for “[a]ny issuer, the outstanding securities of which are owned exclusively by persons who, at the time of acquisition of such securities, are qualified purchasers”). Generally, a “qualified purchaser” is any person or family-owned company owning more than $5 million in investments or any person who owns and invests on a discretionary basis $25 million or more. Id. § 80a-2(a)(51)(A).
registration under the Investment Company Act, even if the fund has more than 100 investors.

The fourth traditional federal securities regulation relevant to hedge funds is the Investment Advisers Act of 1940 (Advisers Act). The Advisers Act requires all investment advisers, including hedge fund advisers, to register with the SEC and disclose information such as compensation, the adviser’s balance sheet, the scope of the adviser’s authority, and other data. 102 Prior to the Dodd–Frank Act, however, an exemption excused advisers managing less than fifteen clients from having to register with the SEC under the Advisers Act (private-adviser exemption). 103 To qualify for the private-adviser exemption, hedge fund advisers historically argued that each hedge fund they managed only counted as one client (rather than counting every investor in every hedge fund managed by the adviser). 104 So, hedge funds could manage up to fourteen different hedge funds, regardless of the number of investors in each fund, and still be exempt from registration under the Advisers Act. 105 Then, in 1984, to the delight of hedge fund advisers, the SEC passed a safe harbor rule that explicitly allowed private fund managers to count each fund managed as one client. 106

Nevertheless, Title IV of the Dodd–Frank Act (PFIARA) has repealed and amended sections of the Advisers Act. 107 Most important to this Note (and as discussed below), the Dodd–Frank Act eliminated the private-adviser exemption from the Advisers Act. 108 Although hedge funds have eluded traditional federal

102. See 15 U.S.C. § 80b-3(a)–(c) (detailing disclosure and registration requirements for investment advisers under the Advisers Act).

103. See Wagner, supra note 29, at 12 (stating that the exemption under the Advisers Act allowed an exemption for advisers with fewer than 15 clients).

104. See id. (“Traditionally, hedge fund advisers avoided registration under the Advisers Act by arguing that fund managers maintain only one client, the hedge fund itself.”).

105. See Kaal, supra note 27, at 414 (“Even the largest hedge fund managers usually ran fewer than fifteen hedge funds and were, therefore, exempt.”).

106. See Mota, supra note 38, at 62 (“In 1985, the SEC adopted a rule allowing investment advisers to count each pooled investment vehicle as a single client.”).

107. See infra Part IV and accompanying text (discussing the implications of the Dodd–Frank Act’s amendments to the Advisers Act relating to the regulation of hedge funds).

108. See infra Part IV and accompanying text (same).
securities laws for decades, the Dodd–Frank Act’s amendments to the Advisers Act are the most direct regulation of hedge funds to date.

B. 1990s: Major Hedge Fund’s Collapse Spurs Regulation Debate

For decades, hedge funds avoided SEC registration and most regulation under the traditional federal securities laws without controversy. In the late 1990s, however, the collapse of the massive hedge fund Long-Term Capital Management (LTCM) sparked debate regarding the need for hedge fund regulation.109

A group of highly reputable traders formed LTCM in 1994.110 The fund had starting equity of $1.3 billion ($100 million of which was contributed by the general partners) and required outside investors to invest at least $10 million.111 At LTCM’s peak, in 1997, the fund grew to larger than $7 billion after the fund made returns of 19.9% in 1994, 42.8% in 1995, 40.8% in 1996, and 17.1% in 1997.112 LTCM typically used an investment strategy that held “long positions in bonds that it considered undervalued and short positions in bonds that it considered overvalued.”113 Based on the yield spread between its positions in high- and low-risk bonds, LTCM would essentially bet on the spread to widen or narrow using derivatives contracts.114

109. See Edwards, supra note 20, at 200 (noting that the collapse, near bankruptcy, and bailout of LTCM in late 1990s sparked conversation about whether there was a need for additional hedge fund regulation).

110. See id. at 199 (noting that the general partners included a former head of bond trading at Salomon Brothers, a former vice chairman of the Federal Reserve Board, and two Nobel Prize recipients for work in the pricing of financial instruments).

111. See id. at 197 (explaining that LTCM was formed in February 1994 with equity of $1.3 billion, of which $100 million came from its general partners, and the fund “required a minimum investment of $10 million, and no withdrawals for three years”).

112. Id.

113. Id. at 197–98.

114. See id. at 198 (explaining that LTCM would buy “high-yielding, less liquid bonds, such as Danish mortgage-backed securities” and then sell short “low-yielding, more liquid bonds, such as U.S. government bonds”). Then, if LTCM thought “the yield spread between the high and low risk bonds . . . was excessively wide,” the fund would bet on the spread to narrow using derivatives contracts. Id.
In early 1998, LTCM became convinced, for a number of reasons, the yield spread between its high- and low-risk bonds was too wide; thus, LTCM bet on the yield spread to narrow.\textsuperscript{115} LTCM borrowed $125 billion from banks (on top of the fund's then $4.8 billion AUM) and increased its leverage ratio (debt-to-equity ratio) to more than 20-to-1.\textsuperscript{116} This leverage ratio, which is extraordinarily large for any hedge fund, would magnify gains or losses depending on the widening or narrowing of LTCM's yield spread.\textsuperscript{117} Later that year, a number of circumstances instilled fear in global bond investors and there was a "stampede to quality" bonds.\textsuperscript{118} Thus, LTCM's yield spread widened (instead of narrowed) and the fund's failed bet was exposed.\textsuperscript{119}

In September 1998, the Federal Reserve Bank of New York became worried about LTCM's creditors (including banks and securities firms) that would suffer losses as a result of LTCM's collapse.\textsuperscript{120} A creditor consortium, including the government, decided that the collapse of LTCM posed "systemic risk" (due to the number of overexposed parties and the amount of money involved) and agreed to a bailout of over $3.6 billion.\textsuperscript{121} Systemic risk is defined as the risk that "an economic shock such as a

\begin{itemize}
\item \textsuperscript{115} See id. at 198 ("LTCM believed that in late 1997 and early 1998, partly as a consequence of the collapse of Asian countries in the summer of 1997, the yield spread between high and low risk bonds . . . was excessively wide.").
\item \textsuperscript{116} Id.
\item \textsuperscript{117} See id. (describing LTCM's leverage ratio in 1998 as "high by any standard"). Because the leverage ratio was so high, "[e]ven a small reduction in yield spreads would mean huge profits for LTCM." Id. On the reserve side, losses would be magnified as well. Id.
\item \textsuperscript{118} See id. at 199 ("As fear spread of what the market repercussions to . . . market breakdowns might be . . . there was a stampede to 'quality.'"). This stampede to quality meant bond investors began to dump their more risky bonds for safer (yet lower-yielding) bonds. Id. Among the leading reasons for global fear in the bond market was the result of Russia devaluing its currency and refusing to honor contracts sold to customers. Id.
\item \textsuperscript{119} See id. ("This sharp widening of yield spreads caused by a stampede to liquidity and quality was just the opposite of what LTCM was betting on.").
\item \textsuperscript{120} See id. at 200 (explaining that in September 1998, the Federal Reserve Bank of New York became aware of LTCM's potential collapse and held meetings to discuss the situation).
\item \textsuperscript{121} See id. (describing how a creditor consortium decided to bailout LTCM to the tune of $3.6 billion after the group decided LTCM's collapse posed systemic risk based on the amount of money involved and the number of overexposed lenders involved).
\end{itemize}
market or institutional failure triggers (through panic or otherwise) either the failure of a chain of markets or institutions or a chain of significant losses to financial institutions, resulting in increases in the cost of capital or decreases in its availability.”

Systemic risk can be paralleled to a domino effect in which a trigger event (here LTCM’s collapse) “causes a chain of bad economic consequences” that have the potential to bring down other financial institutions and overall markets. LTCM’s collapse was the first event that clearly demonstrated hedge funds could have systemic risk consequences. The magnitude of LTCM’s exposure to other market participants showed that a massive hedge fund’s failure could have devastating effects on the overall market.

As a result of LTCM’s collapse, the government issued two reports detailing what went wrong with LTCM and how the hedge fund industry could be better regulated. The first report by the President’s Working Group on Financial Markets recommended that more information on hedge funds should be disclosed publicly to prevent hedge funds from overleveraging their investments. This report also urged lenders to establish better standards for evaluating hedge funds when extending credit. The second report, conducted by the United States General Accounting Office (GAO), confirmed that LTCM’s massive size and leverage created systemic risk that posed a

123. Id. at 198.
124. See id. at 203 (“In LTCM, the potential for systemic risk existed not by reason of its intrinsic status as a hedge fund but by the sheer size of its exposure to other institutions and market participants.”).
125. See LEDERMAN, supra note 4, § 3:3 (“In the wake of LTCM, two significant governmental studies were issued in 1999.”).
127. See id. at 31 (“Improving transparency through enhanced disclosure to the public should help market participants make better, more informed judgments about market integrity and the creditworthiness of borrowers and counterparties.”). This is because the report thought “[t]he central public policy issue raised by the LTCM episode is how to constrain excessive leverage more effectively.” Id.
128. See id. at 30 (explaining that there was a breakdown in market discipline of lending practices during the LTCM situation).
threat to the financial system. Nevertheless, these reports focused on improving public information about hedge funds through disclosure rather than calling for the exemptions for hedge funds in traditional securities laws to be amended or repealed.

In response to these reports and the crash of LTCM, Congress proposed two bills that suggested information-gathering strategies to prevent another major hedge fund collapse, rather than direct regulation. The first bill, the Hedge Fund Disclosure Act (1999 Disclosure Bill), required “unregulated hedge funds” to submit certain information to the Board of Governors of the Federal Reserve System (Federal Reserve Board). The bill defined an unregulated hedge fund as any private fund with $3 billion or more in capital that was not registered under the Investment Company Act; this also included any family or group of pooled hedge funds with AUM of $20 billion or more. These unregulated hedge funds would have to make public quarterly reports including the funds’ total assets, derivatives positions, leverage ratios, and other measures of market risk identified by the Federal Reserve Board (and other government actors such as the SEC). Hedge funds could request that some proprietary information, such as investment

129. See Lederman, supra note 4, § 3:3 (“The second study, conducted by the United States General Accounting Office (GAO), also focused on LTCM’s extensive leverage and its potential adverse impact on the financial system as a whole.”).

130. See Mota, supra note 38, at 63 (explaining that of the reports and recommendations that were issued in the aftermath of LTCM, none of them recommended “changes to . . . the exemptions for hedge funds under” traditional federal securities laws).

131. See Lederman, supra note 4, § 3:3 (explaining that two bills came in response to LTCM’s collapse and the ensuing government studies).


133. See id. § 4 (describing that unregulated hedge funds have to submit quarterly reports to the Federal Reserve Board).

134. See id. § 3(3) (providing the definition for unregulated hedge fund under the proposed legislation).

135. See id. § 4(a) (detailing the information that an unregulated hedge fund would have to provide to the Federal Reserve Board on a quarterly basis).
strategies, be kept confidential from the public. Congress recognized that hedge funds had the potential to affect systemic risk, but it did not think direct regulation was the best option. Still, the proposed legislation called for “reliable information about hedge fund activities” to ensure that the government could prevent (or, if necessary, control) the collapse of any major hedge funds that could cause a “severe burden on the United States financial system.”

The second bill introduced was the Derivatives Market Reform Act of 1999 (1999 Reform Bill). It aimed to reduce systemic risk in the financial markets through “enhance[d] oversight over certain derivatives dealers and hedge funds.” Titles I and II of the 1999 Reform Bill dealt largely with derivatives markets. Title III, however, shared many of the same reporting requirements for certain hedge funds as the proposed 1999 Disclosure Bill. The 1999 Reform Bill required quarterly reporting to the SEC for “unregistered hedge funds,” defined as “any pooled investment vehicle, or group or family of pooled investment vehicles, that has total AUM of $1 billion or more,” and is not registered under the Investment Company Act. Thus, the threshold for reporting was $2 billion lower than

---

136. See id. § 4(c) (explaining that a hedge fund may request that “proprietary information concerning investment strategies and positions” in a quarterly report “be segregated in a confidential section of the report which shall not be available to the public”).

137. See id. § 2 (stating that the congressional findings noted that hedge funds can potentially “pose a threat to the safety and soundness of the United States and international financial systems,” but “market forces, rather than government regulations, are the best tools for constraining hedge funds”).

138. Id. § 2(8).

139. Id. § 2(7).

140. See H.R. 3483, 106th Cong. (1999) [hereinafter 1999 Reform Bill] (“Amend[ing] federal securities laws to enhance oversight over certain derivatives dealers and hedge funds, reduce the potential for such entities to increase systemic risk in the financial markets, enhance investor protections, and other purposes.”).

141. Id.

142. See id. §§ 101, 201 (providing reform for federal securities laws dealing with certain derivatives dealers and broker-dealer oversight).

143. See LEDERMAN, supra note 4, § 3:3 (noting the similarities in the quarterly reporting requirements, for certain hedge funds, of the 1999 Reform Bill compared to those of the 1999 Disclosure Bill).

the 1999 Disclosure Bill, and it had no “family or group” of pooled hedge funds distinction. The quarterly reports would include detailed financial information and “[a] description of the models and methodologies that the pooled investment vehicle use[d] to calculate, assess, and evaluate market risk.”145 This information would be made public by the SEC and shared amongst various federal agencies.146 The 1999 Reform Bill also allowed hedge funds to request that information, such as trading strategies, be kept confidential.147

Ultimately, Congress did not enact either of the two bills.148 The collapse of LTCM was soon regarded as a one-off that was unlikely to occur again.149 Many saw LTCM as an outlier to the hedge fund industry and believed banks (and other lenders) had tightened their lending practices enough to avoid another such build-up of excessive leverage.150 The hedge fund industry vowed to become more transparent to its investors, providing additional comfort that increased hedge fund regulation was unnecessary.151 Although no direct regulation of hedge funds resulted from the collapse of LTCM, the situation sparked more serious debate for hedge fund regulation.152

145. Id. § 301(k)(1)(A)–(F).
146. See id. § 301(k)(3) (describing the availability of the quarterly reports upon the SEC’s receipt under the 1999 Reform Bill).
147. See id. § 301(k)(4) (stating the procedures for the confidentiality of proprietary information under the 1999 Reform Bill).
148. See LEdeman, supra note 4, § 3:3 (noting that neither the 1999 Disclosure Bill nor the 1999 Reform Bill were enacted).
149. See id. (“As the events of 1998 receded in time, there was a growing appreciation that the facts surrounding LTCM were not representative of hedge funds in general.”).
150. See id. (explaining that in 1998 the International Monetary Fund and The President’s Working Group noted that banks, and other lenders, had improved their management of hedge fund exposures through better credit practices after the collapse of LTCM).
151. See id. (discussing how the “hedge fund industry itself responded with initiatives to improve risk management and provide greater transparency to investors”).
152. See Wagner, supra note 29, at 17 (suggesting that 2004 hedge fund regulation proposals were, in part, a result of reports and meetings that occurred after the collapse of LTCM).
C. 2004: Push for Direct Hedge Fund Regulation

The debate regarding hedge fund regulation did not seriously resurface until 2003 when the SEC called for a Hedge Fund Roundtable.\(^{153}\) At that time, the technology bubble had burst and investors were looking for alternative ways to invest their money.\(^{154}\) As a result, the SEC noticed that hedge funds were enjoying an influx of capital and the industry was growing.\(^{155}\) This led to the rise of so-called “funds of funds,” which invest in a variety of different hedge funds to maintain exposure to the hedge fund industry while diversifying portfolio allocations.\(^{156}\) The SEC became concerned that funds of funds could directly expose less wealthy—and often less-sophisticated—individual investors to hedge funds through public offerings.\(^{157}\) The SEC acknowledged, however, most of these offerings were limited to institutional investors (pension funds, public companies, etc.).\(^{158}\) Meanwhile, institutional investors also were investing more in general hedge funds, so less-sophisticated individual investors, who had invested in various institutional investors, were now being indirectly exposed to hedge funds.\(^{159}\) And, in the

\(^{153}\) See Mota, supra note 38, at 63 (noting that the SEC held a Hedge Fund Roundtable in 2003 to discuss the possibility of a hedge fund study).

\(^{154}\) See Lederman, supra note 4, § 3:4 (noting that when the bull market of the 1990s came to an end, with the burst of the technology bubble, more investors looked to alternative investments, such as hedge funds).

\(^{155}\) See Registration Under the Advisers Act of Certain Hedge Fund Advisers, 69 Fed. Reg. 72,054, 72,056 (Dec. 10, 2004) (to be codified at 17 C.F.R. pts. 275, 279) [hereinafter 2004 Hedge Fund Rule] (noting that in 2003 the hedge fund industry assets had grown 30% in the previous year and 260% in the previous five years).


\(^{158}\) See id. (explaining that only institutional investors participated in the offerings of most funds of hedge funds in 2003).

\(^{159}\) See id. at 72,058 (explaining that in the few years before 2003, “a growing number of public and private pension funds, as well as universities, endowments, foundations, and other charitable organizations, ha[d] begun to invest in hedge funds or ha[d] increased their allocations to hedge funds”).
background of all of this, the SEC had seen an increase in their enforcement actions against fraudulent hedge fund advisers.\textsuperscript{160}

By September 2003, the SEC issued a report called \textit{Implications of the Growth of Hedge Funds} (2003 Report)\textsuperscript{161} that recommended direct regulation of hedge funds through the Advisers Act based on various public policy concerns.\textsuperscript{162} The 2003 Report cited the growth of hedge funds, growth in hedge fund fraud, and broader exposure to hedge funds as reasons for direct regulation.\textsuperscript{163} The 2003 Report noted that prior SEC staff reports had studied the systemic risks posed by hedge funds, but this report chose to focus on “[t]he growth and investor protection implications of hedge funds.”\textsuperscript{164}

Accordingly, in 2004, the SEC promulgated a rule under the Advisers Act entitled the Registration Under the Advisers Act of Certain Hedge Fund Advisers (2004 Hedge Fund Rule).\textsuperscript{165} The SEC cited the 2003 Report’s policy reasons for the administrative action.\textsuperscript{166} The 2004 Hedge Fund Rule made changes to how an adviser could qualify for registration exemption under the Advisers Act.\textsuperscript{167} Recall that hedge fund advisers were traditionally exempt from SEC registration under the Advisers Act if they managed less than fifteen clients—and prior to the

\begin{itemize}
\item[160.] See id. at 72,056 (stating that by 2003 the SEC had seen a “growth in the number of [SEC] hedge fund fraud enforcement cases”).
\item[161.] 2003 SEC REPORT, supra note 156.
\item[162.] See id. at 88 (noting that the 2003 Report’s primary recommendation to the SEC was to “consider mandating federal registration of hedge fund investment advisers under the Advisers Act” based on a variety of concerns).
\item[163.] See id. at 76–88 (outlining the growth of hedge funds, the growth in hedge fund fraud, and the broader exposure to hedge funds as reasons for recommending regulation of hedge funds under the Advisers Act to the SEC).
\item[164.] Id. at 3.
\item[165.] See 2004 Hedge Fund Rule, 69 Fed. Reg. 72,054, 72,054 (Dec. 10, 2004) (to be codified at 17 C.F.R. pts. 275, 279) (stating that the SEC adopted the 2004 Hedge Fund Rule to require certain hedge funds to register with the SEC under the Advisers Act).
\item[166.] See id. at 72,055–059 (citing growth of hedge funds, growth in hedge fund fraud, and broader investor exposure to hedge funds as the primary reasons for promulgating the 2004 Hedge Fund Rule).
\item[167.] See 17 C.F.R. § 275.203(b)(3)-2(a) (2004), invalidated by Goldstein v. SEC, 451 F.3d 873 (D.C. Cir. 2006) (requiring hedge funds to count “shareholders, limited partners, members, or beneficiaries” as clients for purposes of § 203(b)(3) of the Advisers Act).
\end{itemize}
2004 Hedge Fund Rule, hedge fund advisers could count each fund they managed as one client rather than counting every individual investor in each managed fund.168 This enabled hedge fund advisers to manage fourteen separate funds, each with multiple investors, and still qualify for the private-adviser exemption under the Advisers Act.169 The 2004 Hedge Fund Rule, however, required hedge fund advisers to “look-through” each of their managed funds and actually count every individual client in each fund.170 Consequently, the 2004 Hedge Fund Rule required most hedge funds to register with the SEC by February 1, 2006.171

The SEC passed the 2004 Hedge Fund Rule by a narrow 3–2 vote.172 The two dissenting SEC commissioners did not think registration under the Advisers Act was the best alternative and argued that the SEC should have “collected and analyzed the existing information [on hedge funds] and determined what new information would be useful before imposing mandatory registration.”173 The dissent also suggested this was a misuse of the SEC’s already limited resources.174 Not surprisingly, most hedge fund advisers were unhappy about having to register with the SEC, and it did not take long before a prominent hedge fund

168. See supra notes 103–07 and accompanying text (discussing historical exemptions for hedge fund advisers under the Advisers Act).

169. See Verret, supra note 2, at 806 (noting that before the 2004 Hedge Fund Rule, “[e]ven the largest hedge fund managers usually ran fewer than fifteen hedge funds and were, therefore, exempt”).

170. See 2004 Hedge Fund Rule, 69 Fed. Reg. at 72,066 (requiring hedge fund advisers “to ‘look-through’ the funds to count the number of investors as ‘clients’ for purposes of the private-adviser exemption” under the Advisers Act).


172. See Kaal, supra note 27, at 415 (“The [2004 Hedge Fund Rule] was eventually issued by a [3–2] vote . . . .”). This was not a party-line vote as Chairman William H. Donaldson (Republican) joined Harvey J. Goldschmid (Democrat) and Roel C. Campos (Democrat) in favor of the 2004 Hedge Fund Rule. Cynthia A. Glassman (Republican) and Paul S. Atkins (Republican) dissented.


174. See id. at 72,090 (arguing that the 2004 Hedge Fund Rule would result in a misuse of the SEC’s limited resources).
manager challenged the regulation in court. In *Goldstein v. SEC*, on June 23, 2006, the Court of Appeals for the District of Columbia invalidated the look-through provision of the 2004 Hedge Fund Rule. With the 2004 Hedge Fund Rule vacated, hedge fund managers reverted back to counting each fund they managed as one client to qualify for the private-adviser exemption under the Advisers Act.

After *Goldstein* vacated the 2004 Hedge Fund Rule, the SEC did not pursue an appeal. The SEC, however, remained focused on protecting hedge fund investors from fraudulent hedge fund advisers. In 2007, instead of attempting another round of direct regulation, the SEC proposed and adopted Rule 206(4)-8 (Hedge Fund Anti-Fraud Rule) under § 206 of the Advisers Act. Specifically, the Hedge Fund Anti-Fraud Rule prevents hedge fund advisers from “making false or misleading statements to investors” or “otherwise defrauding” them. The SEC noted that it “would not need to demonstrate that an adviser violating [the Hedge Fund Anti-Fraud Rule] acted with scienter [(knowledge or

175. See Verret, supra note 2, at 809 (“In June of 2006, Philip Goldstein and his hedge fund Opportunity Partners L.P. challenged the SEC’s equation of ‘client’ with ‘investor’ in the new regulation.”).

176. See *Goldstein*, 451 F.3d 873, 884 (D.C. Cir. 2006) (holding the 2004 Hedge Fund Rule, requiring that investors in a hedge fund be counted as clients of the fund’s adviser for purposes of the private-adviser exemption from registration under the Advisers Act, was invalid).

177. See id. (vacating the 2004 Hedge Fund Rule).

178. See Testimony Concerning the Regulation of Hedge Funds: Hearing Before the S. Comm. on Banking, Hous., & Urban Affairs, 109th Cong. 9 (2006) (statement of Christopher Cox, Chairman, SEC) (recommending that the SEC promulgate an anti-fraud rule under the Investment Advisers Act after the decision in *Goldstein*).


180. See Prohibition of Fraud by Advisers to Certain Pooled Investment Vehicles, 72 Fed. Reg. 44,756, 44,756 (Aug. 9, 2007) (to be codified at 17 C.F.R. pt. 275) [hereinafter Hedge Fund Anti-Fraud Rule] (adopting an anti-fraud rule that "prohibits advisers to [hedge funds] from making false or misleading statements to, or otherwise defrauding, investors or prospective investors in those [hedge funds]").

181. Id. For the full codification of the SEC’s Hedge Fund Anti-Fraud Rule, see 17 C.F.R. § 275.206(4)-8 (2011).
The SEC decided that using a negligence standard for determining the liability of fraudster hedge fund advisers is appropriate under the Hedge Fund Anti-Fraud Rule. Furthermore, the Rule extends to all hedge fund advisers, including those exempt from SEC registration. The Hedge Fund Anti-Fraud Rule, therefore, has a lower standard than the “catch-all” anti-fraud securities rule under the Exchange Act, Rule 10b-5, which requires scienter for liability. Although the SEC did not ultimately prevail with direct adviser registration through the 2004 Hedge Fund Rule, the Hedge Fund Anti-Fraud Rule is a serious disincentive for fraudulent activity by hedge fund advisers.

IV. Direct Hedge Fund Regulation under Title IV (PFIARA) of the Dodd–Frank Act

Part IV.A looks at the basics of the Financial Crisis and the role hedge funds played in it. Part IV.B details the implications of the enactment of Title IV of the Dodd–Frank Act (PFIARA).

A. Two Failed Hedge Funds Kick-Off Financial Crisis

Although the causes of the Financial Crisis are myriad and complex (and largely beyond the scope of this Note), the general background of the crisis will help shed light on hedge fund regulation under the Dodd–Frank Act. By 2006, there was a flood of capital into the U.S. real estate market because of low interest rates and investor optimism surrounding “seemingly ever-rising housing prices.” Through securitization, which “involves

---

183. See id. at 44,759 (“[The SEC] believe[s] use of a negligence standard also is appropriate as a method reasonably designed to prevent fraud.”).
184. See id. at 44,758 (“[The Hedge Fund Anti-Fraud Rule] applies to both registered and unregistered investment advisers.”).
185. Tellabs, Inc. v. Makor Issues & Rights, Inc., 551 U.S. 308, 318 (2007) (“To establish liability under § 10(b) [of the Exchange Act] and Rule 10b-5, a private plaintiff must prove that the defendant acted with scienter, a mental state embracing intent to deceive, manipulate, or defraud.”) (citing Ernst & Ernst v. Hochfelder, 425 U.S. 185, 193–94 (1976)).
186. BAINBRIDGE, supra note 6, at 4.
pooling income-generating assets and then selling interests in the pool that derive their value from those underlying assets,"\textsuperscript{187} a broad range of investors became exposed to mortgage-backed securities (MBSs).\textsuperscript{188} In MBSs, mortgages (the underlying, income-generating asset) are pooled together and then investors can buy securities representing claims to the underlying income stream.\textsuperscript{189}

Banks used credit enhancements for MBSs, such as guarantees covering defaulting mortgages in the pool, to encourage investor confidence.\textsuperscript{190} Banks also hired credit-rating agencies to rate the securities for marketing purposes.\textsuperscript{191} The credit-rating agencies rated the MBSs using models that assumed the U.S. housing market would continually rise and mortgage default rates would remain low.\textsuperscript{192} This led to more investors exposing themselves to what they thought was a low-risk, always appreciating U.S. housing market. As demand for MBS-type investments increased, “more complex pools-of-pools (collateralized debt obligations, or CDOs)—and even pools-of-pools-of-pools (so called CDO squared)—emerged.”\textsuperscript{193} And, as money rolled into the housing market, banks “started to lend heavily to subprime mortgage borrowers with weak credit ratings” to sustain the process.\textsuperscript{194}

By 2007, mortgage defaults increased, for a variety of reasons, and these MBSs (and CDOs) began to rapidly lose

\textsuperscript{187} Christopher M. Bruner, Corporate Governance Reform in a Time of Crisis, 36 J. Corp. L. 309, 313 (2011).
\textsuperscript{188} See id. at 314 (“Between 1996 and 2007 the stock of outstanding securitized credit in the United States would expand almost five-fold . . . .”).
\textsuperscript{189} See Choi & Pritchard, supra note 65, at 164 (describing a general asset-backed security).
\textsuperscript{190} See Bruner, supra note 187, at 313 (explaining that mortgage-backed securities (MBSs) were “often bolstered by credit enhancements, including guarantees obligating the sponsoring bank to cover defaulting mortgages in the pool”).
\textsuperscript{191} See Robert C. Pozen, Too Big to Save? 49 (2010) (noting that banks hired credit-rating agencies to analyze and rate MBSs in order to be competitive with other securities).
\textsuperscript{192} See Bruner, supra note 187, at 314 (discussing how credit ratings for MBSs “were built on quantitative models assuming low default rates and rising home values”).
\textsuperscript{193} Id. at 313.
\textsuperscript{194} Bainbridge, supra note 6, at 4.
Consequently, credit-rating agencies downgraded many of their ratings for MBSs because their underlying assumptions proved misguided. Because many hedge funds and banks had taken large positions in MBSs and CDOs, this created systemic risk. At the same time, many investors had participated in credit default swaps. A credit default swap—in which one party agrees to pay the principal amount if a home mortgage defaults in exchange for a stream of payments from the counterparty—is supposed to serve a risk-hedging function for MBSs in case of default (almost like insurance). Nevertheless, many hedge funds (and other investors) used credit default swaps for speculative purposes to profit from the defaulting MBSs and declining housing market. When the housing market crashed, however, many of these credit default swaps became worthless because counterparties could not pay the large volume of credit default swaps coming due at one time. Thus, without a hedge for the toxic mortgage-backed investments that so many financial institutions held, a domino effect of financial crisis spread across the United States.

Notably, the systemic risk implications of the housing market, and the involvement of hedge funds in the crisis, came to

---

195. See id. (explaining that in 2006 and 2007 “mortgage defaults increased significantly” and “[t]he resulting deterioration in mortgage performance adversely affected mortgage-backed securities and their more complicated variants”).

196. See POZEN, supra note 191, at 60 (explaining that as subprime mortgages began to default, the credit-rating agencies downgraded the ratings of many MBSs).

197. See CHOI & PRITCHARD, supra note 65, at 166 (“The correlated positions of banks and hedge funds in CDOs and MBSs created systemic risk.”).

198. See Bruner, supra note 187, at 314 (noting that “outstanding credit default swaps—derivative contracts equally suitable for hedging risks on mortgage-related securities and speculating against them—literally skyrocketed” between 2001 and 2007).

199. See CHOI & PRITCHARD, supra note 65, at 166 (describing a general credit default swap and its typical uses).

200. See supra note 198 and accompanying text (explaining that credit default swaps can be used for speculative as well as hedging purposes).

201. See CHOI & PRITCHARD, supra note 65, at 166–67 (explaining that when the housing market declined, “the web of credit default swaps started to unravel” and many credit default swaps became essentially worthless).

202. See id. at 167 (discussing that when many banks and other institutions could no longer hedge against the deteriorating MBSs, the crisis spread).
the forefront in July 2007. At that point, two Bear Stearns hedge funds “that had invested heavily in CDOs failed.” In March 2008, Bear Stearns was bailed out when the government orchestrated a buyout by J.P. Morgan. Bear Stearns’s failure, sparked by the collapse of two of its hedge funds, was the beginning of an economic contagion that infected the United States.

Soon more serious systemic risk consequences came to light as a result of the massive amount of leverage in the financial system, “particularly among investment banks and hedge funds.” When financial institutions began suffering huge losses because of deteriorating MBSs, many institutions had to sell liquid assets to maintain required leverage ratios. Because banks and other institutions needed cash on their balance sheets, they began “calling outstanding loans of hedge funds and other institutional investors.” Many of these hedge funds were highly leveraged, so they also had to sell liquid assets to pay the banks. Furthermore, “credit became scarce and interest rates soared on short-term debt” because banks were hesitant to lend in the midst of so much financial uncertainty.

This created a ripple effect in which financial entities began selling liquid assets for cash, especially publicly-traded stocks.

---

203. See Bainbridge, supra note 6, at 4 (noting that in “July 2007, two Bear Stearns hedge funds that had invested heavily in CDOs failed”); see also Beville et. al., supra note 7, at 855–56 (noting that “systemic implications became apparent as large subprime lenders warned of significant losses” and when “Merrill Lynch seized $400 million in assets of a Bear Stearns fund that incurred heavy losses in mortgage-backed investments”).

204. Bainbridge, supra note 6, at 4.

205. See id. (“An ad hoc government rescue was hurriedly put in place, culminating in J.P. Morgan’s acquisition of Bear Stearns.”).

206. See supra note 7 and accompanying text (noting that the collapse of the Bear Stearns hedge funds signaled the beginning of the Financial Crisis).

207. Pozen, supra note 191, at 122.

208. See id. at 123 (explaining that losses in MBSs forced financial institutions to sell assets to maintain required leverage ratios).

209. Id. at 122.

210. See id. (noting that because hedge funds were highly leveraged, they also were forced to sell assets in order to pay the banks calling outstanding loans).

211. Id. at 123.

212. See id. at 122 (“Since the markets for corporate bonds and asset-backed
This process, known as “deleveraging,” creates a cycle in which firms “sell assets, [and] prices decline in response to the increased supply, creating further losses and potentially requiring additional selling.”213 Ultimately, the crisis in the housing market spread to the capital markets, and the U.S. stock market plummeted by over 38% in 2008.214 With systemic risk consequences in full effect, the Financial Crisis developed into one of the worst recessions in the United States’ history.215

B. Dodd–Frank Title IV (PFIARA)

In response to the Financial Crisis, Congress passed the Dodd–Frank Act to promote financial stability in the United States.216 One of the primary ways the Dodd–Frank Act seeks to provide this stability is by monitoring financial markets for systemic risks.217 Importantly, Congress established the FSOC218 to identify systemic risks and “respond to emerging threats” to the U.S. financial system.219 In light of these goals, and to fill what many saw as a regulatory gap, Title IV of the Dodd–Frank Act (PFIARA) was enacted to regulate hedge funds directly.220
1. What Is the Point of the PFIARA?

The PFIARA has two goals: (i) to provide better protection to private fund investors from private fund advisers; and (ii) to assess systemic risk posed by private funds. The PFIARA primarily aims to further the first goal by amending adviser registration and reporting requirements under the Advisers Act. The second goal is to be accomplished by requiring registered advisers to file certain information with the SEC that the FSOC can then use to assess systemic risk. The PFIARA defines a private fund as an investment fund that falls under the Investment Company Act (but for any exemptions under the Investment Company Act). The SEC makes clear that this definition includes hedge funds.

2. Who Has to Register Under the Advisers Act Because of the PFIARA?

The PFIARA requires all hedge fund advisers to register under the Advisers Act unless exempted. Notably, the PFIARA eliminates the private-adviser exemption under the Advisers Act.

---


223. See id. § 404, 124 Stat. at 1571–74 (codified as amended at 15 U.S.C. § 80b-3) (requiring private fund advisers registered under the Advisers Act to submit certain information to the SEC that the FSOC can use to assess systemic risk).


that hedge fund advisers traditionally relied on to avoid SEC registration.\textsuperscript{227} Instead, the PFIARA sets out a narrow list of exemptions, so that more private fund advisers have to register with the SEC.\textsuperscript{228} The first exemption for adviser registration includes any private fund adviser who manages solely private funds and has AUM less than $150 million.\textsuperscript{229} Thus, if a hedge fund adviser manages more than $150 million in assets, regardless of the number of clients in the fund(s), she must register with the SEC under the Advisers Act. States will have responsibility for hedge fund advisers with AUM between $25 million and $100 million.\textsuperscript{230} The determination of AUM is to be made annually.\textsuperscript{231}

The second exemption is for “family office” funds as defined by the SEC.\textsuperscript{232} A family office fund cannot have clients other than “family clients.” Family clients include current and former family members, certain key employees of the family office, charities funded exclusively by family clients, and other entities as deemed appropriate by the SEC.\textsuperscript{233} These funds must be exclusively controlled by one or more family members and wholly owned by family clients.\textsuperscript{234} The SEC notes that the premise behind the

\begin{itemize}
  \item \textsuperscript{227} See id. (eliminating the private-adviser exemption under the Advisers Act).
  \item \textsuperscript{228} See SEC Exemptions Release, supra note 225, at 3 (“The primary purpose of Congress repealing § 203(b)(3) [of the Advisers Act] was to require advisers of ‘private funds’ to register under the Advisers Act.”).
  \item \textsuperscript{230} See id. § 410, 124 Stat. at 1576–77 (codified as amended at 15 U.S.C. § 80b-3a) (providing the assets threshold for federal registration of investment advisers).
  \item \textsuperscript{231} See SEC Exemptions Release, supra note 225, at 90 (stating that an adviser must “annually calculate the amount of private fund assets it manages”).
  \item \textsuperscript{233} See Family Offices, Release No. IA 3220, at 6 (June 22, 2011) (defining the term family client under the family office fund adviser registration exemption).
  \item \textsuperscript{234} See id. at 30 (outlining the family ownership and control requirements for a private fund to qualify for the family office exemption under the Advisers
family office fund exemption is “to allow families to manage their own wealth.” Interestingly, the SEC takes the approach that key employees can participate in family funds, without registering under the Advisers Act, because their position and experience enable them to protect themselves as investors. Furthermore, exempting key employees allows family office funds to attract talented investment professionals. The family office fund exemption, therefore, allows families (and key employees) to manage hedge funds, regardless of AUM, without having to register with the SEC.

The third exemption is for qualifying venture capital funds’ advisers. The venture capital fund must fit into a narrow set of criteria to avail its adviser(s) of this exemption. To be exempt as a venture capital fund adviser, the private fund must generally limit leverage, represent itself as pursuing a venture capital strategy to its investors, and not offer redemption rights to investors (among other criteria). Because of its narrow language, virtually no hedge funds will qualify for the venture capital fund exemption. The final exemption, for certain foreign private advisers, is extremely narrow and beyond the scope of this Note.

---

235. Id. at 28.
236. See id. at 23–30 (discussing the SEC’s rationale for allowing key employees to be included in family office funds).
237. See id. at 28 (explaining that permitting key employees in family office funds “allows family offices to attract talented investment professionals”).
239. See id. (stating the venture capital fund adviser exemption criteria will be promulgated by the SEC).
240. See SEC Exemptions Release, supra note 225, at 9–72 (providing a full summary of the criteria required for the venture capital fund adviser exemption).
241. See Dodd–Frank Act § 403, 124 Stat. at 1571–74 (codified as amended at 15 U.S.C. § 80b-3) (providing an adviser exemption for certain foreign private advisers). To qualify for the foreign private-adviser exemption, a hedge fund adviser must have AUM less than $25 million (among other criteria). Id. Because this threshold is so narrow, this exemption is unlikely to have any significant effect on systemic risk.
3. **PFIARA Goal 1: How Will Investor Protection Be Furthered?**

As mentioned, the PFIARA’s first goal aims to provide greater investor protection by amending adviser registration and reporting requirements under the Advisers Act.\(^{242}\) Because hedge fund advisers managing over $150 million now have to register under the Advisers Act, they have to file Form ADV with the SEC. Form ADV is the form used by investment advisers to register with the SEC.\(^{243}\) The SEC states that the data collected from Form ADV is used “to protect investors” and “to create risk profiles of investment advisers.”\(^{244}\)

Form ADV, divided into Part One and Part Two, is updated by the registered investment adviser at the end of each year (some information requires more frequent updating).\(^{245}\) Part One requires information about the investment adviser’s business, ownership, clients, employees, business practices, affiliations, and any disciplinary events of the adviser or its employees (in addition to other information).\(^{246}\) Part Two requires registered advisers to provide new and prospective clients with a brochure and brochure supplements containing most of the information required in Form ADV’s Part One.\(^{247}\) All of the information disclosed under Form ADV is fully available to the public.\(^{248}\)

Interestingly, the SEC now requires all exempt private funds to file Part One of Form ADV as well.\(^{249}\) And, even though the

---

242. *See supra* notes 221–22 and accompanying text (discussing investor protection as a goal of the PFIARA).


244. **Rules Implementing Amendments to the Investment Advisers Act of 1940, Release No. IA 3221, at 54** (June 22, 2011) [hereinafter SEC Implementing Release].

245. *See id.* at 16 (describing when a registered adviser must update Form ADV).

246. *See Form ADV, supra* note 243 (explaining the general requirements of Form ADV’s Part One).

247. *See id.* (explaining the general requirements of Form ADV’s Part Two).

248. *See SEC Implementing Release, supra* note 244, at 49 (stating that all information contained in Form ADV and filed with the SEC is to be made available to the public).

249. *See id.* at 40 (explaining exempt reporting advisers still have to report
adviser is exempt from registration, the disclosed information is available to the public. The SEC explains that “Congress gave [it] broad authority under §§ 203(l) and 203(m) of the Advisers Act to require exempt reporting advisers to file reports as necessary or appropriate in the public interest or for the protection of investors.” The SEC thinks public reporting requirements will “provide a level of transparency that will help [it] to identify practices that may harm investors, will aid investors in conducting their own due diligence, and will deter advisers’ fraud and facilitate earlier discovery of potential misconduct.”

The SEC plans to monitor the data collected from Form ADV and then conduct “examinations” on advisers that raise red flags. In an examination, the SEC checks the adviser’s books and records for conflicts of interest and other misconduct. Although exempt and registered advisers are subject to examinations, former SEC Chairman Mary Schapiro stated that the SEC does “not intend to conduct routine examinations” of exempt reporting advisers because “[a]s many observers know, the [SEC] has scarce resources and it is important therefore that [it] target those resources toward the advisers actually registered.”

4. PFIARA Goal 2: How Will the Systemic Risks of Private Funds Be Assessed?

The second goal of the PFIARA, the assessment of systemic risk posed by hedge funds, is to be accomplished through a
collection of data for the FSOC to assess. The SEC will collect this data using the newly created Form PF. The SEC requires all registered hedge fund advisers under the Advisers Act to file Form PF. The SEC, however, has differentiated reporting requirements based on whether the hedge fund adviser is a “Small Private Fund Adviser” or a “Large Private Fund Adviser.” A Small Private Fund Adviser of a hedge fund has AUM between $150 million and $1.5 billion. A Large Private Fund Adviser of a hedge fund has AUM over $1.5 billion. Form PF is divided into four sections, but only Section 1 and Section 2 apply to hedge funds.

All registered hedge fund advisers (Large and Small Private Fund Advisers) must fill out Section 1 of Form PF. Section 1a requires registered hedge fund advisers to provide basic information about any hedge funds they manage. Section 2b asks for more detailed information about each fund, such as each fund’s gross and net assets, the aggregate value of its derivatives positions, and its use of leverage. This section also asks for the “percentage of the fund’s equity held by the five largest equity holders.” The SEC says Section 1b “is designed to allow the

255. See supra notes 221–23 and accompanying text (discussing the goals of the PFIARA).
256. See Reporting by Investment Advisers to Private Funds and Certain Commodity Pool Operators and Commodity Trading Advisers on Form PF, Release No. IA 3308, at 7 (October 31, 2011) [hereinafter Form PF Release] (stating that registered advisers must submit Form PF to the SEC to satisfy systemic risk reporting requirements under the PFIARA).
257. See id. at 18 (describing which investment advisers must file Form PF).
258. See id. at 20–21 (differentiating adviser reporting requirements for Form PF based on the size of the hedge fund).
259. See id. at 21 (explaining the Small Private Fund Adviser threshold under Form PF).
260. See id. (explaining the Large Private Fund Adviser threshold under Form PF).
261. See id. at 20–21 (noting which sections of Form PF are applicable to hedge funds).
262. See id. at 63 (stating that all registered hedge funds are required to fill out Section 1 of Form PF).
263. See id. at 63–65 (outlining information required by Section 1a of Form PF).
264. See id. (outlining information required by Section 1b of Form PF).
265. Id. at 65–66.
FSOC to monitor certain systemic risks for the broader private fund industry.”266 The final part of Section 1, Section 1c, gathers data on each separate hedge fund managed by the adviser, such as “each fund’s investment strategies and the percentage of the fund’s assets managed using high-frequency trading strategies.”267 In addition, advisers have to identify each hedge fund’s top trading counterparties and information on trading and clearing practices.268 The SEC states Section 1c is “designed to enable FSOC to monitor systemic risk that could be transmitted through counterparty exposure, track how different trading strategies are affected by and correlated with market stresses, and follow the extent of private fund activities conducted away from regulated exchanges and clearing systems.”269

Section 2 of Form PF requires information solely from Large Private Fund Advisers of hedge funds (AUM greater than $1.5 billion).270 The SEC tailored Section 2 to acquire additional data focused on “relevant areas of financial activity that have the potential to raise systemic concerns.”271 Section 2a requires Large Private Fund Advisers of hedge funds to give very detailed reports on the value of “assets invested (on a short and long basis) in different types of securities and commodities (e.g., different types of equities, fixed income securities, derivatives, and structured products).”272 The SEC believes this will help the FSOC monitor different asset classes typically held by hedge funds and show trends in hedge funds’ exposures.273 Section 2b requires further disclosure on each separate hedge fund managed with a net asset value over $500 million at the end of any month in the prior fiscal quarter.274 Advisers must disclose information such as portfolio liquidity, large institutional positions, posting of

266. Id. at 73.
267. Id. at 74–75.
268. See id. (describing information required by Section 1c of Form PF).
269. Id. at 76.
270. See id. at 21 (explaining large hedge fund advisers must complete Section 2 of Form PF).
271. Id. at 77.
272. Id. at 78.
273. See id. at 82 (explaining the rationale behind Section 2a of Form PF).
274. See id. at 83 (describing the requirements of Section 2b of Form PF).
HEDGE FUND ADVISER REGISTRATION

collateral by counterparties, leverage, and internal risk assessments.²⁷⁵

Large Private Fund Advisers have a quarterly reporting requirement, and Small Private Fund Advisers only have to make annual reports.²⁷⁶ The SEC designed the Large Private Fund Adviser threshold for hedge funds to gather more systemic risk information on a substantial portion of assets in the hedge fund industry.²⁷⁷ The SEC estimates that the $1.5 billion threshold will capture “over 80% of the U.S. hedge fund industry.”²⁷⁸ Comparatively, Small Private Fund Advisers of hedge funds have to report less information, less frequently, than Large Private Fund Advisers. This is because, from a systemic risk monitoring perspective, the SEC does not think additional information or more frequent reporting is justified for hedge funds smaller than $1.5 billion.²⁷⁹

All of the information gathered in Form PF is to remain confidential because of its proprietary and sensitive nature.²⁸⁰ The information, however, may be shared with other federal departments, agencies, or self-regulatory organizations within the scope of their jurisdiction, subject to confidentiality agreements.²⁸¹ The SEC said it will also coordinate with foreign financial regulators using the Form PF data.²⁸² The SEC promises to adopt controls and systems to protect the confidentiality of the collected information.²⁸³

²⁷⁵ See id. at 83–97 (providing a full summary of Section 2b reporting requirements under Form PF).
²⁷⁶ See id. at 50 (stating the frequency of reporting for Form PF based on hedge fund size).
²⁷⁷ See id. at 31 (explaining the rationale behind the Large Private Fund Adviser threshold for hedge funds on Form PF).
²⁷⁸ Id. at 31. This is 2011 data that comes from a hedge fund industry survey that the SEC has access to called HedgeFund Intelligence. Id.
²⁷⁹ See id. at 54 (explaining the rationale behind the smaller hedge fund adviser reporting requirement on Form PF).
²⁸⁰ See id. at 112 (explaining the rationale behind keeping information gathered from Form PF confidential).
²⁸¹ See id. (discussing entities that the Dodd–Frank Act allows Form PF data to be shared with).
²⁸² See id. at 11 (noting that the Dodd–Frank Act “states that FSOC shall coordinate with foreign financial regulators in assessing systemic risk”).
²⁸³ See id. at 115 (explaining potential controls and systems that the SEC may use to protect the confidentiality of Form PF data).
Ultimately, the SEC gives the acquired data to the FSOC. The FSOC uses the data to monitor hedge fund activities and trends in the hedge fund industry that relate to systemic risk. The FSOC interprets the data and decides whether hedge funds’ activities or trends “could create or increase the risk of significant liquidity, credit, or other problems spreading” across U.S. financial markets. Then, the FSOC can make recommendations to applicable regulatory agencies to “apply new or heightened standards and safeguards for a financial activity or practice” that is causing systemic risk. Also, if the FSOC finds that a particular hedge fund—based on “the nature, scope, size, scale, concentration, interconnectedness, or mix of activities” of the hedge fund—poses a systemic threat to the financial stability of the United States, it can require the Federal Reserve Board to supervise the hedge fund.

V. Analysis and Recommendations

Part V of this Note provides a final analysis of hedge fund regulation after the enactment of the PFIARA. Part V.A analyzes whether hedge fund adviser registration under the Advisers Act is necessary in light of the PFIARA’s goals. Part V.B provides recommendations for hedge fund regulation going forward.


287. Id.

A. Is Hedge Fund Adviser Registration Necessary in Light of the PFIARA?

This section analyzes whether hedge fund adviser registration under the Advisers Act is necessary in light of the PFIARA’s two goals: hedge fund investor protection and hedge fund systemic risk assessment. This Note argues that hedge fund adviser registration is unnecessary because the PFIARA’s goals can be met without it. Thus, resources are being wasted from the perspective of the hedge fund industry and the SEC.

1. Hedge Fund Adviser Registration Is Unnecessary for Investor Protection

At first blush, the PFIARA’s hedge fund adviser registration requirements do not seem particularly onerous. Much of the proprietary information that hedge fund advisers have to disclose will not be reported on Form ADV. Instead, items like trading strategies and asset positions will be reported on Form PF and kept confidential. Passed after a moment of crisis, the PFIARA integrated hedge fund adviser registration into the Dodd–Frank Act—an Act largely concerned with controlling systemic risks to avoid a national financial meltdown. Thus, the issue becomes whether adviser registration, from an investor protection standpoint, squares with the core goals of the Dodd–Frank Act and whether adviser registration is even necessary in the hedge fund industry.

The Advisers Act “is mainly a registration and anti-fraud statute.” There is nothing to suggest, however, that hedge funds’ involvement in the Financial Crisis stemmed from fraudster hedge fund advisers. Although the SEC states hedge

---

289. See supra note 221 and accompanying text (stating the two goals of the PFIARA).
290. See supra Part IV.B.3 (discussing reporting information required by Form ADV).
291. See supra Part IV.B.4 (same).
292. See supra notes 216–20 and accompanying text (discussing the goals of the Dodd–Frank Act and the circumstances surrounding its enactment).
fund investor protection is a goal of the PFIARA, it does not offer any evidence that hedge fund adviser fraud played a key role in the Financial Crisis. As discussed, many academics are reluctant to state (with certainty) the exact part hedge funds had in the Financial Crisis. But many academics agree that hedge funds contributed to the systemic risk consequences that resulted in (and from) the Financial Crisis. For example, many suggest that some hedge funds speculated against deteriorating MBSs by using credit default swaps, other hedge funds invested heavily in toxic CDOs (Bear Stearns), and many hedge funds played a large role in the deleveraging process of the financial system that further intensified the crisis. Nonetheless, it has not been proposed that hedge fund adviser fraud played a key role in the Financial Crisis.

Recall, in 2007, the SEC passed the Hedge Fund Anti-Fraud Rule, which provides a route for the SEC to bring actions against fraudster hedge fund advisers using a negligence standard. The Hedge Fund Anti-Fraud Rule, therefore, has a lower standard for liability than Rule 10b-5 under the Exchange Act. Furthermore, after the crash of LTCM, hedge funds vowed to provide their investors with more transparency. Since then, investors and counterparties to hedge funds “demand, and usually receive, disclosure to the extent it helps them assess the merits of their investments.” In addition, the SEC now requires

294. The SEC does not offer any evidence that hedge fund adviser fraud related to the Financial Crisis in its Implementing Release, the SEC Exemptions Release, or the SEC Study on Enhancing Investment Adviser Examinations following the Financial Crisis and the enactment of the Dodd–Frank Act.

295. See supra note 2 and accompanying text (acknowledging that hedge funds’ involvement in the Financial Crisis is not fully understood).

296. See supra Part IV.A (analyzing the roles, of which many academics agree, hedge funds played in the Financial Crisis).

297. See supra Part IV.A (analyzing the roles, of which many academics agree, hedge funds played in the Financial Crisis).

298. See supra notes 178–83 and accompanying text (discussing the SEC’s promulgation of the Hedge Fund Anti-Fraud Rule).

299. See supra note 185 and accompanying text (discussing Rule 10b-5’s liability standard).

300. See supra note 151 and accompanying text (explaining how the hedge fund industry vowed greater transparency to investors after the crash of LTCM).

301. Schwartz, supra note 122, at 218.
exempt hedge fund advisers under the Advisers Act to file certain information about each fund they manage on Form ADV.302 This information is publicly available, and the SEC suggests that it will provide investors with more transparency and deter adviser fraud.303

The hedge fund industry also, from an investor protection standpoint, has many safeguards. Hedge funds require high initial investments that restrict the industry to sophisticated investors.304 Hedge funds also limit interests in their funds to accredited investors305 to avoid reporting requirements under the Securities Act.306 The Dodd–Frank Act actually heightened standards for what qualifies as an accredited investor under the Securities Act in private offerings.307 This will further ensure that hedge funds offer their interests to more sophisticated parties. In sum, hedge fund adviser registration is unnecessary because (i) there is already an adequate anti-fraud rule in place; (ii) hedge funds have increased transparency to their investors; and (iii) hedge funds have a sophisticated investor class that does not need the same protections provided to ordinary investors.

In addition, there are a few other thoughts worth noting. First, there are financial and resource concerns from the perspective of smaller hedge fund advisers, and, more importantly, the SEC. Not all funds can afford to hire new compliance officers to gather the information required by adviser registration under the Advisers Act.308 Although $150 million in

302. See supra notes 249–52 and accompanying text (discussing how the SEC requires exempt private hedge fund advisers under the Advisers Act to report information on Form ADV’s Part One).
303. See supra note 252 and accompanying text (explaining the reasoning behind why the SEC is requiring exempt private hedge fund advisers under the Advisers Act to report certain information that will be publicly available).
304. See supra note 30 and accompanying text (discussing how hedge funds require significant minimum investments).
305. See supra note 79 and accompanying text (explaining the term accredited investor under the Securities Act).
306. See supra note 79 and accompanying text (explaining the term accredited investor under the Securities Act).
308. See Azam Ahmed, For Small Hedge Funds, Success Brings New
AUM seems large on the surface, some smaller hedge fund advisers note that this is just a blip in an almost $1.6 trillion industry. This threshold appears over-inclusive in a sector that the SEC estimates is dominated by funds managing over $1.5 billion in assets (over 80% of all hedge fund AUM). Also, the SEC admits that it has limited resources to carry out hedge fund adviser examinations in light of the PFIARA. In a study required by the Dodd–Frank Act, Study on Enhancing Investment Adviser Examinations, the SEC stated “[it] will not likely have sufficient capacity in the near or long term to conduct effective examinations of registered investment advisers with adequate frequency” after the PFIARA’s enactment. The increased strain on smaller funds and the SEC are additional reasons why hedge fund adviser registration under the Advisers Act is unnecessary.

Also lurking in the background of whether or not hedge fund adviser registration is necessary under the Dodd–Frank Act, is another question: Even if hedge fund adviser fraud was a legitimate reason for passing the PFIARA, is hedge fund adviser registration the best option to combat this supposed problem? Although this raises a host of issues that are likely the topic of another discussion, history suggests adviser registration has its weaknesses. The recent uncovering of the shocking, and financially devastating, Ponzi scheme of Bernie Madoff—who had voluntarily registered his hedge funds with the SEC and

---


309. See id. (explaining that one hedge fund adviser thinks $150 million funds are the “guppies” of the industry and do not pose great risks to the U.S. financial system).

310. See supra note 278 and accompanying text (detailing hedge fund data used by the SEC).


312. The author thanks George Mason University School of Law Professor J.W. Verret for pointing out this interesting dynamic as it related to this Note. Also, the author thanks Professor Verret for noting potential reasons for the hedge fund industry’s apathy towards the Dodd–Frank Act’s adviser registration amendments. See infra notes 314–16 and accompanying text.
whose fraudulent activity went undetected for years\textsuperscript{313}—suggests that adviser registration may not be the government’s solution to the kind of fraudulent activity that it fears, regardless of whether or not this registration is necessary under the Dodd–Frank Act.

Lastly, although some suggest that many hedge funds remain unconcerned about adviser registration,\textsuperscript{314} this may be more telling of another story. In an industry dominated by larger funds,\textsuperscript{315} perhaps existing funds do not mind adviser registration because smaller funds are more likely to bear the brunt of the costs.\textsuperscript{316} Further, SEC registration expenditures will certainly raise entry costs for start-up hedge funds, which may deter future competition. The hedge fund industry’s apathy towards this new legislation, therefore, could be the combination of a couple things: (i) large, existing hedge funds may view adviser registration as a vehicle that will help them maintain their top status by stifling smaller competition; and (ii) the hedge fund industry may not think that adviser registration presents much of an obstacle at all, which calls into question whether such registration serves as a deterrent for issues like adviser fraud.

\section*{2. Hedge Fund Adviser Registration Is Unnecessary for Systemic Risk Assessment}

The collection of systemic risk information related to hedge funds is a logical solution for hedge fund regulation. Hedge funds have evolved into a large part of the U.S. financial markets.

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{313} See \textit{Sec. & Exch. Comm’n, Investigation of the Failure of the SEC to Uncover Bernard Madoff’s Ponzi Scheme} 20–21 (2009), http://www.sec.gov/news/studies/2009/oig-509.pdf (“[T]he SEC received more than ample information... over the years to warrant a thorough and comprehensive examination and/or investigation of Bernard Madoff... for operating a Ponzi scheme, and... despite three examinations and two investigations being conducted, a thorough and competent investigation or examination was never performed.”).
\item \textsuperscript{315} See supra note 278 and accompanying text.
\item \textsuperscript{316} See supra note 308 and accompanying text.
\end{itemize}
\end{footnotesize}
Current hedge fund data is scattered throughout a variety of industry surveys and a lot of information about hedge funds remains opaque. The confidentiality of Form PF protects the secretive nature of the hedge fund industry and preserves the dynamic trading strategies that make hedge funds a viable alternative investment. Also, the FSOC’s wait-and-see approach (monitoring the information, studying it, and making recommendations) allows it to obtain a full grasp on the evolving issues of the hedge fund industry (such as systemic risk concerns) before acting rashly.

Nevertheless, hedge fund adviser registration under the Advisers Act is unnecessary to gather information related to systemic risk. If the government wants to collect systemic-risk-related data from hedge funds, then why not simply collect the information? Congress and the SEC could require hedge funds (deemed to impact systemic risk) to report the exact information in Form PF to the FSOC without hedge fund adviser regulation. This would save the SEC’s (admittedly) limited resources and would reduce adviser registration compliance costs for smaller hedge funds—most importantly, it would accomplish the PFIARA’s goal of assessing the systemic risks that stem from hedge funds. This practical alternative would reduce costs and provide a smooth transition because the FSOC regime is already in place.

B. Recommendations

This Note asserts that hedge fund adviser registration requirements should be eliminated under the PFIARA. As explained, hedge fund adviser registration is unnecessary from the investor protection standpoint and the systemic risk assessment standpoint alike. But continuing to gather systemic-risk data from hedge funds through Form PF is a logical solution

317. See supra note 2 and accompanying text (describing the hedge fund industry as opaque).
318. See supra note 280 and accompanying text (stating the confidentiality of Form PF).
319. See supra notes 284–88 and accompanying text (explaining how the FSOC will utilize Form PF data).
because it fills an information gap for an opaque, but important, industry. The information required by Form PF provides the FSOC with detailed information that will help with the monitoring of systemic risk.

Currently, the SEC thinks hedge funds with AUM exceeding $1.5 billion should provide the most information connected with systemic risk because these funds make up an estimated 80% of the hedge fund industry. Arguments favoring this threshold make sense because it includes a majority of the hedge fund industry. Also, this amount is similar to the thresholds proposed in the 1999 Disclosure Bill (over $3 billion AUM) and the 1999 Reform Bill (over $1 billion AUM), which both dealt with controlling hedge-fund-related systemic risks after LTCM’s collapse. Determining a threshold for what aggregate group of hedge funds would provide the most pertinent information related to systemic risk, however, is beyond the scope of this Note.

Nonetheless, once the threshold is determined, this Note asserts that only data from the group of hedge funds deemed most important to systemic risk assessment should be collected. So, if the government determines hedge funds with more than $1.5 billion AUM are most pertinent to systemic risk assessment, requiring hedge funds with less than $1.5 billion AUM to report data would be over-inclusive and a waste of resources. At the same time, any exemptions (relating to systemic risk data collection) for hedge funds exceeding the determined threshold would render the policy under-inclusive. This Note, therefore, suggests that once the line is drawn, no hedge funds with AUM exceeding the determined threshold should be exempt from providing information on Form PF. This is because “size matters” when determining systemic risk concerns.

For example, family office hedge funds with AUM that would exceed the determined threshold should not be exempt from

320. See supra note 278 and accompanying text (detailing hedge fund data used by the SEC).

321. See supra notes 131–47 and accompanying text (describing the hedge fund reporting thresholds for the two proposed 1999 bills following the crash of LTCM).

322. See Schwarcz, supra note 122, at 203 (explaining that size matters when considering the potential for systemic risk in a hedge fund).
reporting information related to systemic risk. The theory behind exempting family office hedge funds is that these extremely private funds and families (along with key investment employees of the family office fund) should be able to manage their wealth without interference.323 But what makes a tight-knit family fund managing billions of dollars less likely to affect systemic risk than a similarly situated hedge fund whose investors are not blood relatives? Famous hedge fund manager George Soros recently kicked outside investors of his Soros Fund Management hedge fund to the curb.324 Soros does not want to disclose any information in light of the PFIARA’s enactment, so he decided the family office fund exemption was in his best interest.325 The Soros Fund Management hedge fund has AUM of approximately $25 billion.326 What is to say this fund’s failure or risky investment decisions could not affect systemic risk? After all, George Soros is the same hedge fund adviser who almost single-handedly crushed the British pound sterling in currency markets by betting on its devaluation in 1992 (and subsequently made around $1 billion off the bet).327 Thus, to thoroughly evaluate systemic risks posed by hedge funds, all hedge funds that are deemed important to systemic risk assessment (whether alone or in the aggregate) should provide information to the FSOC without exception.

There are problems, however, with this approach. First, compliance costs will still be high for hedge funds. But, given the serious consequences of the Financial Crisis, it is hard to make an argument that more information is not needed on systemic

323. See supra note 235 and accompanying text (discussing the rationale behind the family office fund adviser registration exemption).


325. See id. (“The fund’s Quantum Group will complete its transition to a ‘family office’ ahead of regulatory changes . . . .”).

326. Id.

risk—especially information on a largely opaque hedge fund industry. Second, it is questionable whether the SEC and the FSOC have the resources and expertise to adequately assess the systemic risk information given by hedge funds. This is a legitimate concern, but the only practical alternative is to establish a Self-Regulatory Organization (SRO). Common “[s]ecurities SROs include national securities exchanges and securities associations registered with the SEC, such as the New York Stock Exchange.” Practitioners and academics proposed this idea prior to the Financial Crisis and the Dodd–Frank Act. The premise is that the hedge fund industry could regulate itself through a private SRO that would coordinate with government regulatory agencies. The SRO would be able to respond more quickly to evolving hedge fund trends and would have more expertise in dealing with hedge funds. There is a strong argument for self-regulation through an SRO, but Congress has largely ignored it.

The Dodd–Frank Act required the GAO to conduct a feasibility study for a hedge fund adviser SRO, but, for a variety of reasons, the report was largely dismissive of the idea. The report stated that while an SRO for hedge fund advisers is feasible, it would require legislative action and present challenges. Among other concerns, the report suggested that a hedge fund adviser SRO would present conflict-of-interest issues and funding the SRO would be expensive. The GAO report


330. See 2011 GAO REPORT, supra note 328, at 9 (providing the potential benefits of an SRO).

331. Id.


333. See GAO REPORT, supra note 328, at 11 (explaining that “the general consensus was that forming a private fund adviser SRO . . . could be done but not without challenges”).

334. See id. at 20 (noting potential problems with a hedge fund adviser SRO).
suggested that the SRO might actually “increase the overall cost of regulation by adding another layer of oversight.”335 The report was also concerned that transparency would be limited because the “SRO would be accountable primarily to its members rather than to Congress or the public.”336 Although there is a strong argument for a private SRO, the practical implications of the GAO report suggest that it is an unlikely option.

Finally, there are confidentiality concerns in providing information related to systemic risk to the SEC and the FSOC with Form PF. The hedge fund industry has a competitive advantage in creating unique investment strategies.337 These unique strategies prevail because competitors have not been able to use reverse engineering techniques to copy others’ strategies.338 This is possible because hedge funds have been able to keep proprietary information confidential. Some will argue that allowing the government to share the collected information with other agencies and foreign governments (subject to confidentiality agreements) poses a huge threat to hedge funds. The SEC has stated, however, that it will not require systemic information reporting through Form PF until it has controls and systems in place.339 Thus, the only way to advance this counterargument is to suggest a hypothetical circumstance in which the government does not follow through on its confidentiality promise. Because that argument is largely speculative, it does not hold much weight.

VI. Conclusion

For years, proponents of hedge fund regulation called for Congress to fill what they perceived as a regulatory gap. Although hedge funds avoided most regulation for decades, the Financial Crisis tipped the scales toward direct federal

---

335. Id.
336. Id.
337. See supra notes 51–52 and accompanying text (discussing competitive advantages of hedge funds).
338. See supra notes 51–52 and accompanying text (same).
339. See supra note 283 and accompanying text (explaining that the SEC plans to maintain confidentiality of Form PF data).
regulation. The enactment of the Dodd–Frank Act, and more specifically the PFIARA, forces many hedge funds to register with the SEC under the amended Advisers Act. This Note asserts, for several reasons, that hedge fund adviser registration under the Advisers Act is unnecessary to advance the PFIARA’s goals: hedge fund investor protection and hedge fund systemic risk assessment. Consequently, hedge fund adviser registration is a waste of the hedge fund industry and the SEC’s resources.

This Note, however, suggests that the collection of hedge fund data, related to systemic risk, is necessary in light of the Financial Crisis, but adviser registration is unnecessary to achieve this goal. This Note provides more practical and tailored alternatives to accomplish this task. Mainly, this Note asserts that once a threshold (based on hedge fund size) is determined for an aggregate group of hedge funds most pertinent to systemic risk assessment, the FSOC should collect data directly through Form PF. Collecting data from smaller hedge funds that do not meet the determined threshold is unnecessary because this will produce an over-inclusive regime. On the other hand, this Note also argues that once a proper threshold is established, no hedge funds with AUM exceeding the determined threshold should be exempt from providing information related to systemic risk. This will avoid an under-inclusive element to the regime as well and accomplish the PFIARA’s most important goal—hedge fund systemic risk assessment.
Review: Is Hedge Fund Registration Necessary?

J.W. Verret*

In Is Hedge Fund Adviser Registration Necessary To Accomplish the Goals of the Dodd–Frank Act’s Title IV?,1 Luke Ashworth sets out a formidable case that questions the grounds for the mandatory hedge fund registration requirement contained in the Dodd–Frank Wall Street Reform and Consumer Protection Act (Dodd–Frank Act).2 This review of his Note will offer a description of the piece’s key strengths, suggest some challenges to the solutions it offers, and close with suggestions about areas that might build on Ashworth’s work in future scholarship.

I. Review

A. The Argument

Ashworth lays out a well-researched background of the history of the hedge fund industry, prior attempts to regulate the industry, and the complex regulatory regime that previously exempted the industry from mandatory registration which was amended by the Dodd–Frank Act.

He explores the underlying justification for the mandatory registration rule by describing hedge fund crises at Long-Term Capital Management in 1999 and two Bear Stearns hedge funds during 2007. He notes how the drafters of the hedge fund registration rule were motivated by a desire to prevent investor fraud and to minimize and monitor systemic risk.

* Assistant Professor of Law, George Mason University School of Law.


Ashworth’s skepticism of the mandatory registration requirement as a response to the financial crisis is even stronger than he suggests. The Financial Crisis Inquiry Commission’s (FCIC) Final Report on the Causes of the Financial and Economic Crisis in the United States mentions hedge funds quite infrequently, and even when it does so it does not describe their involvement in any way that would support the mandatory registration provision.3 While hedge funds were active participants in markets for asset-backed securities that quickly lost value in the financial crisis,4 it is unclear that they were more aggressive than other investors. Indeed, many hedge funds also took short positions in real estate-derived assets, and companies invested in them, that can be credited with popping the asset bubble before it grew even bigger.5

It is also unclear how registration would have affected those funds that played a role in the asset bubble. For example, the FCIC Report examines the role of “hedge funds” at Bear Stearns that had significant positions in mortgage-backed securities, which ultimately brought down the investment bank in 2007 and eventually led to the government-facilitated takeover of the bank by JP Morgan in 2008.6 But Bear Stearns was not an unregulated entity—in fact it was much more heavily regulated under the Security and Exchange Commission’s (SEC) Consolidated Supervised Entity (CSE) capital regulation program than the investment adviser registration regime that hedge funds will be subject to post-Dodd–Frank.7

---


4. See, e.g., id. at 136 (noting that hedge funds often invested in asset-backed securities, such as mortgage-backed securities and collateralized debt obligations).


Ashworth accepts that large hedge funds had a prominent role in the financial crisis but argues that the registration requirement is overbroad from an alternative perspective. Ashworth effectively demonstrates that the redundancy created by Dodd–Frank, in which both the Office of Financial Research (OFR) under the Financial Stability Oversight Council (FSOC) and the SEC will serve as information coordinators, will not aid systemic risk oversight and may cause its own problems. The reader is left convinced that information-reporting requirements, for the same purpose but from different regulators, will compound compliance costs and may even create mutually inconsistent reporting requirements.

If systemic risk is the issue, it is unclear whether and to what extent creating multiple information-collection entities, in both the SEC and the OFR at the Treasury Department, will provide additional value. Where the OFR will be able to coordinate with the FSOC, charged with systemic risk determinations, the SEC will be a step removed from the central decision-making and will have to coordinate through multiple layers to obtain useful information. Redundancies in this area will not be costless, but will add multiple layers of compliance requirements and multiply the risk that proprietary trading information will be revealed to market competitors.

Ashworth’s argument would have been edified by considering the SEC’s failure to respond appropriately to tips it received and preliminary inquiries it initiated in the Madoff and Stanford cases, both generally recognized as exceptional failures by the agency to catch multi-billion dollar Ponzi schemes despite the hedge funds being voluntarily registered with the SEC, and despite the Madoff broker-dealer subsidiary having received successful compliance audits from the SEC prior to the frauds

SUPERVISED ENTITY PROGRAM, at v (2008), http://www.sec-oig.gov/Reports/AuditsInspections/2008/446-a.pdf (“The [CSE] program is a voluntary program that was created in 2004 by the [SEC] . . . . This program allows the [SEC] to supervise [certain] broker-dealer holding companies on a consolidated basis.”). Thus, the CSE program allowed the SEC to supervise Bear Stearns prior to the financial crisis in a manner that “extend[ed] beyond the registered broker-dealer to the unregulated affiliates of the broker-dealer to the holding company itself.” Id. After the collapse of Bear Stearns, the SEC stated that even with increased oversight, “[i]t is undisputable that the CSE program failed to carry out its mission in its oversight of Bear Stearns . . . .” Id. at viii.
being revealed.\(^8\) Even worse, after examining the Stanford funds, the SEC compliance staff suggested to the SEC enforcement staff that it was probably a Ponzi scheme, whereupon the SEC enforcement staff nevertheless decided against taking action.\(^9\) Though the legislative history of Dodd–Frank references both objectives of fraud prevention and systemic risk oversight,\(^10\) it remains unclear whether the mandatory registration requirement was more predominantly an investor protection measure or a systemic risk measure. If the latter, a critique of the SEC's response to the Madoff and Stanford cases remains relevant.

\[\text{B. Review of Ashworth's Recommendations}\]

While the Note offers skepticism of the ability of the SEC to monitor systemic risk, it accepts the general notion that large financial players implicate systemic concerns. It is important to note that during the financial crisis, many hedge funds made bets against the real estate bubble, in effect correcting a wayward market.\(^11\) Hedge funds have enjoyed a level of investment flexibility and freedom from regulatory rigidity that has historically led them to take a commanding role in short selling, which helps to foster a vibrant price discovery function and

---


10. See H.R. REP. NO. 111-57, at 866 (2010), http://www.gpo.gov/fdsys/pkg/CRPT-111hrpt517/pdf/CRPT-111hrpt517.pdf (stating that Title IV of the Dodd–Frank Act “expands the advisers’ reporting requirements to the SEC as necessary or appropriate in the public interest and for the protection of investors or for the assessment of risk by the FSOC”).

11. See supra note 5 and accompanying text.
provides liquidity to the market. Therefore, Ashworth’s suggestion for a compromise approach focusing on systemic risk oversight of larger firms must nevertheless take into account the possibility that his compromise approach may also inhibit those functions served by the hedge fund industry.

Ashworth accepts the popular argument that size is the key dimension of systemic risk, relying on work seeking to give preliminary definition to systemic risk by Schwarcz. But it should also be noted that another important vein of scholarship discounts the role of size in systemic risk.

As Ashworth notes, the SEC has left jurisdiction over hedge fund registration for firms with assets under management between $25 million and $100 million to states. The compromise he suggests would be complicated by the SEC’s current allocation of authority between the states and the federal government. Small firms would be registered with states, medium-sized firms would be unregistered, and larger firms would be regulated by the FSOC. This donut hole of unregulated companies may create some potentially costly unintended consequences.

The existence of even light-touch regulation pursuant to the Investment Advisers Act of 1940 or the FSOC may serve merely as a prelude to subsequent enhanced regulation. A few high profile problems in the hedge fund sector could mean the SEC or the FSOC, or the Federal Reserve as regulator of designated systemically significant financial institutions, will feel pressure to enhance its regulatory authority in this area from the existing investment adviser oversight to a new regime mirroring the investment company or publicly-traded company model, in which disclosure methodologies become mandatory, the ability of investment managers to communicate with investors and

14. See Chen Zhou, Are Banks Too Big to Fail? Measuring Systemic Importance of Financial Institutions, 6 Intl J. Cent. Banking 205, 205 (2010) (“Both the theoretical model and empirical analysis reveal that, when analyzing the systemic risk posed by one financial institution to the system, size should not be considered as a proxy of systemic importance.”).
potential investors will be limited to regulatory approved forms, and the investment strategies will be directly or indirectly regulated by the SEC. If that occurs, it would seriously damage the liquidity and price discovery functions the hedge fund industry serves.

II. Groundwork for Future Inquiry

There are two readily apparent avenues for further research that might build on Ashworth’s successful Note in future work. The newly passed Jumpstart Our Business Startups Act (JOBS Act)\(^\text{16}\) includes a lift on the general solicitation ban found in Regulation D, one which hedge funds frequently use in soliciting investments.\(^\text{17}\) Further, the fact that the hedge fund industry itself embraced the mandatory registration regime, or at least did not fight against it, presents a puzzle for which public choice theory may provide an answer.

Henry Manne’s work in the public choice dynamics of securities regulation suggests that one motivating force behind the development of the securities laws is that large incumbent firms seek to stifle competition from newer, upstart entrants to markets by supporting regulations that increase the costs of entry into markets.\(^\text{18}\) If any participants in the hedge fund industry supported the mandatory registration rule, this could explain why.

Another suggestion for further inquiry would be to consider how changes found in the JOBS Act will change Luke’s argument. Among the provisions of the JOBS Act is a requirement that the SEC lift the ban against general solicitations under Regulation D.\(^\text{19}\) This is an exemption frequently used by hedge funds. What does lifting the general

---

19. See supra note 17 and accompanying text.
solicitation ban do to his investor sophistication argument? Will the fact that hedge funds will be able to communicate with potential investors more broadly significantly alter the cost–benefit calculus behind the mandatory registration rule? Or will the fact that hedge funds will still be limited to investments by accredited investors minimize the impact of the JOBS Act on this discussion?

III. Conclusion

Ashworth’s work is provocative and insightful, and offers a unique willingness to question the conventional wisdom behind a popular item in the Dodd–Frank Act. It opens the door for future inquiry into this new avenue of securities regulation and suggests exciting opportunities for thinking about the SEC’s hedge fund regulatory regime with veins of thought in the securities regulatory, finance, and public choice scholarship.
Why Register Hedge Fund Advisers—
A Comment

Lyman Johnson*

I. Introduction

Luke Ashworth has weighed in as a young legal scholar on a subject of exceeding complexity and importance.¹ His work does what a good Note should do in being well-organized, clearly written, providing useful background, engaging existent literature, and advancing (and supporting) a thesis grounded in policy considerations. And Luke’s point of view is as unmistakable as it is provocative: the much-ballyhooed regulation of hedge fund advisers adopted in the landmark 2010 Dodd–Frank legislation² is wasteful, unnecessary, and should be modified.

Let me welcome Luke to the scholarly world, and extend to him the customary return for thoughtful work: a response. My remarks are designed less to offer critical comments than to place his subject and his views into a larger context, and to offer my perspective on why in 2010 we got the fund adviser law we did, and how that law illuminates larger and longstanding concerns about the sometimes incongruent investor protection, capital formation, and public interest goals of U.S. financial regulation.

---

* Robert O. Bentley Professor of Law, Washington and Lee University School of Law, LeJeune Distinguished Chair in Law, University of St. Thomas (Minneapolis) School of Law. The author thanks Professor Wulf Kaal for his helpful comments.


II. Investor Protection

Luke faults the investor protection rationale underlying the hedge fund adviser registration and disclosure requirements of Dodd–Frank. He points out that there was already a Securities and Exchange Commission (SEC) anti-fraud rule, adopted in 2007, and that hedge funds supposedly draw sophisticated investors who can fend for themselves. I think the investor protection case for the registration requirement is a bit stronger than Luke believes, although that in turn serves to highlight in the fund area a question that long has loomed over federal securities regulation more generally: does mandated information disclosure always protect investors, is that its real goal, and is advancing that objective always consistent with other worthy policies? But before briefly making a few points about investor protection, let me address the seeming incongruity of the traditional investor protection aim of federal securities law generally with the specific concerns of Dodd–Frank and the acute financial crisis that preceded it.

The reason for joining longstanding SEC concerns about investor protection with the sprawling systemic-financial-risk regulation of Dodd–Frank is hidden in plain view early in Luke's paper. Prior to Dodd–Frank, the SEC had failed for some time in its efforts to regulate investment pools and their advisers, as it consistently espoused investor protection in its quest to do so.

---

4. See Ashworth, supra note 1, at 654 (“In sum, hedge fund adviser registration is unnecessary because (i) there is already an adequate anti-fraud rule in place; (ii) hedge funds have increased transparency to their investors; and (iii) hedge funds have a sophisticated investor class that does not need the same protections provided to ordinary investors.”).
WHY REGISTER HEDGE FUND ADVISERS

The horrific 2008 crisis, viewed on this issue alone, was an unexpected and serendipitous regulatory opening, a crisis not to be wasted by a frustrated administrative agency. After all, more stringent securities market regulation typically is not adopted in booming and flush times; it follows on the heels of financial distress and can be overbroad and too far sweeping. In the aftermath of the historic financial meltdown of 2008–09, with victims and villains aplenty, and regulatory nuance in short supply—in contrast to the abundant regulatory shame—who could be opposed to better investor protection, whether or not faulty protection of hedge fund investors had really contributed to the crisis or, by being improved, would prevent another? For the SEC, the regulatory iron in Congress was hot and so the agency opportunistically struck, but for a cause it had sought long before our recent calamity. Luke rightly seeks in vain for the connection between the 2008 crisis and the need for fund adviser registration on investor protection grounds because there likely is none. If protection for investors was needed, it was in spite of, not because of, the crisis. But calculated timing and implausible policy connection to crisis aside, maybe there still is something to be said for better protecting fund investors.

Hedge funds and their advisers almost invariably are organized as limited partnerships or limited liability companies. Whether organized in a state in the United States or under the law of the Cayman Islands, hedge fund vehicles are not stringently regulated under any of these laws. Under U.S. law, moreover, notably Delaware, states have adopted a highly

7. For example, Professor Roberta Romano has argued that provisions of the Sarbanes–Oxley Act were not a focus of “careful deliberation by Congress” and were “enacted under conditions of limited legislative debate, during a media frenzy involving several high-profile fraud and insolvency cases” combined with “what appeared to be a free-falling stock market, and a looming election campaign in which corporate scandals would be an issue.” Roberta Romano, The Sarbanes–Oxley Act and the Making of Quack Corporate Governance, 114 YALE L.J. 1521, 1527 (2005).

8. See 2003 SEC REPORT, supra note 6, at 88–89 (recommending regulation of hedge funds by amending the Advisers Act of 1940 as a result of the SEC’s Hedge Fund Roundtable).

9. See Franklin Edwards, Hedge Funds and the Collapse of Long-Term Capital Management, 13 J. ECON. PERSPECTIVES 189, 190 (1999) (“[A] hedge fund can be organized as a limited liability company, [and] most are organized as limited liability partnerships . . . .”).
flexible, contractarian approach, allowing the fund sponsor to craft a deal document that contractually specifies investor rights. If it is not set forth in the partnership or operating agreement, investors do not have it, including ongoing access to full information. Moreover, the sponsors themselves can reduce or even eliminate their fiduciary duties, thereby removing even traditional state law safeguards for egregious adviser conduct.

It is possible that, as a result of Dodd–Frank, information in Form ADV, filed with the SEC, coupled with on-site agency examinations, will afford greater transparency to regulators and investors, deter adviser wrongdoing, and facilitate earlier discovery of past misconduct. This may extend, for example, to determining: the amount and type of a fund’s leverage, that a fund’s portfolio assets are of the kind and amount (appropriately valued) as reported and are safe, whether funds performed as stated, whether insider trading or other wrongful activity is engaged in, whether conflicts of interest exist and are handled appropriately, and so on. Investors may be able to, and perhaps they should, take better measures to fend for themselves and

10. See, e.g., Elf Atochem N. Am., Inc. v. Jaffari, 727 A.2d 286, 290 (Del. 1999) (noting that “the policy of freedom of contract underlies” both Delaware limited liability companies as well as limited partnerships).

11. For example, section 18-1101 of the Delaware Limited Liability Company Act reiterates that Delaware gives “maximum effect to the principle of freedom of contract and to the enforceability of limited liability company agreements.” DEL. CODE ANN. tit. 6, § 18-1101(b) (2012).


14. This is a professed SEC objective for the regulation. See Rules Implementing Amendments to the Investment Advisers Act of 1940, Release No. IA 3221, at 125–30 (June 22, 2011) [hereinafter SEC Implementing Release], http://www.sec.gov/rules/final/2011/ia-3221.pdf. On October 9, 2012, the SEC, Office of Compliance Inspections and Examinations, sent a letter to newly registered investment advisers explaining that certain high-risk areas will be focused on during agency examinations. See Letter from Drew Bowden, Deputy Dir., Office of Compliance Inspections & Examinations, Sec. & Exch. Comm’n (Oct. 9, 2012), http://www.sec.gov/about/offices/ocie/letter-presence-exams.pdf. These areas are marketing, portfolio management, conflicts of interest, safety of client assets, and valuation. Id.
elicit stronger protections (including monitoring mechanisms), but a tough financial cop on the beat can also be on the lookout to head off trouble and to publicly punish (and deter) troublemakers and thereby can also help to upgrade industry norms.

The preexisting anti-fraud rule noted by Luke likely curbs some wrongdoing, but it does not mandate the fuller disclosure of information required by Form ADV, either to investors or the SEC. And its proscription of misconduct is limited to its quite general phrasing. By analogy, the venerable Rule 10b-5 is not thought, in its role as an “anti-fraud” rule, to make other mandated disclosures unnecessary under the Securities and Exchange Act of 1934. That rule supplements but does not supplant more detailed disclosure obligations. Of course, any disclosure regime—particularly one aimed at an industry that historically has been opaque and loathes to divulge much information—runs the risk that information will be “sanitized” to screen out (or artfully describe) negative information or, due to lags in reporting, simply be unhelpfully “stale.” Hedge fund advisers are likely to be cautious in their initial disclosures and are sure to pay great attention to evolving practices of other registrants.

And it may be open to serious doubt whether so-called “sophisticated” or “accredited” investors truly are financially sophisticated and able to self-help as they are assumed to be. No doubt fund investors are not as financially illiterate as retail investors, as was so alarmingly noted in a 182-page study released by the SEC on August 30, 2012. But given the wealth and income definitions for accredited investors—largely unaltered over the last thirty years, anyway—the behavioral

15. For the exact language of the preexisting anti-fraud rule for hedge funds, see 17 C.F.R. § 275.206(4)-8 (2011).
16. 17 C.F.R. § 240.10b-5.
18. See, e.g., id. § 78m (stating various disclosure obligations that certain issuers of securities, such as publicly-traded companies, must file with the SEC).
20. The original definition of “accredited investor” found in Regulation D was adopted in 1982. 17 C.F.R. § 230.506. It was altered modestly by § 413(a) of Dodd–Frank to exclude the value of an investor’s primary residence in
and regulatory premise is that relatively well-off investors must be considerably more financially astute than the ordinary investor.21 Yet, although some funds, and certain sectors in particular, can and do provide mouth-watering gains when measured over selected periods, on average hedge funds, as an aggregate investment class, have underperformed risk-free Treasury bills due to an upsurge in the sheer number of funds since 2000.22 And many fund investors seem to follow the typical retail investor strategy of chasing “yesterday’s returns” and preferring large, well-known funds.23 More information alone, of course, may do nothing to stop this dynamic, but consistently low returns over a long enough period may lead to a market solution, with investors eventually migrating to other asset classes. But it does suggest that if numerous accredited investors are not particularly savvy on this front, they may be more susceptible to wrongdoing than is commonly supposed. And they certainly make attractive candidates for a good fleecing.24


22. See Fiammetta Rocco, The Success of Hedge Funds: Masterclass, ECONOMIST, July 7, 2012, at 77–78 (noting that although some hedge fund investors do very well, “on average hedge funds have underperformed even risk-free Treasury bills . . . . [b]ecause the bulk of investors’ capital has flooded in over the past ten years”) (reviewing SIMON LACK, THE HEDGE FUND MIRAGE: THE ILLUSION OF BIG MONEY AND WHY IT’S TOO GOOD TO BE TRUE (2012)).

23. Id.

24. The North American Securities Administrators Association reported in August 2012 that, for the second year in a row, Regulation D Rule 506 private offerings were, along with real estate investment schemes, the most reported products at the center of state securities enforcement actions. See Richard Hill, JOBS Act Opens New Path for Scammers, NASAA Says in Annual List of Top Schemes, 44 SEC. REG. & L. REP. 1618, 1618 (2012).
WHY REGISTER HEDGE FUND ADVISERS

(JOBS Act)\(^{25}\) in an effort to bolster another policy goal, capital formation among smaller issuers\(^{26}\)—who likely will target less savvy investors than those in hedge funds. In effect, this was a legislative effort to “jumpstart” economic activity. Although investor protection and capital formation are both key policy goals of federal securities laws,\(^{27}\) they can be in conflict. Efforts to “ease” the burden on capital formation frequently translate into less disclosure, reduced regulatory scrutiny, heightened fraud, diminished investor protection, and resulting public cynicism.\(^{28}\) The recent SEC proposal under the JOBS Act to permit general solicitation in connection with exempt offerings if, among other requirements, the securities are only sold to accredited investors\(^{29}\)—even if, apparently, they are “publicly” offered—will apply to hedge funds relying on exemptions under the Investment Company Act of 1940.\(^{30}\) Although the advertising rules of the Advisers Act of 1940\(^{31}\) remain applicable to fund advisers,\(^{32}\) they

---

26. Id.
29. See Eliminating the Prohibition Against General Solicitation and General Advertising in Rule 506 and Rule 144A Offerings, Securities Act Release No. 33-9354, 77 Fed. Reg. 54,464, 54,464 (proposed Aug. 29, 2012) (“The proposed amendment to Rule 506 would provide that the prohibition against general solicitation and general advertising contained in Rule 502(c) of Regulation D would not apply to offers and sales of securities made pursuant to Rule 506, provided that all purchasers of the securities are accredited investors.”). As of February 6, 2013, no amendment to Rule 506 had been made in this regard.
30. 15 U.S.C. §§ 80a-1 to 80a-64.
31. Id. §§ 80b-1 to 80b-21.
protect existing clients not prospective ones. It is hard to believe, moreover, that widespread advertising will not lead to greater interest in hedge funds among those who, although “accredited,” do not, in their pursuit of juiced returns and the status of being “in” a hedge fund, fully understand what they are doing and the true nature of the risks they are taking, including the way in which their exposure to leverage risk is elevated through the funds’ use of debt and the great difficulty of exiting the investment. Also, particularly challenging for investors is understanding exactly how a fund is valuing its assets, an area in which practices vary widely.

Whether sponsor investor-verification techniques and enhanced industry “best practices” will successfully filter unqualified investors lured by enticing advertising remains to be seen. This is of great concern to state securities regulators. Sponsors themselves are concerned about this issue. The Hedge Fund Association has sought greater clarification from the SEC as to how funds should verify investor accreditation, inasmuch as the SEC’s proposed rule failed to offer a safe harbor. Such a rule also would position funds to defend against charges of wrongdoing in bringing in nonaccredited investors, by permitting them to argue that they followed SEC guidance. Such guidance, moreover, if given, might become a de facto industry standard, thereby impeding development of more rigorous practices.

The hedge fund anti-fraud rule of 2007, even coupled with the 2010 adviser registration requirement criticized by Luke, may do little to curb likely damage to investors in an overly zealous policy pursuit of “capital formation.” In this way, Luke’s pointed investor protection concern on the fund-adviser front joins a larger and longer debate over whether and how the federal

Jan. 19, 2013) (“[A]dvisers that are registered with the SEC must continue to comply with rules relating to advertising under the Investment Advisers Act of 1940 . . . .”) (on file with the Washington and Lee Law Review).

securities laws’ philosophy of full disclosure can achieve that goal while also balancing it with other policy goals with which it sometimes clashes.

**III. Systemic Risk**

If Luke is tough on investor protection, he is more generous on systemic risk. However, although he credits the systemic risk concerns of Dodd–Frank as worthwhile, Luke thinks adviser registration is not necessary to advance this goal. Instead, he argues, the Financial Stability Oversight Council (FSOC) can collect pertinent data directly from fund advisers through Form PF,34 without the intervention of the SEC. Even this more direct approach, however, invites certain concerns.

Given the ninety-day or longer lags in receiving data, information provided to the SEC and FSOC can be far more unhelpfully stale than fresher “real time” data. A lot of undetected systemic risk can develop in the meantime. This is somewhat meliorated by an obligation to “promptly” update specified information that becomes “materially inaccurate.”35 The key premise, of course, is that if adverse developments occur, managers will report them immediately, rather than later. A key risk, however, is that managers may tend not to regard adverse developments as material, perhaps out of simple hope and belief that conditions will change or from a simple desire not to alert the SEC. Moreover, one wonders whether staff at the SEC and the FSOC can efficiently handle a large influx of data and, upon distilling it, accurately identify disturbing aggregate trends. It is easier in hindsight to see potentially destabilizing patterns than it is while awash in data and searching for, but not wishing to overreact to, possible indicators of trouble. Only time will tell on this front.

But concerns about hedge funds’ role in systemic risk may mask a deeper clash of interests, i.e., that between the interests

---


35. SEC Implementing Release, supra note 14, at 114.
of fund investors and the larger public interest as reflected in systemic (social) risk concerns. Hedge funds command large pools of capital, deploy significant leverage, engage in rapid trading activity, and may profit from and magnify volatility. Whether this creates or exacerbates unacceptable systemic (social) risk is unclear but should be ascertained. What is clear is that funds need investors. Thus, if funds are the sort of financial vehicle that contributes to systemic instability, it is because the very financial wherewithal to do so is supplied by investors. The public policy problem, ironically, may not be to protect investors from fund advisers but, in effect, may be to protect the financial system from risks caused by inappropriate use of the very concentrated capital supplied by investors. Viewed this way, hedge funds may present less of an “adviser vs. investor” regulatory issue than a “fund vs. society” concern.

Here, it might be better to acknowledge outright that mandated disclosure for systemic risk concerns is just that and not cloak it in investor protection rhetoric. As has been noted about the federal Bureau of Corporations, established in 1903, its purpose similarly was aimed at “collecting industrial data and investigating corporate trade practices as a deterrent against illicit corporate activities.” The information was aimed at consumption outside the firm, not at those investing in it. And disclosure was required to deter activity that might enrich investors, not damage them. As with hedge fund regulation, then, that early 20th century disclosure law was designed to

36. See Edwards, supra note 9, at 189–91 (describing general characteristics of hedge funds).


39. See Joo, supra note 37, at 1582 (“Progressives wanted disclosure because they believed it would create accountability, not to shareholders, but to ‘the public.’”).

40. See id. at 1583 (“[T]he Bureau required a corporation’s ‘disclosure’ not to empower or protect that corporation’s shareholders, but to discourage the corporation from anticompetitive activity—activity that could have enriched its shareholders.”).
Why Register Hedge Fund Advisers

protect the public interest, not investors. This suggests a potentially deeper clash of policy goals in adviser regulation. What is good about hedge funds—for at least some, perhaps many, investors—may not be good for larger social interests.

Fortunately, we can sidestep this issue for now because what Dodd–Frank really does on the adviser regulation front is simply mandate data collection. The requirements, although law, are really in the nature of an extended study. The aim, paradoxically, is to learn more about the historically opaque hedge funds and their activities precisely to ascertain whether and how to regulate them. It may well be concluded that no further regulatory action is needed or even that a rollback is in order. Or, perhaps particular areas of concern will be targeted for action. But first we need to know how fund advisers themselves are responding, an area in which Professor Wulf Kaal is gathering and assessing useful data. Kaal’s early findings in a survey of about ninety hedge fund managers is that the industry appears to be only “modestly affected” by the new requirements. Approximately seventy-two percent of managers reported no expected strategic response to Dodd–Frank, with smaller funds more likely to be affected. Although earnings and profits were expected to decline due to higher compliance costs—with almost half of responding advisers outsourcing compliance work—seventy-five percent of managers did not expect investor returns to suffer. Over forty percent have changed their marketing materials, a figure that may rise yet again if existing restrictions on advertising are dropped, as recently proposed.


42 Id. (manuscript at 59).

43 Id. (manuscript at 56); see also Wulf A. Kaal, The Effect of the Dodd–Frank Act on the Hedge Fund Industry 41 (Feb. 2, 2013) (unpublished manuscript) (asserting that certain data suggest that the compliance and administrative costs created by the Dodd–Frank Act may disproportionately affect smaller hedge funds) (on file with author).

44 Kaal, supra note 41 (manuscript at 42).

45 Id. (manuscript at 53).

46 Id. (manuscript at 42).

47 See supra note 29 and accompanying text.
And from there, we can see what the SEC and FSOC—or Congress—does with the new trove of data, either on the investor protection or systemic risk fronts. Luke’s skepticism on the former may be borne out, thereby further spotlighting the latter. And this would then open an even more potentially disquieting debate about the social responsibilities and legal rights of those who facilitate rapid, large capital movements in a way that may destabilize modern markets. Dodd–Frank certainly takes no position on, but may set up, that far more difficult conversation.
Sheltering Psychiatric Patients from the *DeShaney* Storm: A Proposed Analysis for Determining Affirmative Duties to Voluntary Patients†

Claire Marie Hagan*

*Table of Contents*

I. Introduction ................................................................. 726

II. Substantive Due Process and Affirmative Duties .......... 731
   A. Substantive Due Process: A Charter of Negative Duties ........................................... 733
   B. Affirmative Duty Exceptions: *Youngberg* and *DeShaney*........................................... 734
      1. *Youngberg v. Romeo*................................................. 736
      2. *DeShaney v. Winnebago*................................. 737
   C. Affirmative Duties After *DeShaney* ......................... 742
      1. The Functional Custody Exception .................... 744
      2. The State-Create d Danger Exception ............... 746

III. The Landscape of Mental Health Law and Public Psychiatric Hospitalization .................................. 748
   A. The Current Law of Voluntary and Involuntary Admissions .......................................... 748
   B. Historical Context of Mental Health Law and Hospitalization ...................................... 752

† This Note received the 2012 Washington and Lee Law Council Law Review Award for outstanding student Note.

* Candidate for J.D., Washington and Lee University School of Law, May 2013; M.S.W., University of Michigan, December 2006; B.A., University of Notre Dame, May 2005. Many, many, many thanks to: my editors, Professor Timothy Jost and Ned Hillenbrand; to Professors Susan Stefan and J.D. King for commenting on this Note; to my friend and teacher, Fr. Tom Blantz, C.S.C., at Notre Dame; and to my former colleagues in the Department of Psychiatry at the Johns Hopkins Hospital.
IV. Circuit Split over Affirmative Duties to Voluntary Patients

A. A Status-Based Approach to Affirmative Duties
   1. Sixth Circuit: *Higgs v. Latham* .......... 758
   2. First Circuit: *Monahan v. Dorchester*.......... 761

B. A Fact-Based Approach to Affirmative Duties .......... 766
   1. Eleventh Circuit: *Spivey v. Elliot* .......... 767
   2. Eighth Circuit: *Kennedy v. Schafer* and
      *Shelton v. Arkansas Department of Human Services* .......... 768

V. Concerns about Overreliance on Voluntary/Involuntary Distinction .......... 775

A. Competency: Can Acutely Mentally Ill Patients
   Give Informed Consent? .......... 775

B. Coercion: How do External Forces Influence
   a Patient’s Voluntary Status? .......... 777

VI. Proposed Analysis for Determining Duties to a Voluntary Patient .......... 781

A. The Proposed Analysis .......... 781

B. Benefits of this Approach .......... 783
   1. The Proposed Analysis Appropriately
      Considers the Voluntary Distinction .......... 784
   2. The Proposed Analysis Retains Doctrinal
      Integrity .......... 786
   3. The Proposed Analysis Will Not Unduly
      Expand Affirmative Duties .......... 788

VII. Conclusions .......... 790

I. Introduction

The government owes us nothing. That is, the Constitution does not require the government to protect us or to provide services. Instead, the Constitution restricts the government from

acting. For example, citizens enjoy the sense of security provided by their local police force, but the police are not constitutionally obligated to respond to every emergency call. As a corollary to this “no-duty” rule, state actors generally cannot be liable for inaction. Plaintiffs who sue the state on an affirmative-duty based claim face a formidable challenge, and yet plaintiffs frequently assert the claims.

The no-duty rule is not without exception. In Youngberg v. Romeo, the Supreme Court ruled that states owe involuntarily committed state-hospital patients affirmative duties of care, protection, and rehabilitation. In the 1970s and 1980s, the Court also recognized affirmative duties for prisoners, for pre-trial detainees, and for arrestees. But in DeShaney v. Winnebago County Department of Social Services, the Supreme Court (explaining that the negative-liberties model of the Constitution provides no basis for substantive benefits).

2. Id.


4. See DeShaney v. Winnebago Cnty. Dep't of Soc. Servs., 489 U.S. 189, 196–97 (1989) (“[T]he State cannot be held liable under the [Due Process] Clause for injuries that could have been averted had it chosen to provide them.”).

5. See 1 Martin A. Schwartz, Section 1983 Litigation: Claims and Defenses § 3.09[D] (4th ed. 2012) (“[Section] 1983 claimants continue to file large numbers of due process duty to protect claims.”).


7. This Note uses the term “state” to include state and local entities and their agents.


9. See Estelle v. Gamble, 429 U.S. 97, 103 (1976) (establishing that, under the Eighth Amendment, states owe affirmative duties to provide medical care to prisoners).

10. See Bell v. Wolfish, 441 U.S. 520, 545 (1979) (establishing states’ affirmative duty to provide safe conditions to pretrial detainees).


12. See DeShaney v. Winnebago Cnty. Dep't of Soc. Servs., 489 U.S. 189, 191 (1989) (holding that the state’s failure to protect an individual from private harm did not violate the Due Process Clause).
sharply cabined liability premised on affirmative-duty theories and fortified the general no-duty rule. Nonetheless, *DeShaney* reiterated *Youngberg*'s rule that involuntary patients enjoy affirmative rights. And circuit courts generally interpret *DeShaney* as creating a rule that affirmative duties arise in three discrete scenarios: (1) formal custody; (2) functional custody; and (3) state-created danger.

State hospitals present a unique context for analyzing affirmative duties. Unlike criminal custodial settings, state laws set forth procedures for people to either voluntarily enter state hospitals or to be involuntarily committed. The “voluntary” and “involuntary” labels seem, at first blush, to describe a particular patient’s relationship to the state: Involuntary patients are held against their will, and voluntary patients fully consent to hospitalization. The “voluntary” label often provides a faulty description, however, because state-hospital patients may lack competency to give informed consent, and coercive forces may taint their consent.

This Note focuses on affirmative duties to voluntary state-hospital patients. If the state involuntarily commits a patient, formally taking custody, then the state has an affirmative duty to

13. See Peter Irons, Brennan vs. Rehnquist 107 (1994) (“The Court went out of its way to decide the *DeShaney* case. . . . Rehnquist’s majority opinion in the case suggests that he wanted to send a message to federal judges.”).
14. See id. at 199 (discussing established affirmative-duty contexts).
15. See infra Part II.C.1 (discussing the functional custody exception); infra Part II.C.2 (discussing the state-created danger exception). For clarity, in this Note, “formal custody” refers to incarceration, involuntary commitment, pre-trial detention, and arrest. “Functional custody” refers to analogous situations in which the state, through affirmative acts, creates a custodial relationship. “State-created danger” refers to situations in which the state acts either to create a danger or to render an individual more vulnerable to a danger.
17. See infra Part V (raising concerns with the voluntary distinction).
18. This Note uses the terms “mental health patient” and “state-hospital patient” to refer to individuals receiving inpatient treatment in state-operated psychiatric facilities, either voluntarily or involuntarily. Patients in private facilities are outside this Note’s scope. Private facilities and employees generally are not state actors, and therefore are not subject to § 1983. See Civil Action for Deprivation of Rights, 42 U.S.C. § 1983 (2006) (imposing liability against state actors).
protect and care for that person. But what about voluntary patients? Circuits are split on this question. Does the state owe different obligations to patients based simply on a formal status—voluntary or involuntary? What if state law allows a facility to hold a voluntary patient for seventy-two hours after the patient decides he wants to leave? What if a voluntary patient is so psychiatrically ill that state law forbids discharge? What if the hospital knows that a patient is likely to harm another patient?

Courts disagree on how to answer these questions. Circuits generally take one of two approaches: (1) a strict status-based test, using an individual’s formal “voluntary” or “involuntary” status to determine if the state owes affirmative duties, or (2) a fact-intensive inquiry of whether the individual was truly a voluntary patient when the harm occurred. Both analyses, however, emphasize the individual’s commitment status—voluntary or involuntary—as the chief element.

Should courts employ the voluntary/involuntary distinction as the engine driving the analytic train? This Note argues: No. First, the distinction may amount to an artificial signifier. Affirmative duty analysis seeks to understand the relationship

---


20. See infra Part IV (reviewing the circuit split).

21. See infra Part IV.A.2 (discussing the First Circuit’s decision in Monahan v. Dorchester Counseling Ctr., 961 F.2d 987 (1st Cir. 1992)).

22. See infra Part IV.B.3 (discussing the Third Circuit’s decision in Torisky v. Schweiker, 446 F.3d 438 (3d Cir. 2006)).

23. See infra Part IV.B.2 (discussing the Eighth Circuit’s decisions in Kennedy v. Schafer, 71 F.3d 292 (8th Cir. 1996), and Shelton v. Ark. Dep’t Human Servs., 677 F.3d 837 (8th Cir. 2012)).

24. See infra Part IV.A.3 (discussing the Fifth Circuit’s decision in Walton v. Alexander, 44 F.3d 1297 (5th Cir. 1995) (en banc)).

25. See infra Part IV (analyzing the different approaches circuit courts take in analyzing affirmative duties to state-hospital patients).

26. See infra Part IV.A (reviewing cases that adopt a status-based analysis).

27. See infra Part IV.B (discussing cases that adopt a fact-based analysis).

and course of dealings between the petitioner and the state.\textsuperscript{29} While the labels “voluntary” and “involuntary” superficially signal whether a patient’s admission is a product of consent or of confinement, these labels can be misleading.\textsuperscript{30} Second, the voluntary distinction loses sight of alternative bases for affirmative rights. When courts focus on whether a state-hospital patient is voluntary, either by status or de facto, the court may overlook or neglect to explore fully the functional custody and state-created danger exceptions.\textsuperscript{31}

This Note puts forth an alternative analytic structure for deciding whether state-hospital patients can establish that the state owed affirmative \textit{Youngberg} duties.\textsuperscript{32} Beginning with the presumption that the state owes no affirmative duty,\textsuperscript{33} the court next considers each \textit{DeShaney} exception in turn. First, if an individual is committed involuntarily, then \textit{Youngberg} duties exist based on formal custody. Second, if an individual is a voluntary patient, then \textit{Youngberg} duties exist if the state exercises functional custody by restricting the person’s liberty. Third, if an individual is a voluntary patient, then \textit{Youngberg} duties exist if the state creates or increases a danger threatening the individual.

Part I of this Note introduces the issues to be addressed. Part II reviews substantive due process and its general role in restricting government action. Part II then analyzes \textit{Youngberg} and \textit{DeShaney} before briefly reviewing the current state of affirmative duty law, including both the functional custody and the state-created danger doctrines. Part III focuses on mental health law, providing a background on the laws governing

\textsuperscript{29} See Daniels v. Williams, 474 U.S. 327, 332 (1986) (explaining that the Fourteenth Amendment governs the relationship between individuals and state governments).

\textsuperscript{30} See \textit{infra} Part V.A (examining competency and coercion in voluntary admissions).

\textsuperscript{31} See \textit{infra} notes 266, 339–42 and accompanying text (discussing the circuits’ focus on voluntariness, rather than on functional custody or state-created danger); \textit{infra} notes 214–20 and accompanying text (discussing Judge Suhreinrich’s concurrence in \textit{Higgs}, which argued that the majority should have considered functional custody and state-created danger).

\textsuperscript{32} \textit{Infra} Part VI.

\textsuperscript{33} See \textit{infra} Part VI.A (containing this Note’s recommended analysis).
voluntary and involuntary admission and on the state hospital context in light of recent historical trends. Part IV reviews the circuit split over whether voluntary state-hospital patients are owed affirmative duties. Part V raises concerns about circuits’ focus on the voluntary/involuntary distinction. This Note concludes in Part VI by proposing an analysis for approaching affirmative duties in the state hospital context, arguing that this approach gives appropriate weight to voluntary/involuntary issues, retains doctrinal integrity, and will not overburden states.

II. Substantive Due Process and Affirmative Duties

The Fourteenth Amendment’s Due Process Clause protects individuals from unreasonable or oppressive government action.34 State actors violate an individual’s substantive due process rights if they act so unreasonably or so oppressively that no amount of procedural protections could justify their action.35 The scope of these substantive rights, however, is quite limited.36 Constitutional law does not remedy every state actor’s wrong.37

Due process of law originated in English law as a legal maxim meant to restrain the sovereign.38 At common law, the substantive prong protected property and contract rights.39 American jurisprudence continued recognizing due process as a “significant constitutional limitation[]” on executive and

34. See Daniels, 474 U.S. at 331 (explaining that the Due Process Clause provides procedural and substantive protections against government action depriving individuals of life, liberty, or property in an arbitrary or oppressive manner). This Note focuses only on the substantive prong of due process.

35. Id.

36. Id.

37. See id. at 332 (“[The Constitution] does not purport to supplant traditional tort law in laying down rules of conduct to regulate liability for injuries that attend living together in society.”).


39. See id. at 98–102 (summarizing the common law roots of due process and the doctrine’s evolution in American jurisprudence).
legislative power. Over time, the doctrine evolved, with substantive due process now protecting individual interests “relating to marriage, family, procreation, and the right to bodily integrity.” The Court has long held that individuals possess a fundamental right to personal autonomy, including in a medical setting.

Patients who suffer some injury while in a state hospital may claim that the state violated their substantive due process rights. Due process rights may be vindicated via claims brought under 42 U.S.C. § 1983. And in the last few decades, § 1983 has “emerged as a potent weapon for state hospital patients.” These constitutional claims are popular for mental health advocates seeking “system-wide changes,” because § 1983 actions offer injunctive relief and legal fees in addition to compensatory damages.

---

40. Id. at 9.
42. See Union Pac. Ry. v. Botsford, 141 U.S. 250, 251 (1891) (“No right is held more sacred, or is more carefully guarded . . . than the right of every individual to the possession and control of his own person, free from all restraint or interference of others.”).
43. See Schloendorff v. Soc’y of N.Y. Hosp., 105 N.E. 92, 93 (N.Y. 1914) (“Every human being of adult years and sound mind has a right to determine what shall be done with his own body.”). The Supreme Court has adopted Judge Cardozo’s articulation in Schloendorff of the individual’s right to control his body from unwanted medical intervention. Cruzan v. Mo. Dep’t of Health, 497 U.S. 261, 269 (1990).
44. See U.S. CONST. amend. XIV, § 1 (“[N]or shall any State deprive any person of life, liberty, or property, without due process of law.”).
45. See Civil Action for Deprivation of Rights, 42 U.S.C. § 1983 (2012) Every person who, under color of any statute, ordinance, regulation, custom, or usage . . . subjects, or causes to be subjected, any citizen of the United States . . . to the deprivation of any rights . . . secured by the Constitution and laws, shall be liable to the party injured.
47. See id. at 138 (explaining why the remedies available pursuant to a § 1983 claim may be more attractive than state law tort remedies when legal advocates hope to effect broad change in state hospitals).
This Part first reviews the doctrine of substantive due process. Generally, this doctrine restricts, rather than mandates, government action.\(^{48}\) Then, this Part reviews the two Supreme Court cases governing when a state owes affirmative duties to state-hospital patients: *Youngberg* and *DeShaney*.\(^{49}\) Finally, this Part summarizes the affirmative-duty rules that circuit courts have developed after *DeShaney*.\(^{50}\)

### A. Substantive Due Process: A Charter of Negative Duties

Substantive due process foists negative duties on the government, curtailing its power.\(^{51}\) Judge Posner, writing for the Seventh Circuit, articulated the Fourteenth Amendment’s scope in a case involving a dangerous, mentally ill person who—subsequent to being released by the state—murdered the plaintiff’s decedent:

There is a constitutional right not to be murdered by a state officer, for the state violates the Fourteenth Amendment when its officer, acting under color of state law, deprives a person of life without due process of law. But there is no constitutional right to be protected by the state against being murdered by criminals or madmen. It is monstrous if the state fails to protect its residents against such predators but it does not violate the due process clause of the Fourteenth Amendment or, we suppose, any other provision of the Constitution. The Constitution is a charter of negative liberties; it tells the state to let people alone; it does not require the federal government or the state to provide services, even so elementary a service as maintaining law and order.\(^{52}\)

The Supreme Court consistently affirms this characterization of the Fourteenth Amendment as “a charter of negative liberties.”\(^{53}\)

---

48. *See infra* Part II.A (explaining the general no-duty rule).
49. *See infra* Part II.B (examining *Youngberg* and *DeShaney*).
50. *See infra* Part II.C (discussing post-DeShaney developments).
B. Affirmative Duty Exceptions: Youngberg and DeShaney

The general rule that the Due Process Clause imposes negative duties is not absolute. In discrete contexts, the Clause imposes affirmative duties, calling on the state to proactively serve or protect.54

Circuit courts have labored to define the scope of affirmative duties for several decades. In Martinez v. California,55 Justice Stevens suggested in dicta that a state might owe an affirmative duty of protection if the state (1) becomes aware of a special danger threatening a specific individual and (2) indicates a willingness to protect that person.56 Many circuits interpreted this language to create a “special-relationship” doctrine, whereby the government undertakes an obligation to protect persons with whom it shares a special relationship.57 These circuits began

("Neither the text nor the history of the Due Process Clause supports petitioner's claim that the governmental employer's duty to provide its employees with a safe working environment is a substantive component of the Due Process Clause."); Davidson v. Cannon, 474 U.S. 344, 348 (1986) ("The guarantee of due process has never been understood to mean that the State must guarantee due care on the part of its officials.").

54. See DeShaney, 489 U.S. at 198 ("It is true that in certain limited circumstances the Constitution imposes upon the State affirmative duties of care and protection with respect to particular individuals.").

55. See Martinez v. California, 444 U.S. 277, 284 (1980) (holding that the state was not liable for a murder committed by a parolee five months after the state released him).

56. See id. at 285

[The parole board was not aware that appellants' decedent, as distinguished from the public at large, faced any special danger. We need not and do not decide that a parole officer could never be deemed to “deprive” someone of life by action taken in connection with the release of a prisoner on parole.

57. See, e.g., Balistreri v. Pacifica Police Dep't, 855 F.2d 1421, 1425–26 (9th Cir. 1988) (ruling that the plaintiff's allegations that the state owed an affirmative duty of police protection after the state issued a restraining order and received notice of plaintiff's danger was sufficient to state a claim under the special-relationship doctrine); Estate of Bailey v. Cnty. of York, 768 F.2d 503, 510–11 (3d Cir. 1985) (vacating and remanding the district court's dismissal of a § 1983 claim and ruling that plaintiff might be able to prove the state violated an affirmative duty arising under the special-relationship doctrine); Jones v. Phyfer, 761 F.2d 642, 644–45 (11th Cir. 1985) ("What is required in a 42 U.S.C. § 1983 action is the establishment of a special relationship between the victim and the criminal or between the victim and the state, or some showing that the
finding affirmative duties grounded in substantive due process following *Martinez*.

In the context of mental health patients, two Supreme Court cases shape the legal doctrine. First, in *Youngberg*, the Court held that the government takes on affirmative duties when it involuntarily confines a person to a mental health facility. Second, in *DeShaney*, the Court hemmed in the scope of affirmative duties.

---

58. See *Youngberg v. Romeo*, 457 U.S. 307, 324 (1982) (ruled that an individual who has been involuntarily committed to a psychiatric hospital “enjoys constitutionally protected interests in conditions of reasonable care and safety, reasonably nonrestrictive confinement conditions, and such training as may be required by these interests”).

59. *Id.*

60. See *DeShaney v. Winnebago Cnty. Dep’t of Soc. Servs.*, 489 U.S. 189, 191 (1989) (holding that the state’s failure to protect an individual from private harm did not violate the Due Process Clause).

61. See *id.* at 196 (“[O]ur cases have recognized that the Due Process Clauses generally confer no affirmative right to governmental aid, even where such aid may be necessary to secure life, liberty, or property interests of which the government itself may not deprive the individual.”).
I. Youngberg v. Romeo

In 1982, the Supreme Court first considered whether an involuntarily confined mental health patient possesses liberty interests under substantive due process in Youngberg v. Romeo. Youngberg involved a “profoundly retarded” thirty-three-year-old man who was civilly committed to Pennhurst, a state psychiatric facility, because he posed an imminent danger to himself and others. At Pennhurst, hospital staff proposed a treatment plan designed to reduce Mr. Romeo’s aggressive and violent behaviors, but never implemented it. Mr. Romeo repeatedly suffered injuries, some of which were self-inflicted, during the hospitalization. Mr. Romeo’s mother filed suit against the hospital’s directors and supervisors, alleging in part that the defendants violated Mr. Romeo’s Fourteenth Amendment Due Process rights.

The Supreme Court ruled that the State had violated Mr. Romeo’s substantive due process rights by not fulfilling its affirmative obligation to protect and care for him. Justice Powell, writing for the majority, analogized involuntary state hospitalization to incarceration. Just as the state owes affirmative duties to prisoners in public prisons, it also owes duties to involuntary patients in public hospitals. The hospital must provide food, shelter, clothing, and medical care to all

63. Id. at 309–10.
64. Id. at 310–11.
65. See id. at 310 (stating that petitioner’s complaint alleged that Mr. Romeo was injured “on at least sixty-three occasions”).
66. Id.
67. See id. at 324 (“Respondent thus enjoys constitutionally protected interests in conditions of reasonable care and safety, reasonably nonrestrictive confinement conditions, and such training as may be required by these interests.”).
68. Id. at 315.
69. See id. at 315–16 (“If it is cruel and unusual punishment to hold convicted criminals in unsafe conditions, it must be unconstitutional to confine the involuntarily committed—who may not be punished at all—in unsafe conditions.”).
patients. But Mr. Romeo’s rights included more than these basic services. The Court ruled that states owe involuntarily committed individuals certain rights (Youngberg rights): (1) the right to reasonable care and safety; (2) the right to reasonably nonrestrictive conditions; and (3) the right to any training or rehabilitation associated with these interests.

The decision does not address whether Youngberg rights extend to voluntary patients. Some lower courts interpreted Justice Powell’s opinion as applying to all mental health patients, irrespective of the method of admission—voluntary or involuntary. Additionally, some circuit courts continued finding affirmative duties based on Martinez and the special-relationship doctrine.

In 1989, the Supreme Court revisited its affirmative duty rulings in DeShaney, a case set outside of a psychiatric institution, but nonetheless broadly addressing the scope and nature of substantive due process protections.

2. DeShaney v. Winnebago County Department of Social Services

In DeShaney, the Supreme Court considered whether the Substantive Due Process Clause requires states to protect

---

70. See id. at 315, 324 (noting in dicta that the State conceded that it owes these duties to all state-hospital patients).

71. See id. at 314–25 (finding each of these liberty interests). The Court further determined that a “professional judgment” standard should apply to determine whether the state violated its duties to protect or care for a patient. See id. at 321–22 (reasoning that deferring to the hospital clinicians’ judgment is appropriate because it allows hospitals to concentrate on treating patients according to their individual needs without imposing an onerous blanket rule).

72. See Savidge v. Fincannon, 836 F.2d 898, 907 n.44 (5th Cir. 1988) (rejecting the argument that Youngberg applies only when an individual has been confined through formal proceedings); Ass’n for Retarded Citizens of N.D. v. Olson, 561 F. Supp. 473, 485 (D.N.D. 1982) (“[I]f the plaintiffs had voluntarily consented to their admission to the Grafton state school, it would not follow that all their rights to liberty under the due process clause were waived.”), aff’d and remanded, 713 F.2d 1384 (8th Cir. 1983).

73. See supra note 57 (listing pre-DeShaney circuit cases on special relationships).

74. See infra Part II.B.2 (examining the DeShaney ruling).
individuals from private—rather than state-inflicted—harm. DeShaney is not about mental health patients. Joshua DeShaney was a four-year-old boy who became involved with Wisconsin’s Department of Social Services (DSS) in January 1983 because of his father’s abuse. During 1983, Joshua was hospitalized three times for injuries; hospital staff suspected child abuse each time and contacted DSS. DSS placed Joshua under the hospital’s temporary custody for three days after the first report, but returned Joshua to his father’s home three days later, and began sending a social worker out to visit the home monthly. Over a period of a little more than a year, Joshua kept showing up in the emergency room, injured; the hospital kept reporting the injuries to DSS; and DSS kept sending the social worker—and nothing more. The social worker documented “suspicious injuries on Joshua’s head” and “continuing suspicions” that Joshua was being physically abused. Despite these reports, DSS took no further action. In March 1984, Joshua went into a life-threatening coma. Emergency surgery revealed brain damage resulting from repeated head injuries over a long period. Now “profoundly retarded,” he would likely spend the rest of his life in an institution. Joshua’s mother sued DSS on his behalf, alleging that DSS knew of or should have reasonably known about the abuse, and so the State violated Joshua’s due process rights by failing to protect him.

75. See DeShaney v. Winnebago Cnty. Dep’t of Soc. Servs., 489 U.S. 189, 194 (1989) (characterizing the issue presented as “when, if ever, the failure of a state or local governmental entity or its agents to provide an individual with adequate protective services constitutes a violation of the individual’s due process rights”).
76. Id. at 192–93.
77. Id.
78. Id.
79. Id.
80. Id.
81. Id.
82. Id. at 193.
83. Id.
84. Id.
85. Id. at 193–94.
The Supreme Court disagreed. Chief Justice Rehnquist, writing for the majority, analyzed the scope of substantive due process and the nature of its protections. The majority began its analysis with the principle that the Due Process Clause functions as a limitation on the government. The Clause restricts the government—it does not empower the government by requiring states to protect society or provide services. From this premise, the majority reasoned that a state cannot be liable for private violence because states have no duty to protect against this violence. Youndberg, according to the majority, “stand[s] for the proposition that when [a] State takes a person into its custody and holds him there against his will, the Constitution imposes upon it a corresponding duty to assume some responsibility for his safety and general well-being.”

The Court also expressly rejected the argument that affirmative duties might derive from a special relationship. The special-relationship doctrine implied that affirmative duties arise from the state’s knowledge of danger and its indicated willingness to help. Chief Justice Rehnquist pointed out that the doctrine’s analysis misconstrues the nature and purpose of affirmative duties, which are meant to protect individuals from harms caused

86. See id. at 194–96 (reviewing the Clause’s text, history, and supporting doctrine).
87. See id. at 196 (“The Due Process Clauses generally confer no affirmative right to governmental aid.”). Some scholars criticize Chief Justice Rehnquist’s interpretation of the Due Process Clause, arguing that it “fundamentally distorts the meaning . . . by abstracting the language from its historical context.” Steven J. Heyman, The First Duty of Government: Protection, Liberty and the Fourteenth Amendment, 41 DUKE L.J. 507, 561 (2001). In his Article, Professor Heyman analyzes the Fourteenth Amendment’s origins from English law, as well as the congressional debates and framers’ intentions in adopting the Amendment. Id. He concludes that the view that the Due Process Clause provides only negative duties is “indefensible, and that the history of the Fourteenth Amendment in no way forecloses recognition of a constitutional right to protection.” Id. at 512.
89. Id. at 199–200. Justice Brennan criticized this characterization of Youndberg in his dissent. Infra notes 103–04 and accompanying text.
90. See id. at 197–98 (discussing the special-relationship argument advanced here by petitioners and accepted in several circuit courts after Martinez, but ultimately rejecting the argument).
91. Id. at 197 n.4.
by affirmative state acts that render people unable to act on their own behalf. Without affirmative state action, duties will not arise.

Justice Brennan, joined by Justices Marshall and Blackmun, dissented and criticized the majority’s analytic posture. Justice Brennan did not disagree that, as a general principle, the Due Process Clause imposes no obligation on the government to provide services. But the majority’s “initial fixation” with this principle led it astray. Approaching affirmative-duty claims with “suspicion,” the majority framed the case as one about inaction: the state’s failure to protect Joshua. Any actions that the state did take became tangentially important. To the majority, the question was whether a state should be liable for inaction. The majority viewed the facts through a skewed lens, focusing in on what did not happen—the ways the state did not exercise custody or increase Joshua’s vulnerability, “foreshadow[ed]—perhaps even preordain[ed]” the conclusion. Justice Brennan argued that a proper analysis should begin differently and ask what actions the state did take.

Justice Brennan also took issue with the majority’s characterization of Youngberg—that the state owed affirmative duties because it confined Mr. Romeo’s liberty through civil commitment. Justice Brennan asserted that Mr. Romeo’s constitutional right did not spring from the government’s prior

92. Id. at 200.
93. Id.
94. Id. at 203–12 (Brennan, J., dissenting).
95. See id. at 203 (stating that the Court’s precedent supports this principle).
96. See id. at 205 (stating that the majority’s preoccupation with the negative rights principle leaves the majority “unable to appreciate” Court precedent related to when affirmative rights arise).
97. Id. at 204.
98. Id.
99. See id. (arguing that the majority’s baseline perspective of no positive rights brings about “a concomitant suspicion of any claim that seems to depend on such rights”).
100. Id.
101. Id. at 205.
102. Id. at 206.
act of committing him. The affirmative duty arose because the state “separated him from other sources of aid,” obliging the state to provide substitute aid. Mr. Romeo was unable to care for himself because he was mentally retarded, not because he was confined.

In a separate dissent, Justice Blackmun highlighted the case’s tragic facts, “displaying emotion rarely seen in Supreme Court opinions.” He argued that the majority’s “sterile formalism” in distinguishing between action and inaction was inappropriate and marked “a sad commentary on American life, and constitutional principles.” DSS knew that Joshua was in danger, and DSS chose “inaction.” The majority, he asserted, misread “the broad and stirring Clauses of the Fourteenth Amendment,” which were designed to prevent such a rigidly formalistic interpretation of the law. By adopting this formalistic distinction between action and inaction, Justice Blackmun alleged, the majority opinion recalled “the antebellum judges who denied relief to fugitive slaves.”

Despite passionate criticisms in the dissents and from legal scholars, DeShaney remains good law. The Substantive Due

103. See id. (“This restatement of Youngberg’s holding should come as a surprise.”).
104. Id.
105. Id.
108. Id. at 212.
109. Id.
110. Id.
111. See, e.g., CHEMERINSKY, supra note 106, at 182 (arguing that DeShaney represents a “particularly troubling” example of how a “conservative attack” on the Constitution in recent decades has caused the Supreme Court to limit constitutional liberties and protections); Laurence H. Tribe, The Curvature of Constitutional Space: What Lawyers Can Learn from Modern Physics, 103 HARV. L. REV. 1, 10 (1989) (criticizing the majority’s analysis because its “stilted, pre-modern paradigm” fails to understand how Wisconsin “may itself have played a major role in shaping the world it observes”).
112. See infra notes 118–19 (stating that DeShaney remains the Supreme Court’s preeminent ruling on affirmative duties under substantive due process).
Process Clause does not generally require states to affirmatively protect individuals, and so a state is not liable if an individual becomes injured after the state fails to protect him. DeShaney's tone strongly suggests that it hoped to firmly limit the scope of affirmative duties. The Chief Justice, however, identified several contexts that give rise to affirmative duties on the state: (1) in previously enumerated contexts involving formal custody—incarceration, civil commitment, pre-trial detention, and police custody; (2) when a state takes an individual into functional custody through an affirmative action, restricting that person's ability to protect himself; and (3) when a state creates a danger or causes an individual to be more vulnerable to a danger.

C. Affirmative Duties After DeShaney

DeShaney remains the Supreme Court's authoritative ruling on affirmative duties under the Substantive Due Process Clause. The Court has subsequently affirmed DeShaney's

114. See id. at 202–03 (remarking that "natural sympathy" pushes judges and lawyers to seek compensation for Joshua, but cautioning that the state actors "should not have [liability] thrust upon them by this Court's expansion of the Due Process Clause").
115. Id. at 198–99. DeShaney noted that some circuit court precedent also includes foster children in this list, but the Court declined to rule on this issue. Id. at 201 n.9. Most circuits addressing this issue since DeShaney have ruled that states owe affirmative duties to children placed in foster homes. See Martin A. Schwartz & Kathryn R. Urbonya, Section 1983 Litigation 40–41 & n.213 (Fed. Judicial Ctr., 2d ed. 2008) (collecting relevant circuit cases).
116. See DeShaney, 489 U.S. at 200 (“[I]t is the State's affirmative act of restraining the individual's freedom to act on his own behalf—through incarceration, institutionalization, or other similar restraint of personal liberty—which is the 'deprivation of liberty' triggering the protections of the Due Process Clause.” (emphasis added)).
117. See id. at 201 (“While the State may have been aware of the dangers that Joshua faced in the free world, it played no part in their creation, nor did it do anything to render him any more vulnerable to them.” (emphasis added)).
118. See Erwin Chemerinsky, Government Duty to Protect: Post-DeShaney Developments, 19 Touro L. Rev. 679, 683 (2003) (“DeShaney is the touchstone for all subsequent discussions about the affirmative duty to provide protection under due process.”).
holding, expressing its “reluctance to expand the concept of substantive due process because guideposts for responsible decisionmaking in this unchartered area are scarce and open-ended.” Expanding due process through judicial decisions creates the risk that the judiciary’s policy preferences will define the scope of protected liberties.

Despite the Court’s self-imposed restraint, it has not scaled back established affirmative duties. Since DeShaney, the Supreme Court has reiterated that affirmative duties exist in the enumerated and recognized contexts of formal confinement—prisons, pre-trial detainees, arrestees, and “persons in mental institutions.” The functional custody and state-created danger exceptions identified in DeShaney give plaintiffs alternative avenues for holding the government constitutionally liable for a private harm.

In the years since DeShaney came down, circuit courts have fashioned different variations on the functional custody and state-created danger exceptions.

119. See Collins v. City of Harker Heights, 503 U.S. 115, 126–27 (1992) (ruling that the Due Process Clause does not impose an affirmative duty on municipalities to provide minimal safety levels for employees); Town of Castle Rock v. Gonzales, 545 U.S. 748, 773 (2005) (Stevens, J., dissenting) (calling it “perfectly clear” that the Constitution imposes no general obligation on the state to provide police protection).

120. Collins, 503 U.S. at 125; see also Washington v. Glucksberg, 521 U.S. 702, 720 (1997) (asserting the importance of judicial restraint, because “[b]y extending constitutional protection to an asserted right or liberty interest, we, to a great extent, place the matter outside the arena of public debate and legislative action”).

121. See Moore v. City of E. Cleveland, 431 U.S. 494, 502 (1989) (plurality opinion) (calling substantive due process a “treacherous field for this Court” and arguing that “history counsels caution and restraint” in identifying new substantive liberty interests).

122. Collins, 503 U.S. at 127.

123. See Erwin Chemerinsky, The State-Created Danger Doctrine, 23 Touro L. Rev. 1, 1 (2007) (reviewing the theories available for establishing that the government owes an individual an affirmative duty after DeShaney).
1. The Functional Custody Exception

_DeShaney_ suggests that a state acquires an affirmative duty to care for and protect an individual when the state acts to restrain that person’s liberty.\(^{124}\) Most circuits have subsequently interpreted this language in _DeShaney_ as establishing affirmative duties when a state takes “functional custody”\(^{125}\) over an individual.\(^{126}\)

Circuits generally apply the functional custody exception and find an affirmative duty to protect when the state “affirmatively places the individual in a position of danger [that] the individual would not have otherwise faced.”\(^{127}\) Some circuits explicitly require an element of involuntariness: the state must have taken custody over the individual against his will.\(^{128}\)

---

\(^{124}\) See _DeShaney_ v. Winnebago Cnty. Dep’t of Soc. Servs., 489 U.S. 189, 200 (1989) (“[I]t is the State’s affirmative act of restraining the individual’s freedom to act on his own behalf—through incarceration, institutionalization, or other similar restraint of personal liberty—which is the ‘deprivation of liberty’ triggering the [duty].” (emphasis added)).

\(^{125}\) Many courts and scholars refer to the situation in which a state restricts an individual’s liberty through an affirmative act as a “special relationship.” See, e.g., Patel v. Kent Sch. Dist., 648 F.3d 965, 971 (9th Cir. 2011) (stating that _DeShaney_ creates a “special relationship exception” when a state restricts a person’s liberty such that the person cannot protect himself). To avoid confusion with the special-relationship doctrine that arose under _Martinez_—which _DeShaney_ explicitly rejects—this Note uses the term “functional custody” to denote situations in which a state’s action restricts a person’s liberty and thereby creates a constitutional obligation for the state to protect that person. See _supra_ notes 57, 90–92 and accompanying text (reviewing the special-relationship exception after _Martinez_ and _DeShaney’s_ subsequent rejection of this doctrine).

\(^{126}\) See SCHWARTZ & URBONYA, _supra_ note 115, at 40 (discussing functional custody).

\(^{127}\) Wallace v. Adkins, 115 F.3d 427, 429 (7th Cir. 1997). See also Armigo v. Wagon Mound Pub. Schs., 159 F.3d 1253, 1261 (10th Cir. 1998) (“[L]iability may attach to a state actor for the violence of a third party if the state restrained the plaintiff’s personal liberty and that restraint hindered the plaintiff’s freedom to act to protect himself.”).

\(^{128}\) See Randolph v. Cervantes, 130 F.3d 727, 730 (5th Cir. 1997) (stating that functional custody requires that the state’s action be “involuntary or against his will” (internal citations omitted)). See also Petition for Writ of Certiorari at 7–8, Kovacic v. Villarreal, No. 10-1235, 2011 WL 1393814, at *7–8 (Mar. 8, 2011) (requesting clarification on the ”swirling, murky waters” of functional custody doctrine), _cert. denied_, 131 S. Ct. 2985 (2011).
a state exercises may also affect this analysis, as circuits sometimes hinge functional custody on whether the individual “depend[ed] completely on the state to satisfy [his] basic human needs” and the state’s action prevented other sources of care.129

Individuals have asserted claims based on functional custody in various circumstances and found varying success. Most circuits recognize the state’s act of placing a child in foster care to be sufficient for establishing affirmative duties.130 By contrast, state laws mandating school attendance usually fail to establish functional custody—meaning that public schools do not owe affirmative duties to students—because the attendance laws do not sufficiently restrain the child.131


130. See Doe ex rel. Johnson v. S.C. Dep’t of Soc. Servs., 597 F.3d 163, 175 (4th Cir. 2010) (“We now hold that when a state involuntarily removes a child from her home, thereby taking the child into its custody and care, the state has taken an affirmative act to restrain the child’s liberty, triggering the [affirmative] protections of the Due Process Clause.”); Nicini v. Morra, 212 F.3d 798, 808 (3d Cir. 2000) (en banc) (“[W]hen [a] state places a child in state-regulated foster care, the state has entered into a special relationship with that child which imposes upon it certain affirmative duties.”); Norfleet ex rel. Norfleet v. Ark. Dep’t of Human Servs., 989 F.2d 289, 293 (8th Cir. 1993) (“[A] special custodial relationship . . . was created by the state when it took [a child] from his caregiver and placed him in foster care.”); Yvonne L. ex rel. Lewis v. N.M. Dep’t of Human Servs., 959 F.2d 883, 892–93 (10th Cir. 1992) (finding a clearly established right for foster children to be placed in safe conditions); K.H. ex rel. Murphy v. Morgan, 914 F.2d 846, 853 (7th Cir. 1990) (determining that states owe affirmative duties to foster children, reasoning that “[o]nce the state assumes custody of a person, it owes him a rudimentary duty of safekeeping no matter how perilous his circumstances when he was free”); Meador v. Cabinet for Human Res., 902 F.2d 474, 476 (6th Cir. 1990) (“[D]ue process extends the right to be free from the infliction of unnecessary harm to children in state-regulated foster homes.”).

131. See, e.g., Doe ex rel. Magee v. Covington Cnty. Sch. Dist. ex rel. Keys, 675 F.3d 849, 857 (5th Cir. 2012) (en banc) (“We reaffirm, then, decades of binding precedent: a public school does not have a DeShaney special relationship with its students requiring the school to ensure the students’ safety from private actors.”); Hasenfus v. LaJeunesses, 175 F.3d 68, 73–74 (1st Cir. 1999) (concluding that compulsory attendance laws did not create an obligation for a school to protect a student from suicide); Maldonado, 975 F.2d at 732–33 (“[T]he amount of freedom on the part of the student and the degree of parental involvement and control necessarily dictate that the state does not become the primary caretaker simply by mandating compulsory school attendance.”). But while compulsory
2. The State-Created Danger Exception

Most circuits also interpret DeShaney as acknowledging a “state-created danger exception.”132 This doctrine holds that a state assumes affirmative duties to protect a person when the state itself creates or increases the danger that ultimately causes the person’s harm.133 This theory of liability predated DeShaney in some circuits. Judge Posner, who articulated the conception of the Constitution as a “charter of negative liberties,” pointed out an oft-quoted “snake pit” exception: “[I]f the state puts a man in a position of danger from private persons and then fails to protect him, it will not be heard to say that its role was merely passive; it

---

132. See, e.g., Schwartz v. Booker, No. 11-1583, 2012 WL 6604196, at *4 (10th Cir. Dec. 19, 2012); Pena v. DePrisco, 432 F.3d 98, 109 (1st Cir. 2005) (stating that a government actor may be liable if he creates a danger through an affirmative act); Butera v. District of Columbia, 235 F.3d 637, 651 (D.C. Cir. 2001) (“We . . . hold[] that, under the State endangerment concept, an individual can assert a substantive due process right to protection by the District of Columbia from third-party violence when District of Columbia officials affirmatively act to increase or create the danger that ultimately results in the individual’s harm.”). The Fifth Circuit has rejected the state-created danger exception. Beltran v. City of El Paso, 367 F.3d 299, 307 (5th Cir. 2004). The First Circuit has discussed and apparently recognized the doctrine, but never found resulting liability. See Rivera v. Rhode Island, 402 F.3d 27, 35 (1st Cir. 2005) (discussing the status of the state-created danger doctrine within the circuit). The Fourth Circuit has also discussed the doctrine, but limits affirmative duties to facts involving some degree of custody. See Pinder v. Johnson, 54 F.3d 1169, 1175 (4th Cir. 1995) (en banc) (“This Court has consistently read DeShaney to require a custodial context before any affirmative duty can arise under the Due Process Clause.”).

133. See Chemerinsky, supra note 123, at 3 (stating that circuits continue developing the state-created danger doctrine).
is as much an active tortfeasor as if it had thrown him into a snake pit.”

No uniform test for state-created danger has developed, and the Supreme Court, at this point, has declined to weigh in. While circuits have crafted slightly different tests for what circumstances constitute state-created danger, their tests generally share several elements: (1) the state, through some affirmative action, created or increased a risk of harm; (2) the plaintiff—as opposed to the general public—must have been rendered more vulnerable to the harm; (3) the state must have known, or reasonably should have known, about the danger; and (4) the state’s conduct must shock the conscience.

Circuits continue to settle the general contours of both the state-created danger and functional custody exceptions. Context clearly plays a central role in both exceptions, and the context of voluntary psychiatric hospitalization raises unique concerns, as discussed in the next Part.

134. Bowers v. DeVito, 686 F.2d 616, 618 (7th Cir. 1982). Another frequently cited illustration from Judge Posner explains, “The state, having saved a man from a lynch mob, cannot then lynch him, on the ground that he will be no worse off than if he had not been saved.” K.H. ex rel. Murphy v. Morgan, 914 F.2d 846, 849 (7th Cir. 1990).

135. See McClendon v. City of Columbia, 305 F.3d 314, 324–25 (5th Cir. 2002) (reviewing different approaches taken by sister circuits).

136. See, e.g., Bright v. Westmoreland Cnty., 443 F.3d 276, 281–82 (3d Cir. 2006) (laying out the circuit’s test and emphasizing that the state must have misused—rather than failed to use—its authority); Armijo v. Wagon Mound Pub. Schs. 159 F.3d 1253, 1262–63 (10th Cir. 1998) (“[P]laintiff must show: (1) the charged state actors created the danger or increased the plaintiff's vulnerability; (2) the plaintiff [belonged to] a limited and specifically definable group; (3) the defendants' conduct put the plaintiff at substantial risk of serious, immediate, and proximate harm; (4) the risk was obvious or known.”); Kallstrom v. City of Columbus, 136 F.3d 1055, 1067 (6th Cir. 1998) (stating that states assume affirmative duties “by substantially increasing the likelihood that a private actor would deprive [an individual] of their liberty interest in personal security”); Mitchell v. Duval Cnty. Sch. Bd., 107 F.3d 837, 839 (11th Cir. 1997) (“In order for a plaintiff to hold the state liable under the ‘special danger’ analysis, he must show that the state affirmatively placed him in a position of danger which was distinguishable from that of the general public.”).
III. The Landscape of Mental Health Law and Public Psychiatric Hospitalization

This Part begins by reviewing the general laws governing voluntary and involuntary hospitalization. Then, it discusses forces that transformed state hospitals in the last several decades, affecting the voluntary-hospitalization context.

A. The Current Law of Voluntary and Involuntary Admissions

Patients with acute mental health needs can be admitted to a psychiatric facility in a variety of ways. If the individual...
agrees to the admission and gives informed consent, then he may be admitted voluntarily.139

Voluntary patients can be further subdivided into two categories: pure voluntary or conditional voluntary. Under a “pure” or “informal” voluntary admission, a patient can leave the hospital whenever he chooses.140 This freedom to walk out of the hospital may compromise therapeutic interventions, and so states often restrict pure voluntary admissions either by law or by policy and favor “conditional” or “formal” voluntary admissions.141 Typically, a formally voluntary patient who requests to leave the facility may be detained for a statutorily defined period—usually a few days—during which time the clinical staff evaluates whether the patient requires involuntary commitment.142 Based on this evaluation, clinical staff may institute involuntary commitment proceedings if warranted, and otherwise must discharge the patient once the statutory holding period elapses.143

139. See Appelbaum & Gutheil, supra note 46, at 43 (noting that all states currently allow voluntary admission).

140. Id. See, e.g., Mich. Comp. Laws § 330.1412 (2012) (“An informal voluntary patient shall be allowed to terminate his hospitalization and leave the hospital at any time during the normal day shift hours of the hospital, and the hospital shall so inform the patient at the time he is hospitalized.”).

141. Appelbaum & Gutheil, supra note 46, at 43. See, e.g., Mich. Comp. Laws § 330.1411 (restricting informal voluntary hospitalizations by requiring that “the hospital director consider[] the individual to be clinically suitable for that form of hospitalization”).


The term “voluntary” is misleading. These admissions often involve coercive factors like criminal charges or family pressures. Additionally, not only do the locked doors mean that voluntary patients are physically restricted from walking out of the facility, but most voluntary patients are legally restricted from leaving the hospital at will.

Alternatively, involuntary civil commitment procedures are available to hospitalize individuals who either refuse to consent to treatment or who lack capacity to consent. Some baseline constitutional standards limit state laws. States cannot involuntarily confine an individual simply because he has a mental illness: commitment requires that a mental illness is causing a person to pose some danger to himself or to others. Civil commitment also demands that a state prove, by clear and convincing evidence, that commitment is appropriate under state law.

Within these constitutional constraints, states enact varying laws. Most states permit emergency commitment based on a clinical psychiatric evaluation of statutory criteria, allowing for short-term, temporary hospitalization until a judicial hearing can be convened. Commitment hearings are formal proceedings


145. Appelbaum & Gutheil, supra note 46, at 39. See also supra note 142 (collecting state statutes permitting state hospitals to hold voluntary patients for evaluation after the patient requests to be discharged).

146. See Appelbaum & Gutheil, supra note 46, at 40–42 (discussing the history of, and rationales for, civil commitment).

147. See id. at 42–45 (reviewing the standards governing civil commitment).

148. See O’Connor v. Donaldson, 422 U.S. 563, 576 (1975) (“[A] State cannot constitutionally confine without more a nondangerous individual who is capable of surviving safely in freedom.”).

149. See Addington v. Texas, 441 U.S. 418, 431 (1979) (ruling that due process demands clear and convincing evidence because this standard “strikes a fair balance between the rights of the individual and the legitimate concerns of the state”).

150. See Appelbaum & Gutheil, supra note 46, at 42–43 (explaining that emergency commitments vary in length from two days to three weeks and
where judicial officers preside and patients are afforded some due process rights.151 State laws on involuntary commitment standards require different levels of “dangerousness” caused by a mental illness. Substantively, these laws generally encompass three elements: (1) the individual has a mental illness; (2) the individual is dangerous; and (3) the individual needs treatment.152 Some states are increasing commitment standards providing information on different clinical and evidentiary requirements under state law to institute emergency commitment). See, e.g., ARIZ. STAT. ANN. § 36-535(B) (2012) (6 business days); CONN. GEN. STAT. § 17a-502(a) (2012) (15 days); GA. CODE ANN. § 37-3-62 (2012) (15 days); 405 ILL. COMP. STAT. 5/3-706 (2012) (5 days).

151. See Appelbaum & Gutheil, supra note 46, at 43 (explaining that in the vast majority of states, judges—rather than administrative officers or juries—decide civil commitment hearings). See, e.g., ARIZ. REV. STAT. ANN. § 36-505 (2012) (guaranteeing patients an independent evaluation of their condition); CAL. WELF. & INST. CODE § 5256.4 (2012) (guaranteeing patients counsel, the right to present evidence, and the right to cross-examination); GA. CODE ANN. § 37-3-1(8) (2012) (requiring that patients have “effective assistance of counsel,” that hearings be recorded, and giving patients subpoena power, among other procedural rights); MD. CODE ANN., HEALTH–GEN. § 10-631 (2012) (requiring that involuntarily admitted patients receive notice in plain language informing them of their rights to consult counsel, and requiring that if the patient cannot understand the notice contents, his parent, guardian, or next-of-kin receive the information).

152. See Appelbaum & Gutheil, supra note 46, at 43–45 (reviewing the differences among state civil commitment laws). For specific examples of state statutes setting forth the requirements for involuntary civil commitment, see DEL. CODE ANN. tit. 16 §§ 5001, 5010 (2012) (requiring that mental illness “renders [a] person unable to make responsible decisions with respect to the person’s hospitalization” and that the person pose “a real and present threat of harm without immediate hospitalization”); FLA. STAT. § 394.4667(1) (2012) (requiring serious mental illness and either that the person pose a harm to himself or others, or that he is unable to function in the community); HAW. REV. STAT. § 334-60.2 (2012) (requiring that the person “is imminently dangerous to self or others, is gravely disabled or is obviously ill”); KY. STAT. § 202A.026 (2012) (requiring that the person present either a danger or threat of danger to self or others, and can reasonably benefit from treatment); MD. CODE ANN., HEALTH–GEN. § 10-632(g) (2012) (requiring a “mental disorder,” which requires inpatient care, and that the person “present[] a danger to the life or safety of the individual or of others”); MICH. COMP. LAW § 330-1401 (2012) (defining “person requiring treatment” to include an individual who, because of mental illness, cannot attend to his basic needs, or who cannot understand his need for treatment, is treatment noncompliant, and has consequently been violent or been placed in psychiatric facilities, prison, or jail twice in past forty-eight months).
by demanding proof of some threat, attempt, or actual occurrence of harm.\textsuperscript{153}

States continue to develop laws governing civil commitment. For example, most states have adopted “outpatient commitment” laws, which require certain patients to comply with psychiatric treatment in the community.\textsuperscript{154} As states rework the laws governing commitment, the context of voluntary hospitalization evolves.\textsuperscript{155}

\textbf{B. The Historical Context of Mental Health Law and Hospitalization}

Mental health law and public hospitalization should be understood in historical context.\textsuperscript{156} Sovereigns have confined mentally ill persons for centuries.\textsuperscript{157} At common law, the

\textsuperscript{153}. See \textsc{Appelbaum \& Gutheil}, \textit{supra} note 46, at 43 (explaining that heightened proof standards are intended to make the commitment determination an objective decision).

\textsuperscript{154}. See \textit{id.} at 48 (discussing outpatient commitment laws). Only a few states use these laws regularly because most states lack needed clinical and administrative structures at this time. \textit{Id.} The criteria for which individuals qualify for outpatient commitment vary; some states require a high likelihood that a patient will relapse into acute symptoms, while other states require a pattern of dangerousness. \textit{Id.} For examples of state statutes governing outpatient commitment orders and standards, see \textsc{Ala. Code} §§ 22-52-10.2 to -10.3 (2012); \textsc{Fla. Stat.} § 394.4655 (2012); \textsc{Ga. Code Ann.} §§ 37-3-90, 37-3-93 (2012) (permitting courts to order outpatient commitment for one-year periods if a physician concludes from examination that an individual “is a mentally ill person requiring involuntary treatment” and outpatient treatment is available); \textsc{50 Pa. Con. Stat.} § 4406(b) (2012) (“[A] court may permit partial hospitalization or outpatient care, or if at any time thereafter the director shall determine such partial hospitalization or outpatient care to be beneficial to the person so committed, the same may be permitted by said court upon application by the director.”).

\textsuperscript{155}. See \textsc{John Q. LaFond \& Mary L. Durham}, \textit{Back to the Asylum: The Future of Mental Health Law \& Policy in the United States} 53–56 (1992) (reviewing different proposals for modifying civil commitment law).

\textsuperscript{156}. See \textsc{Paul S. Appelbaum}, \textit{Almost a Revolution: Mental Health Law \& the Limits of Change} 12–13 (1994) (arguing that examining the evolution of mental health laws in recent decades “strengthen[s] our understanding of the complex interaction of forces involved when the law is applied to the mentally ill”).

\textsuperscript{157}. See Note, \textit{Civil Commitment of the Mentally Ill: Theories and
government was obligated to “take care of those who could not take care of themselves.” American courts continued applying this rule, but legal procedures changed as psychiatry evolved. In the twentieth century, mental health law and public psychiatric facilities transformed dramatically with regards to both the voluntary nature and the total number of admissions.

First, in the twentieth century voluntary admission became available and popular. During the first half of the century, almost every psychiatric patient was admitted involuntarily. Involuntary admissions were preferred for administrative convenience and a general belief that “the presence of mental illness per se rendered a person incompetent to consent to hospitalization.” State hospitals were considered important institutions for social reform, with psychiatric professionals bestowed significant power over their patients.

Beginning in the 1950s, psychiatric professionals and patients’-rights advocates fought to increase voluntary treatment. States responded, dramatically revising the laws governing admissions. By the early 1970s, voluntary admissions became more common than involuntary admissions. Currently, most patients are admitted voluntarily when both

---

158. See, e.g., Beverly’s Case, (1603) 76 Eng. Rep. 1118 (K.B.) 1124–26; 4 Co. Rep. 123 b, 126 a–127 b (discussing the King’s duties to “lunatics” and “idiots”).

159. See Commitment Theories & Procedures, supra note 157, at 1288 (explaining that “lunatics” came to be seen as ill, rather than as cursed).

160. See supra note 46, at 38 (“[T]he idea that the mentally ill might be able to sign themselves into psychiatric hospitals voluntarily is a relatively new one.”).

161. See id. (“[B]y 1949 only 10% of patients were voluntarily admitted [to psychiatric hospitals].”). No state permitted voluntarily admission until 1881. Id.

162. Id.

163. See supra note 155, at 84 (stating that before civil rights reforms, state actors could “make virtually all decisions for patients”).

164. See id. (explaining that civil libertarians opposed involuntary confinement, while psychiatrists favored voluntary admissions because they led to more effective treatment).

165. Id.
private and public/state hospitals are considered. Some states even adopted statutory presumptions or policy statements encouraging voluntary—rather than involuntary—admission and treatment. In state hospitals, however, involuntary admissions have recently become more common than voluntary.

Second, in the late-twentieth century the number of state hospital beds—and subsequently, the number of state-hospital patients—declined dramatically. In the mid-1950s, over 500,000 mentally ill persons sat “warehoused” in state institutions for years without effective treatment and lacking a mechanism to obtain release. By 2003, the number of state hospital beds had fallen by over 90%, down to 40,000 beds.

Deinstitutionalization grew from a combination of factors: changing societal attitudes about institutionalization; financial incentives to treat people in the community; new antipsychotic medications, allowing outpatient treatment; new federal

166. Id.

167. See, e.g., ALASKA STAT. § 47.30.655(1) (2012) (“Persons [should] be given every reasonable opportunity to accept voluntary treatment before involvement with the judicial system.”); COLO. REV. STAT. § 27-65-101(d) (2012) (declaring that the legislature intends to “encourage the use of voluntary rather than coercive measures to provide treatment and care”).

168. See LAFOND & DURHAM, supra note 155, at 43 (noting the trend); see also infra notes 183–84 and accompanying text (explaining the relationship between deinstitutionalization and involuntary admissions, as well as the effect on the patient populations in state hospitals).

169. See LAFOND & DURHAM, supra note 155, at 85 (describing “inhumane conditions in mental hospitals” in the 1940s).

legislation aimed at building outpatient treatment centers; and civil liberties lawyers’ advocacy. The civil liberties lawyers facilitated deinstitutionalization through several key lawsuits challenging hospital conditions and commitment procedures. These suits facially addressed civil rights violations in state hospitals, but were actually “targeted at closing down the hospitals.”

Deinstitutionalization caused “staggering” effects on the mentally ill population and mental health services. Although a full understanding of deinstitutionalization is beyond the scope of this Note, one important effect is relevant: The availability of state psychiatric hospitalization declined dramatically. States have slashed the number of state hospital beds. Because beds are scarce, states limit admission to patients with the severest symptoms. The structure of healthcare funding for psychiatric


172. See O’Connor v. Donaldson, 422 U.S. 563, 576 (1975) (ruling that civil commitment requires the patient be dangerous and unable to live safely in the community); Wyatt v. Alderholt, 503 F.2d 1305, 1313–14 (5th Cir. 1974) (upholding injunction ordering Alabama state hospitals to improve the condition and staffing ratios).


174. Id. at 50.

175. See Substance Abuse & Mental Health Servs. Admin., Funding & Characteristics of State Mental Health Agencies 69 (2009) [hereinafter SAMHSA Funding Report] (discussing the trend in most states to decrease state hospital services).

176. Id. One effect of deinstitutionalization is that many people with severe mental illness are now institutionalized in prisons and jails rather than hospitals. See Grant H. Morris, Refusing the Right to Refuse: Coerced Treatment of Mentally Disordered Persons 150–51 (2006) (stating that approximately 283,000 people with mental illness were incarcerated in a 1998 study—often for petty crimes—meaning that jails have become the largest mental health providers in the country).

177. See LaFond & Durham, supra note 155, at 51–52 (describing how deinstitutionalization and the consequent shortage of state hospital services
services encourages this practice by covering a predetermined amount of care in private facilities.\textsuperscript{178} State governments bear the entire cost of patients’ treatment in state hospitals, incentivizing states not to admit patients with available insurance coverage into the state-funded public facilities.\textsuperscript{179} As a result, the sickest individuals—those who run out of covered inpatient services—are more likely to be admitted to state hospitals.\textsuperscript{180}

The legal system also caused changes in state hospitals. State hospitals are also devoting more beds to forensic patients who come from the criminal system.\textsuperscript{181} In 2008, one-third of state-hospital patients had been criminally committed to the hospital.\textsuperscript{182} State psychiatric hospitals also recently witnessed a dramatic rise in the proportion of involuntarily committed patients, a trend that contrasts with the general rise of voluntary admissions noted above.\textsuperscript{183} The trend back to involuntary commitment in state hospitals stems from increasing forensic populations and general goals underlying deinstitutionalization.\textsuperscript{184} The proportional rise of involuntary patients indicates a corollary rise in the proportion of patients posing a risk of harm—a necessary precondition to involuntary commitment. State hospitals, then, face more violent patient populations.\textsuperscript{185}

The state hospital today differs sharply from the state hospital in previous decades.\textsuperscript{186} The sweeping reforms, sharp

\begin{itemize}
\item \textsuperscript{179} \textit{Id}.
\item \textsuperscript{180} \textit{Id}.
\item \textsuperscript{181} See SAMHSA \textsc{Funding Report}, \textit{supra} note 175, at 71 (stating that between 1993 and 2007, state hospital expenditures for forensic services increased from 10.7\% to 36\%).
\item \textsuperscript{182} \textit{Id}.
\item \textsuperscript{183} See \textsc{Lafond \& Durham}, \textit{supra} note 155, at 43 (suggesting a trend toward involuntary admissions in state hospitals).
\item \textsuperscript{184} \textit{Id} (explaining that in some states, public hospitals serve only involuntary patients in order to keep the patient census low).
\item \textsuperscript{185} \textit{Id} at 53.
\item \textsuperscript{186} See William H. Fisher, Jeffrey L. Geller \& John A. Pandiani, \textit{The
budget cuts, and concentration of more violent and more acutely ill patients suggest that courts should carefully guard patients' rights. But voluntary patients' rights are murky, with circuits divided on how to properly analyze affirmative duties.  

**IV. Circuit Split over Affirmative Duties to Voluntary Patients**

This distinction between patients admitted voluntarily versus involuntarily bears significance beyond just criteria for admission or discharge. Courts use this voluntary/involuntary distinction to determine whether the state owes a particular patient Youngberg rights, but the circuits disagree on how to properly analyze voluntary patients' rights. This circuit split predates DeShaney, but since DeShaney was handed down, that decision drives the train in courts' analyses. This Part examines two broad approaches that circuit courts take. First, several circuits look solely to a patient’s formal status as a voluntary or involuntary patient. Second, other circuits scrutinize the facts in each case, looking at the circumstances of a patient’s hospitalization to determine whether the patient’s commitment was voluntary or involuntary.

---

*Changing Role of the State Psychiatric Hospital, 28 HEALTH AFFAIRS 676, 676–81* (arguing that state hospitals evolved from being the primary treatment source for the mentally ill, and now primarily manage those people who cannot be managed elsewhere). The authors identify four groups of patients whom they predict will likely “define the state hospital’s mission for the foreseeable future”: (1) people with criminal justice histories; (2) forensic patients; (3) sexually dangerous persons; and (4) patients who are difficult to discharge, often because of the severity of their illness. *Id.* at 679–80.

187. *See infra* Part IV (discussing the circuit split).

188. *See supra* note 71 and accompanying text (stating Youngberg's rule that involuntarily committed mental health patients are owed certain affirmative duties); Lanman v. Hinson, 529 F.3d 673, 682 n.1 (6th Cir. 2008) (noting the circuit split on whether states owe affirmative duties to voluntary patients).

189. *See Lanman*, 529 F.3d at 682 n.1 (discussing the circuit split both before and after DeShaney).

190. *See infra* Part IV.A (reviewing cases applying status-based tests).

191. *See infra* Part IV.B (reviewing cases applying fact-based tests).
A. A Status-Based Approach to Affirmative Duties

One group of circuit courts applies the general no-duty rule broadly when voluntary patients bring Youngberg-type claims premised on affirmative duties of care or protection. Basing their affirmative-duty analysis on the patients' formal admission status, these circuits interpret DeShaney to preclude affirmative duties for everyone except civilly committed patients.

1. Sixth Circuit: Higgs v. Latham

The Sixth Circuit applied a status-based analysis focused on the voluntary/involuntary distinction in Higgs v. Latham. Josephine Higgs was admitted to Western State Hospital in complicated circumstances. Josephine was initially a patient at Grayson County Hospital. Her husband observed her mental illness worsening during her treatment at Grayson, he petitioned a court to order Josephine to be involuntarily hospitalized at Western State, and the court granted this order. Pursuant to this order, Josephine was transported via ambulance from Grayson to Western State, and was strapped down due to state officials' concern that she would harm herself. Upon arriving at Western State, however, Josephine was allowed to sign herself in as a voluntary patient because the Western State staff were...
never informed about the prior court proceeding. \textsuperscript{199} Had the hospital known about the court order, it would not have allowed Josephine to sign in as a voluntary patient. \textsuperscript{200} After Josephine was admitted, another patient sexually assaulted her, prompting Josephine and her husband to bring this § 1983 suit for monetary and injunctive relief. \textsuperscript{201} The case first went before a magistrate judge, who denied the State's motion for summary judgment, finding a genuine factual dispute over whether Josephine was, in fact, a voluntary patient. \textsuperscript{202} The district court reversed, granting summary judgment to the State on the grounds that Josephine was a voluntary patient with no constitutional right to affirmative care and protection. \textsuperscript{203}

The Sixth Circuit agreed that Josephine, as a voluntary patient, had no positive rights. \textsuperscript{204} The court began by analyzing \textit{Youngberg} and \textit{DeShaney}. \textsuperscript{205} The court interpreted \textit{Youngberg} as concluding that involuntary patients have affirmative rights, but as silent regarding voluntary patients. \textsuperscript{206} \textit{DeShaney}, however, does apply to voluntary patients, and under \textit{DeShaney}, affirmative duties depend "on the kind of restraint that disables a person from caring for himself." \textsuperscript{207} Applied to this case, the court reasoned that the state’s court order did not amount to an actual restraint because it was “unexecuted and unknown” to the hospital. \textsuperscript{208} For the hospital to take on affirmative duties, the hospital must have restrained Josephine. \textsuperscript{209} And because the

\textsuperscript{199} \textit{Id.} at *1–2.
\textsuperscript{200} See \textit{id.} at *1 (containing the testimony about the admission procedures).
\textsuperscript{201} \textit{Id.}
\textsuperscript{202} \textit{Id.} at *2.
\textsuperscript{203} \textit{Id.}
\textsuperscript{204} See \textit{id.} at *4 (“If the district court was correct in concluding that Mrs. Higgs was a voluntary patient at Western State Hospital, then it follows that she had no constitutionally based right of action against any of the defendants.”).
\textsuperscript{205} See \textit{id.} at *2–4 (examining these cases).
\textsuperscript{206} \textit{Id.} at *2.
\textsuperscript{207} \textit{Id.} at *3.
\textsuperscript{208} \textit{Id.}
\textsuperscript{209} \textit{Id.} at *4.
hospital did not know about the court order and because Josephine signed voluntary admission papers, “there was no ‘affirmative act’ by the hospital to deprive her of liberty,” and so the state’s duty-to-protect was never triggered.210

The majority rejected Josephine’s arguments that she was not competent to be voluntarily admitted and that hospital staff coerced her into signing the voluntary paperwork.211 First, regarding Josephine’s “allegedly confused state of mind,” the court refused to engage in this “highly problematic exploration of the state of mind of an acutely ill mental patient.”212 Second, the court rejected arguments that “advice” given by a nurse at Grayson that “it would be better for [Josephine] to admit herself voluntarily to Western State rather than being involuntarily committed there” amounted to constructive confinement.213

Judge Suhrheinrich concurred in the ruling but disagreed with the majority’s analysis of the voluntary/involuntary distinction,214 arguing that the majority misinterpreted DeShaney.215 To start, Judge Suhrheinrich contended that “DeShaney [does not] control the outcome of this case.”216 First, DeShaney does not preclude voluntary patients from establishing claims based on functional custody or on state-created danger.217 With respect to functional custody, Judge Suhrheinrich suggested that affirmative duties may have been triggered on the several occasions when Josephine requested to leave Western State, “and permission was refused.”218 With respect to state-created danger, affirmative duties may have been triggered when she was court-ordered for temporary commitment, given that Josephine probably would not have gone to Western State or signed

---

210. Id.
211. See id. at *4–5 (discussing Josephine’s claims that her admission was not voluntary).
212. Id. at *5.
213. Id.
214. Id. at *6 (Suhrheinrich, J., concurring).
215. Id.
216. Id.
217. See id. (“DeShaney’s custody test for liability may therefore be met even after it is shown that a patient was voluntarily admitted.”).
218. Id. at *6 n.1.
voluntary admission papers otherwise.\textsuperscript{219} Overall, the concurrence is noteworthy for its recommendation that voluntary patients should be able to make out affirmative-duty claims based on functional custody or state-created danger.\textsuperscript{220}

2. First Circuit: Monahan v. Dorchester Counseling Center\textsuperscript{221}

The First Circuit adopted a restrictive approach similar to the Higgs majority in Monahan v. Dorchester Counseling Center.\textsuperscript{222} Mr. Monahan was a voluntary patient at a state-operated group home called Millie’s Cottage.\textsuperscript{223} On the day the injury occurred, Mr. Monahan, accompanied by Cottage staff, went to a hospital emergency room and was evaluated for possible admission.\textsuperscript{224} The hospital staff determined that inpatient treatment was unnecessary and that he could return to the Cottage.\textsuperscript{225} While state employees were driving him back from the emergency room, Mr. Monahan jumped out of the car and walked toward an interstate highway.\textsuperscript{226} The driver, a state employee, did not call the police or try to stop Mr. Monahan; instead, he drove back to the group home and called a hospital, sending two other state employees to look for Mr. Monahan.\textsuperscript{227} During this time, a car struck Mr. Monahan, causing serious injury.\textsuperscript{228} Mr. Monahan then sued state officials for injunctive and compensatory relief stemming from his injuries, claiming they violated a duty to provide adequate

\begin{itemize}
\item \textsuperscript{219} Id.
\item \textsuperscript{220} Id. at *6 (“DeShaney compels us to go beyond asking whether Higgs was a voluntary admittee. DeShaney demands that we examine the limitations imposed on Higgs while she was a resident at the state hospital.”).
\item \textsuperscript{221} Monahan v. Dorchester Counseling Ctr., 961 F.2d 987, 991 (1st Cir. 1992) (ruling that a voluntary patient had no constitutional right to protection because the state did not restrict his liberties through civil commitment procedures).
\item \textsuperscript{222} Id.
\item \textsuperscript{223} Id. at 988.
\item \textsuperscript{224} Id.
\item \textsuperscript{225} Id.
\item \textsuperscript{226} Id. at 988–89.
\item \textsuperscript{227} Id. at 989.
\item \textsuperscript{228} Id.
\end{itemize}
supervision and treatment. The district court granted the defendants' motion to dismiss for failure to state a claim and also denied Mr. Monahan's motion to submit an amended complaint. Mr. Monahan appealed to the First Circuit.

In analyzing whether the state had violated any duty to Mr. Monahan, the First Circuit examined the relationship between the patient and the state. The facts differ from Youngberg because Mr. Monahan was not involuntarily committed. The facts also differ from DeShaney because Mr. Monahan was living in a state-operated facility, rather than a private home. Despite these differences from DeShaney's facts, the court applied DeShaney's rationale—that the Due Process Clause does not generally impose affirmative duties—and determined that the state had assumed no such duty. "Because the state did not commit Monahan involuntarily, it did not take an 'affirmative act' of restraining his liberty."

The court rejected Mr. Monahan's arguments that DeShaney's exceptions for functional custody or state-created danger applied, basing its rejection on Mr. Monahan's voluntary status. First, while acknowledging that a state assumes affirmative duties when it uses coercive measures to restrict an individual's liberty, the court determined that the state had not coerced Mr. Monahan because it had not initiated formal commitment proceedings. The court must look to whether the state's affirmative actions caused Mr. Monahan to give up some liberties. Mr. Monahan's mental illness—not the state—

---

229. Id.
230. See id. at 990 (discussing the history in the lower court).
231. Id.
232. See id. (comparing the instant case to Youngberg).
233. See id. ("His relationship to the state was therefore considerably closer than that of the plaintiff in DeShaney.").
234. See id. at 990–91 (stating that under DeShaney, Mr. Monahan "failed to state a viable claim for denial of substantive due process").
235. Id. at 991.
236. See id. at 992 n.5 (interpreting DeShaney's language to mean that "where the state's coercive power is not involved, there can be no constitutional (as opposed to tort) right to careful treatment").
237. Id. at 992.
238. Id. Although the circuit had ruled in pre-DeShaney cases that
deprived him of his liberty. Second, the state-created danger argument failed for similar reasons. The state’s actions might have made Mr. Monahan more vulnerable to harm. Again, Mr. Monahan “voluntarily availed himself of a Commonwealth service,” and so the state did not take on affirmative duties. Overall, then, the circuit’s analysis turned primarily on Mr. Monahan’s voluntary status—the state owed him no affirmative duties because the state had not initiated formal commitment proceedings.


The Fifth Circuit, sitting en banc, adopted a similar analysis in Walton v. Alexander. Christopher Walton was a voluntary residential student at the Mississippi School for the Deaf. Another student sexually assaulted Christopher several times. The state became aware of the first assault and instituted precautions, such as separating the students into different dormitories. But then budgetary constraints caused the school to close one of the dormitories and the boys were placed back in the same building, where the student sexually assaulted Christopher again. The district court denied the defendant’s

voluntary patients might have affirmative constitutional rights because of severe symptoms or de facto involuntary conditions, the court suggested that DeShaney demands different analysis. Id.

239. See id. (“His helplessness was not attributable to the [State taking] him into custody involuntarily.”).

240. See id. at 993 (agreeing with Mr. Monahan’s argument that “the Commonwealth could plausibly be said to have rendered him more vulnerable to danger”).

241. See id. (“The Commonwealth did not force Monahan, against his will, to become dependent on it.”).

242. Walton v. Alexander, 44 F.3d 1297, 1306 (5th Cir. 1995) (en banc) (ruling that a state has no duty to protect persons within its custody unless the state has taken affirmative steps to involuntarily confine them).

243. Id.

244. Id. at 1299.

245. Id. at 1299–1300.

246. Id.

247. Id. at 1300.
motion for summary judgment on qualified immunity grounds. The defendant filed an interlocutory appeal, in which a panel majority held that Superintendent Alexander had a special relationship with Christopher and, therefore, Christopher was entitled to *Youngberg* rights. 248 The Fifth Circuit then took up this interlocutory appeal en banc, reversing the panel majority’s decision and the district court’s denial of immunity. 249

The panel majority had ruled that Christopher was owed affirmative duties because the state had exercised functional custody. 250 Looking to language in *DeShaney* suggesting that states trigger affirmative duties by restraining someone’s liberty in a manner similar to incarceration, the court examined the factual setting of Christopher’s relationship to the state. 251 Many facts indicated “a significant custodial component” in this relationship: Christopher lived at the school; the school enforced strict rules; Christopher was “not free to leave” the school at will; and economically, most families had “no other viable option” for educating handicapped children. 252 These facts combined to make Christopher “dependent on the School for his basic needs and [he] lost a substantial measure of his freedom to act.” 253 The state, therefore, had functional custody and was obligated to provide Christopher with a reasonably safe environment. 254

The en banc panel disagreed. Summarizing its analysis, the Fifth Circuit explained, “we have followed [*DeShaney*’s] language strictly.” 255 If a state has affirmatively exercised its power to take custody over an individual against his will, then the state assumes affirmative duties. 256 But Christopher had “voluntarily subjected himself” to the state’s custody and had “the option of

248. *See id.* (discussing the prior history).
249. *Id.* at 1299.
251. *See id.* (describing the facts that indicate a custodial relationship).
252. *Id.*
253. *Id.*
254. *Id.*
256. *Id.*
leaving at will." Therefore, the state had no affirmative duty to protect him.

Judge Parker concurred specially in an opinion joined by three other circuit judges. The concurrence disagreed with DeShaney's proper interpretation, asserting that the en banc majority read DeShaney "erroneously" to require "a bright line rule that represents an extreme constitutional viewpoint." But this bright-line rule arbitrarily limits constitutional rights without accurately reflecting the facts of any given case.

The concurrence argued that the court should apply a factor-based test assessing the quality and nature of the relationship between patient and state. Reasoning that this test comports with DeShaney, Judge Parker looked to the DeShaney opinion. While DeShaney limited affirmative duties to contexts such as involuntary commitment and incarceration, the majority was explicit that "other similar restraint[s] of personal liberty" will trigger affirmative duties. Thus, the concurrence argued, courts

257. See id. at 1305 (analyzing Christopher’s case in light of DeShaney).
258. Id.
259. Id. at 1306 (Parker, J., concurring).
260. Id.
261. See id. at 1309 (stating that the majority’s rule arbitrarily assigns constitutional rights, and providing hypothetical illustrations of how the bright-line test fails to reliably and accurately measure state control).
262. See id. at 1309–10
   Instead of asking whether a person was taken into custody involuntarily, we should consider several factors to determine whether a special relationship exists in a particular case: 1) the authority and discretion state actors have to control the environment and behavior of the individuals in their custody, 2) the responsibilities assumed by the State, 3) the extent to which an individual in state custody must rely on the State to provide for his or her basic needs, and 4) the degree of control actually exercised by the State in a given situation.
263. See id. at 1307–08 (analyzing DeShaney’s majority opinion).
264. Id. (citing DeShaney v. Winnebago Cnty. Dep’t of Soc. Servs., 489 U.S. 189, 200 (1989)). The concurrence pointed out the absurdity created by the majority’s holding: "[S]tate actors entrusted with the responsibility to care for and protect our most vulnerable citizens may do so with constitutional impunity," while incarcerated "criminals are wrapped in the protective cloak of the constitution." Id. at 1310.
must analyze the restraints on liberty in order to determine whether constitutional duties are at play.265

The Walton majority, however, like the Monahan and Higgs courts, determined that affirmative-duty analysis turns on formal status. These courts interpret DeShaney to require involuntary commitment procedures before a state takes on affirmative duties.266 And, as the Higgs and Walton concurrences point out (albeit, critically),267 these courts avoid a fact-heavy analysis. Only facts about commitment proceedings are relevant, and the patient’s formal status is determinative.

**B. A Fact-Based Approach to Affirmative Duties**

Other circuit courts reject a strict status-based approach, concluding that voluntary status does not per se preclude Youngberg rights. Prior to DeShaney, several circuit courts ruled that voluntary status is irrelevant in these cases.268 After DeShaney, the Third, Eighth, and Eleventh Circuits continue to look beyond formal voluntary/involuntary labels.

---

265. Id. at 1310. Applying its test, the concurrence concluded that the state exercised sufficient control over Christopher to trigger affirmative duties. Id. However, the concurrence determined that the state actors were not liable because the plaintiffs failed to prove the state violated its duty by acting with deliberate indifference. Id. at 1307.

266. See supra notes 207, 235, 256 and accompanying text (discussing how the First, Sixth, and Fifth Circuits applied DeShaney to require involuntary commitment).

267. See supra notes 214–20 and accompanying text (summarizing Judge Suhrheinrich’s concurrence in Higgs); supra notes 259–65 and accompanying text (summarizing Judge Parker’s concurrence in Walton).

268. See, e.g., Soc’y for Good Will to Retarded Children v. Cuomo, 737 F.2d 1239, 1246 (2d Cir. 1984) (ruling that patients’ voluntary status is “irrelevant” in determining whether Youngberg rights apply, reasoning that Youngberg’s analysis applies equally to voluntary and involuntary patients); Goodman v. Parwatikar, 570 F.2d 801, 804 (8th Cir. 1978) (ruling that, although a state hospital has no obligation to admit anyone, the state owes affirmative duties to those individuals that it does admit); Harper v. Cserr, 544 F.2d 1121, 1124 (1st Cir. 1976) (ruling that a voluntary patient may assert a constitutional right to affirmative protection by establishing a sufficient level of helplessness).
1. Eleventh Circuit: Spivey v. Elliot

The Eleventh Circuit ruled in *Spivey v. Elliot* that whether a child receiving treatment voluntarily in a state-operated residential program has *Youngberg* rights depends on the level of control the state exercises. Tremain Spivey was a residential student at Georgia’s School of the Deaf, living at the school five days per week, with his mother’s consent. A thirteen-year-old classmate sexually assaulted Tremain on several occasions. Tremain’s mother subsequently withdrew him from the school and filed a § 1983 suit against the school, claiming that the school violated its duty to provide a safe environment under *Youngberg*. The district court granted the defendant’s motion for summary judgment on qualified immunity grounds. Tremain appealed.

The Eleventh Circuit issued two opinions responding to the State’s argument that Tremain had no right to its protection because he was at the school voluntarily. In its first ruling, the Eleventh Circuit disagreed with the State, interpreting *DeShaney* to mean that a state owes affirmative duties when it exercises sufficient control and dominion over an individual. The control and dominion requirement might be met whether a patient is voluntary or involuntary. Regarding Tremain’s voluntary status, “[t]he outcome of the case cannot turn on that

---

269. *Spivey v. Elliot*, 29 F.3d 1522, 1526 (11th Cir. 1994) (ruling that the state owed affirmative duties to a voluntary patient under *DeShaney’s* functional custody exception).

270. *Id.*

271. *Id.*

272. *Id.* at 1523.

273. *Id.*

274. *Id.* at 1523–24.

275. *Id.* at 1524.

276. *Id.*

277. *Id.* at 1526.

278. See *id.* (“The question is not so much how the individual got into state custody, but to what extent the State exercises dominion and control over that individual.”).

279. *Id.*
distinction."\textsuperscript{280} The court, however, affirmed the district court’s grant of summary judgment to the State on qualified immunity grounds.\textsuperscript{281}

In a later ruling, the court revisited its decision sua sponte, backing away from its affirmative duty analysis.\textsuperscript{282} Pointing out that Tremain’s claim failed on qualified immunity, the court decided that its affirmative duty analysis was unnecessary and should not serve as precedent.\textsuperscript{283} Therefore, while Spivey’s first ruling—that affirmative duties may be found independently of the voluntary/involuntary distinction—is not binding, the analysis is informative as to how the Eleventh Circuit might rule in the absence of qualified immunity issues.\textsuperscript{284}

\noalign{\vspace{1em}}
\noalign{\vspace{1em}}
\noalign{\vspace{1em}}

2. Eighth Circuit: Kennedy v. Schafer\textsuperscript{285} and Shelton v. Arkansas Department of Human Services\textsuperscript{286}

The Eighth Circuit applied an analysis similar to the Eleventh Circuit’s first Spivey decision in *Kennedy v. Schafer*.\textsuperscript{287}

\noalign{\vspace{1em}}
\noalign{\vspace{1em}}
\noalign{\vspace{1em}}

\textsuperscript{280.} Id.

\textsuperscript{281.} See id. at 1527 (deciding that the petitioner failed to show that his constitutional right was clearly established at the time the harm occurred).

\textsuperscript{282.} See Spivey v. Elliot, 41 F.3d 1497, 1498 (11th Cir. 1995) (“Upon reconsideration on the suggestion of other members of this Court, we now think it enough to decide that there was no clearly established constitutional right allegedly violated by the defendants.”).

\textsuperscript{283.} See id. at 1499 (“[T]his panel has chosen to withdraw all of its prior opinion which relates to whether the complaint alleges a constitutional right.”).

\textsuperscript{284.} The Supreme Court has ruled that courts should analyze whether a constitutional duty is established before reaching qualified immunity questions. See Siegert v. Gilley, 500 U.S. 226, 232 (1991) (discussing the proper structure of analysis).

\textsuperscript{285.} Kennedy v. Schafer, 71 F.3d 292, 295 (8th Cir. 1995) (remanding case in which voluntarily admitted psychiatric patient committed suicide and instructing lower court that if based on patient’s condition at time of suicide the facility could have lawfully detained her, then state had affirmative duty), cert. denied, 518 U.S. 1018 (1996).

\textsuperscript{286.} Shelton v. Ark. Dept’ of Human Servs., 677 F.3d 837, 839, 842 (8th Cir. 2012) (affirming that state owed no affirmative duties when voluntarily admitted psychiatric patient attempted suicide and, upon finding her alive, medical staff did not administer mouth-to-mouth resuscitation, leading to the patient’s death).

\textsuperscript{287.} *Kennedy*, 71 F.3d. at 295.
The patient in this case was Kathleen Kennedy, a fifteen-year-old girl who was admitted voluntarily to Hawthorn Children’s Psychiatric Hospital in Missouri. The hospital knew Kathleen was suicidal and her clinicians ordered suicide precautions, requiring that the nursing staff keep Kathleen within their eyesight. About a week after these precautions began, the hospital unit was short-staffed, and the nursing supervisor declined to schedule another nurse to work the evening shift. That evening, April 8, 1992, the nursing staff did not keep Kathleen within their eyesight. After 2:30 PM, staff did not see Kathleen for over three hours. At 5:10 PM, staff discovered that Kathleen had committed suicide. Her parents sued the State under § 1983, claiming that the hospital violated its duty to provide a safe and humane environment. The district court granted summary judgment to the State on the grounds that it did not owe Kathleen affirmative Youngberg duties because of her voluntary status, and alternatively on qualified immunity grounds. Kathleen’s parents appealed.

In analyzing whether the state owed Kathleen such a duty, the Eighth Circuit focused on the degree of control the state exercised over Kathleen. Its analysis resembles the Eleventh Circuit’s initial approach in Spivey, which found that Georgia’s control over Tremain Spivey placed him within DeShaney’s functional custody exception. The Eighth Circuit, however,
looked to whether Missouri exercised such control over Kathleen that “she had become, in effect, an involuntary patient.”299

Pointing out that Missouri law restricts patients who pose a substantial risk of harming themselves from leaving the hospital,300 and that this statutory language appeared on the voluntary admission papers Kathleen’s mother had signed, the court reasoned that Kathleen might not have enjoyed an absolute right to leave the hospital.301 The court remanded the case for a factual determination of how much control the state exercised over Kathleen.302

To support its analysis, the court pointed to DeShaney’s language that a state owes affirmative duties when it exercises functional custody, thereby creating a situation “sufficiently analogous” to involuntary hospitalization.303 It applied this language, however, not asking whether Kathleen’s situation was “analogous” to an involuntary patient’s, but by asking whether Kathleen had, in fact, become an involuntary patient.304 Through this analytic step, the court explicitly avoided addressing whether Youngberg might apply to voluntary patients.305 And while the

---


300. Id. at 295–96. The court arrived at this interpretation of Missouri law by reading three separate statutory provisions together. First, state statutes give the hospital discretion to refuse to discharge a minor psychiatric inpatient who is substantially at risk of harming herself, covering patients at risk for suicide. Mo. Rev. Stat. §§ 632.155(2), 632.005(10)(a) (2012). Second, if state psychiatric workers become aware that any person—including non-patients—is likely to cause serious harm because of a psychiatric illness, then the state actor is under a duty to evaluate the person’s condition. Id. § 632.300(1). Third, if the state actor concludes that this person poses an “imminent” risk of substantial harm because of a psychiatric illness, then the actor is obligated to initiate procedures for involuntary hospitalization. Id. § 632.300(2).

301. Kennedy, 71 F.3d at 295.

302. Id. at 296.

303. Id. at 295 (quoting DeShaney v. Winnebago Cnty. Dep’t of Soc. Servs., 489 U.S. 189, 201 n.9 (1989)).

304. See id. (stating that factors such as Kathleen’s clinical condition and state law “may have converted her status to that of an involuntary patient”); id. at 296 (“Facts change, and legal status follows facts.”).

305. Id. at 295 (“[T]his disposition makes it unnecessary to address the question whether a voluntary mental patient enjoys the same due process protections as an involuntary patient.”).
Kennedy court looked beyond Kathleen’s formal status, it nevertheless grounded its decision on her de facto status as either a voluntary or an involuntary patient.306

The Eighth Circuit recently revisited its Kennedy decision in Shelton v. Arkansas Department of Human Services, clarifying the circuit’s approach to voluntarily admitted patients.307 The facts in Shelton resemble those in Kennedy, with several key differences.308 Brenda Shelton signed in as a voluntary patient at Arkansas State Hospital because she was suicidal.309 The clinical staff initially placed her on “suicide watch,” but—unlike Kathleen Kennedy—these precautions were eventually removed.310 Three days later, nursing staff found Brenda in her room and discovered that Brenda had hanged herself.311 Brenda was unconscious, but still alive.312 The clinical staff, however, refused to provide mouth-to-mouth resuscitation.313 Medical equipment that might have helped revive Brenda was locked in a storage room; the room was unavailable because a nurse had locked the key inside the room.314 Brenda died a few days later.315 The administrator of her estate sued the State, the hospital, and several clinicians individually, alleging, inter alia, that the hospital violated Brenda’s substantive due process rights.316 The district court dismissed the lawsuit for failing to state a claim, ruling that the state did not owe Brenda affirmative duties because she was a voluntary patient.317

306. See id. at 296 (stating that Kathleen’s formal status is not determinative, and that the lower court must resolve whether Kathleen was, in effect, involuntary).
308. Compare id. at 839 (explaining the facts in Shelton), with supra notes 288–94 and accompanying text (explaining the facts in Kennedy).
309. Shelton, 677 F.3d at 839.
310. Id.
311. Id.
312. Id.
313. Id.
314. Id.
315. Id.
316. Id.
317. Id. at 839–40.
On appeal to the Eighth Circuit, the arguments concentrated on *Kennedy’s* proper application. Brenda’s estate argued that, under *Kennedy*, affirmative duties were triggered when the hospital staff discovered that Brenda had hanged herself. The court pointed out that the estate “appear[ed] to concede” that duties were not triggered when the clinical staff removed Brenda’s suicide precautions several days before. This fact differentiated Brenda’s circumstances from Kathleen Kennedy’s.

The two cases were further distinguished, the court explained, by whether the alleged constitutional violation occurred before or after the patient attempted suicide. The nurses failed to sufficiently monitor Kathleen before she committed suicide, while the alleged wrong in Brenda’s case occurred after the staff found her still alive. Brenda was “wholly incapacitated” and incapable of further harming herself. Consequently, any state statutes requiring involuntary treatment for persons at risk of self-harm could not have converted Brenda’s voluntary admission to involuntary (as was the case in *Kennedy*). Moreover, the court expressed reluctance to impose potential liability on state actors to emergency situations, which require “split-second, emergency-care decisionmaking.”

318. See id. at 840–42 (considering competing interpretations of the *Kennedy* decision and its impact on the instant case).
319. Id. at 840–41 (describing petitioner’s arguments).
320. Id. at 841.
321. See id. (explaining that *Kennedy* involved “at least a degree of” liberty deprivation because Kathleen was on suicide precautions, differentiating that case from *Shelton*, in which Brenda was not on suicide precautions when the injury occurred).
322. Id.
323. Id.
324. See id. at 842 (stating that, as a factual matter, Brenda posed no risk of additional self-harm because she was unconscious).
325. See id. (reasoning that Brenda’s situation resembled an unconscious patient brought into an emergency room; in such a situation, the state is not constitutionally obligated to provide treatment).
326. Id.
case," the court declined to extend Kennedy’s rule, although it left the rule intact.327

3. Third Circuit: Torisky v. Schweiker328

The Third Circuit also applied a fact-intensive analysis in Torisky v. Schweiker.329 In this case, the guardians of twenty individuals with mental retardation sued Pennsylvania officials after the patients were transferred between facilities.330 The patients were all being treated at the Western Center, and when the state decided to close this facility, it transferred the patients to other state facilities.331 The patients then brought § 1983 claims seeking monetary and injunctive relief.332 The circuit court reviewed the district court’s ruling that voluntary patients share the same constitutional rights under Youngberg that involuntarily committed patients enjoy.333

To resolve this issue, the court examined DeShaney and other circuit rulings. From the case law, the court first concluded that not every mental health patient has Youngberg rights.334 Even though voluntary patients reside in the state’s custody, the state may not have deprived them of their liberty.335 The court

327. Id.
328. Torisky v. Schweiker, 446 F.3d 438, 447 (3d Cir. 2006) (concluding that whether a voluntary patient possesses due process rights requires "looking beyond the label of an individual’s confinement to ascertain whether the state has deprived an individual of liberty in such a way as to trigger Youngberg’s protections").
329. Id.
330. Id. at 441.
331. Id.
332. Id.
333. See id. (defining the issue as “whether a state’s affirmative duty under the Due Process Clause to care for and protect a mental health patient in state custody depends upon the individual’s custody being involuntary”).
334. See id. at 444 (“[T]he substantive rights recognized in Youngberg are limited to persons whose personal liberty has been substantially curtailed by the state.” (quoting Fialkowski v. Greenwich Home for Children, 921 F.2d 459, 465 (3d Cir. 1990))).
335. See id. at 446 (“Thus, a custodial relationship created merely by an individual’s voluntary submission to state custody is not a ‘deprivation of liberty’ sufficient to trigger the protections of Youngberg.”).
extracted a common theme from other circuits’ rulings: whether a
state has deprived an individual’s liberty and triggered
Youngberg’s protections depends on “whether the individual is
free to leave state custody.”336 The court concluded that the
petitioners might be able to prove facts supporting that
Pennsylvania owed them affirmative duties of care.337

The Third Circuit ruled that courts should carefully
scrutinize the facts in order to determine whether voluntary
patients are owed affirmative duties, just as the Eighth Circuit
ruled in Kennedy and implicitly reaffirmed in Shelton.338 But
while Kennedy required a factual determination of whether the
state’s actions had “converted” a voluntary patient into an
involuntary patient,339 the Torisky court did not go so far. The
Third Circuit’s test asks whether the voluntary patient is free to
leave.340 Voluntary patients may be restricted from leaving a
hospital at will without the state initiating formal commitment
proceedings.341 The Third Circuit’s analysis therefore fits more
squarely within DeShaney’s functional-custody exception.342 Its
analysis does not explicitly rely on the voluntary/involuntary

336. Id. at 447. The court looked at cases including Kennedy v. Schafer,
discussed supra in section IV.B.2; Monahan v. Dorchester Counseling Center,
discussed supra in section IV.A.2; and Walton v. Alexander, discussed supra in
section IV.A.3.
337. See id. at 448 (concluding that “a constitutional violation may have occurred”).
338. See id. at 447 (stating that the court must look “beyond the label of an
individual’s confinement”); Shelton v. Ark. Dep’t of Human Servs., 677 F.3d 837,
841–42 (8th Cir. 2012) (applying Kennedy’s rule to the facts in the instant case);
Kennedy v. Schafer, 71 F.3d 292, 296 (8th Cir. 1995) (remanding the case for a
factual determination of Kathleen’s status), cert. denied, 518 U.S. 1018 (1996).
339. See supra notes 303–06 and accompanying text (analyzing Kennedy’s
analysis and ruling).
340. See Torisky v. Schweiker, 446 F.3d 438, 447 (3d Cir. 2006) (stating that
courts must examine the facts of a patient’s custody and determine “whether the
individual is free to leave state custody”).
341. See id. at 446–47 (discussing that patients who voluntarily enter a
hospital may face restrictions on their ability to leave, and noting that
Pennsylvania law allows a hospital to keep a voluntary patient in custody for up
to seventy-two hours).
342. See id. at 444 (referencing DeShaney’s rule that the state owes
affirmative duties to individuals over whom it exercises functional custody).
distinction but the court’s inquiries overlap with Kennedy’s de facto involuntary commitment analysis.

V. Concerns about Overreliance on Voluntary/Involuntary Distinction

Given the division among circuits over what analytic method courts should use to determine whether voluntary patients are owed affirmative duties, this Part suggests that an analysis putting less emphasis on voluntary status is appropriate. Categorizing patients as either voluntary or involuntary is easy, but often fails to accurately and reliably capture the full picture of a patient’s relationship with the state. This Part considers two reasons that a patient’s voluntary status may not accurately describe his relationship to the state: competency and coercion.

A. Competency: Can Acutely Mentally Ill Patients Give Informed Consent?

Voluntary admission requires that an individual express his agreement to be hospitalized by giving informed consent. Unless a court has determined that an individual is incompetent, the law presumes that a person can make personal decisions regarding medical treatment. Competency raises special issues in the setting of a psychiatric hospitalization.

When a person is acutely ill with a psychiatric illness—to the point of requiring inpatient care—is that person competent to give informed consent? For most of the twentieth century, all

343. The concurrence in Walton v. Alexander, 44 F.3d 1297, 1309–10 (5th Cir. 1995) (Parker, J., concurring), highlighted the problem with substituting voluntary status for a more thorough analysis of the facts.

344. See Bruce J. Winick, Competency to Consent to Treatment: The Distinction Between Assent and Objection, 28 Hous. L. Rev. 15, 15–16 (1991) (explaining that informed consent requires: informational disclosure; competency; voluntariness; and a decision).

345. See id. at 21–22 (discussing the presumption that adults are legally competent).

346. See id. at 18 (“[C]ompetency is one of the central questions of mental health law.”).
patients were hospitalized involuntarily, and courts believed that patients’ severe psychiatric symptoms rendered them per se incompetent. The Supreme Court voiced concerns about competence and voluntary consent in Zinermon v. Burch. In that case, a voluntary patient challenged his admission status, arguing that he had lacked the capacity to give informed consent when he signed his voluntary admission paperwork. The Court agreed with the patient, reasoning that mental illness, by its very nature, “create[s] special problems regarding informed consent,” meaning that hospital staff may not be justified in accepting a patient’s proffered consent for treatment “at face value.”

Generally, competence to make treatment decisions hinges on whether a patient can make rational decisions about his treatment or whether he can care for himself. Making these competency determinations imposes considerable burdens on clinicians and healthcare facilities. Finding a patient incompetent often requires clinicians to follow extra procedures in the course of providing treatment, because a surrogate decision maker may now make the patient’s decisions. Patients requiring psychiatric hospitalization may be legally incompetent because of their acute symptoms, but may nevertheless be treated as competent and permitted to sign the paperwork for voluntary admission.

347. See, e.g., Sporza v. German Sav. Bank, 84 N.E. 406, 408 (N.Y. 1908) (describing the state’s history of institutionalizing people adjudicated insane “on account of the necessity of protecting them and the public from their disordered minds and insane acts”).

348. See Zinermon v. Burch, 494 U.S. 113, 138–39 (1990) (affirming that petitioner stated a claim for relief in alleging that Florida violated his procedural due process rights by admitting him as a voluntary patient, when petitioner alleged he was incompetent to give consent at the time of signing the admission paperwork).

349. Id.

350. Id. at 133 n.18.


352. Id. at 143.

353. Id.

354. See id. (stating that in order to ensure a patient receives a fair competency determination, the patient should have a lawyer and access to a
In such a situation, the patient’s “voluntary” label threatens to mislead a court relying on the voluntary/involuntary distinction as a proxy for understanding the relationship between a patient and the state. The fact that a patient was voluntarily admitted means little when the admission was flawed. As the Supreme Court pointed out in Zinermon, a legally incompetent “voluntary” patient may very well be “unlikely to benefit from the voluntary patient’s statutory right to request discharge.”

B. Coercion: How do External Forces Influence a Patient’s Voluntary Status?

Related to capacity and informed consent are concerns that outside forces—such as professionals, family members, the legal system, or mental health policies—influence a patient’s status as voluntary or involuntary.

Apprehensions about clinical staff coercing patients have influenced mental health policies since the 1960s. As discussed

medical expert).

355. See Karna Halverson, Voluntary Admission and Treatment of Incompetent Persons with a Mental Illness, 32 WM. MITCHELL L. REV. 161, 166 (2005) (arguing that “allowing the patient to be voluntarily admitted based on his or her consent without any competency determination leaves too much room for abuse”); Albert B. Palmer & Julian Wohl, Voluntary Admission Forms: Does the Patient Know What He’s Signing?, 23 HOSP. & COMMUNITY PSYCH. 38, 38 (1972) (presenting results from a study, which indicated that only one of forty patients could recount the essential provisions of a signed voluntary admission form); Donald H. Stone, The Benefits of Voluntary Inpatient Psychiatric Hospitalization: Myth or Reality?, 9 B.U. PUB. INT. L.J. 25, 26 (1999) (“Voluntary psychiatric hospitalization should be the result of a competent and informed decision arrived at within a non-coercive environment. Hospitalization based on anything less is not only involuntary, but it is an infringement of personal liberty.”); id. at 36 (“[O]ften a mentally ill person, upon arrival at a psychiatric hospital, is disoriented or distressed. Because the patients are disturbed, confused, frightened, and distraught, there are indications that they are unable to comprehend the major step they take through self-admission.”).


357. See Parry, supra note 351, at 467 (discussing concerns that patients were coerced to accept voluntary hospitalization and care). Beyond questions of coercion in the admission process are questions about coercion in treatment. See Morris, supra note 176, at 171 (stating that professionals have long debated whether involuntarily committed individuals may refuse treatment). In general,
above, voluntary hospitalization became increasingly popular among legal and psychiatric professionals in the last century, and an increasing proportion of patients consented to voluntary admission during this same period.\textsuperscript{358} Similarly, as state-hospital professionals returned to preferring involuntary admissions, more state-hospital patients are being involuntarily admitted.\textsuperscript{359}

This power to influence patients’ decisions blurs the line dividing voluntary and involuntary patients. For example, the Supreme Court noted that if a voluntary patient is actually incapable of making informed decisions, this patient probably is not in a position to exercise his legal rights.\textsuperscript{360} The hospital setting increases the risk that patients’ decisions will be unduly influenced.\textsuperscript{361} Psychiatric clinicians treating acutely ill patients approach issues like coercion differently than do constitutional scholars.\textsuperscript{362} Legal scholars tend to focus on broad principles; clinicians, however, focus more on individual patients and how different actions will affect them.\textsuperscript{363} Clinicians’ results-based
orientation is arguably reasonable; most involuntary patients whose symptoms respond to treatment retrospectively agree that the treatment was in their best interest. But aside from arguments about the utility of coerced treatment, the fact remains that coercion is a frequent part of inpatient psychiatric care—even voluntary treatment. The “voluntary” label masks coercive elements, giving the impression that any given voluntary patient has fully and competently consented to psychiatric hospitalization.

Coercion might arise from a variety of sources. The simple threat of involuntary commitment leads some patients to sign voluntary admission forms. State law may incentivize patients to avoid involuntary commitment to avoid having a commitment order “on the record.” Other patients have described feeling internally coerced by their psychiatric symptoms. The legal system’s announced preference for voluntary admissions—as expressed by state statutes, court opinions, policy declarations, and scholars—may play out by encouraging patients to voluntary treatment, but at the expense of enjoying the legal protections springing from involuntary status.

ensuring that a patient continues engaging in treatment after discharge, and maintaining a safe and therapeutic environment for all patients. Id. at 1174 (reviewing empirical evidence about patients’ attitudes after receiving effective treatment).

Coercion might arise from a variety of sources. The simple threat of involuntary commitment leads some patients to sign voluntary admission forms. State law may incentivize patients to avoid involuntary commitment to avoid having a commitment order “on the record.” Other patients have described feeling internally coerced by their psychiatric symptoms. The legal system’s announced preference for voluntary admissions—as expressed by state statutes, court opinions, policy declarations, and scholars—may play out by encouraging patients to voluntary treatment, but at the expense of enjoying the legal protections springing from involuntary status.

The reasons given for voluntary admission include: (1) it involves less stigma to the patient; (2) it is less coercive; (3) it allows the patient to acknowledge a desire for help and treatment; (4) it respects
Coercion is often indirect. Voluntary patients may be influenced by their fellow patients.\textsuperscript{371} Currently, because state hospitals are admitting a higher proportion of involuntary, dangerous patients, voluntary patients are affected in the hospital’s milieu and often receive less clinical attention than their more aggressive peers.\textsuperscript{372} Moreover, many voluntary patients only agree to hospitalization in the first place because their family or caregiver pressured them to get inpatient treatment.\textsuperscript{373}

These coercive elements, coupled with uncertainties about capacity, suggest that the voluntary/involuntary distinction cannot validly measure affirmative duties. Coercion and capacity go directly to the nature of the hospitalization.\textsuperscript{374} The label “voluntary” is an artificial signifier that communicates limited information about whether a patient truly made an informed, intelligent decision to be institutionalized.

---

\textsuperscript{371.} Miller, \textit{supra} note 362, at 1181.

\textsuperscript{372.} \textit{Id}.

\textsuperscript{373.} \textit{See id.} at 1210 (stating that patients’ families usually favor treatment).

\textsuperscript{374.} \textit{See id.} (stating that “mentally disordered persons are perhaps subject to more coercion than most other groups” and this coercion often leads to “voluntary” admissions).
VI. Proposed Analysis for Determining Duties to a Voluntary Patient

How should courts analyze whether a state-hospital patient was owed affirmative duties of care and protection under substantive due process? This subpart proposes a three-step analysis that avoids the pitfall of overrelying on the voluntary distinction.375

A. The Proposed Analysis

Courts should not use a patient’s formal voluntary status as a proxy for understanding a patient’s relationship to the state. The voluntary/involuntary distinction is unreliable and it unduly constrains the analysis.376 So how can courts determine which voluntary patients are owed Youngberg duties and which are not? DeShaney itself suggests a straightforward approach. Starting with the general no-duty rule, DeShaney then offered three exceptions, each triggering affirmative duties: formal custody, functional custody, and state-created danger.377 This translates cleanly into a three-part test that considers whether any of these exceptions apply when a plaintiff alleges that a State violated an affirmative duty.378

In the state hospital context, this three-part test provides courts with a logical method for analyzing affirmative duties. When a patient sues the State based on a claim that the state failed to provide an affirmative duty of care or protection, the court should begin with the general rule that no duty exists.379

375. Infra Part VI.A.
376. See supra Part V (arguing that the voluntary distinction fails as a screening tool).
377. See supra Part II.B.2 (analyzing DeShaney).
378. See Higgs v. Latham, No. 91-5273, 1991 WL 216464, at *6 (6th Cir. Oct. 24, 1991) (per curiam) (Suhrheinrich, J., concurring) (arguing that the court should “go beyond asking whether [the patient] was a voluntary admittee” and consider whether the state-created danger or functional custody exceptions apply).
379. See supra Part II.A (discussing the general no-duty rule in substantive due process).
Before addressing any qualified immunity arguments, the court should determine whether the plaintiff has established that he was owed affirmative duties at the time of injury.\textsuperscript{380} Next, the court should determine whether one of three exceptions apply.

First, affirmative duties arise if the state civilly commits the patient, taking formal custody.\textsuperscript{381} Patients with a formal “involuntary” status, like Mr. Romeo,\textsuperscript{382} fall into this exception. The voluntary/involuntary distinction is used as a screening tool for this exception, sorting out which patients are in the state’s formal custody. The distinction does not mean that involuntary status is the source of the affirmative duties, only that involuntary patients meet the standard for acquiring affirmative duties, as decided in \textit{Youngberg}.

Second, affirmative duties arise if the state acts to take functional custody over a voluntary patient.\textsuperscript{383} In this analysis, the court may look to state actions that potentially converted the voluntary hospitalization into a de facto involuntary one. For example, courts might find functional custody if the facts show that: the patient lacked capacity at the time he signed voluntary paperwork; staff coerced the patient to sign voluntary paperwork; the patient requested to be discharged and staff declined this request; or the staff exercised an extremely high degree of control over the patient, with the effect of excluding other people from helping the patient. In \textit{Higgs}, for example, the fact that Mrs. Higgs was under a court order for inpatient care and that her requests to be discharged were denied could create an affirmative duty.\textsuperscript{384} By contrast, the functional custody exception should fail in \textit{Monahan} unless Mr. Monahan could show that he was coerced to remain in the state’s care.\textsuperscript{385}

\textsuperscript{380} See Cnty. of Sacramento v. Lewis, 523 U.S. 833, 841 n.5 (1998) (stating that courts should determine whether a plaintiff was owed a constitutional duty before reaching qualified immunity arguments).

\textsuperscript{381} See supra Part II.B.1 (discussing \textit{Youngberg}’s rule that involuntary commitment triggers affirmative duties).

\textsuperscript{382} See supra Part II.B.1 (discussing \textit{Youngberg}).

\textsuperscript{383} See supra Part II.C.1 (discussing the functional custody exception).

\textsuperscript{384} See supra Part IV.A.1 (discussing \textit{Higgs}).

\textsuperscript{385} See supra Part IV.A.2 (discussing \textit{Monahan}).
Third, affirmative duties arise if the state created or increased the danger that caused the patient’s harm.\textsuperscript{386} Courts again should apply the standards developed within their own circuits for state-created danger.\textsuperscript{387} State-created danger may be found if the plaintiff can show that: the state had documented awareness either that the patient was substantially at risk to injure himself or be injured by another person; the staff cut off other sources of aid, such as limiting visitors’ access or not informing family about the patient’s risk; and the state, by a direct action or policy, increased the likelihood that the patient would suffer the foreseen harm. If another patient inflicted the injury, courts should scrutinize whether the hospital reasonably should have known that the wrongdoer might commit this harm. State-created danger might apply in \textit{Shelton}, given that Brenda was an inpatient, cut off from any other sources of aid, and hospital staff refused to administer resuscitation—although this outcome is less clear.\textsuperscript{388} \textit{Kennedy} presents a clearer case of state-created danger because staff—aware that Kathleen presented such a serious risk of suicide that she required constant staff presence—failed to follow these ordered precautions.\textsuperscript{389} The state hospital cannot become a “snake pit” without the state acquiring affirmative duties.\textsuperscript{390}

\textbf{B. Benefits of this Approach}

The Proposed Analysis offers a straightforward application of \textit{DeShaney} and \textit{Youngberg}, nudging courts to look beyond voluntary/involuntary issues. This subpart explains three benefits from this analysis: it considers the voluntary/involuntary

\begin{itemize}
\item \textsuperscript{386} See \textit{supra} Part II.C.2 (discussing the state-created danger exception).
\item \textsuperscript{387} See \textit{supra} note 136 and accompanying text (reviewing state-created danger tests).
\item \textsuperscript{388} See \textit{supra} Part IV.B.2 (discussing \textit{Shelton}).
\item \textsuperscript{389} See \textit{supra} Part IV.B.2 (discussing \textit{Kennedy}).
\item \textsuperscript{390} See Bowers v. DeVito, 686 F.2d 616, 618 (7th Cir. 1982) (analogizing the state-created danger scenario to the government throwing a man into a snake pit).
\end{itemize}
distinction appropriately, it comports squarely with substantive
due process doctrine, and it does not unduly burden states. 391

1. The Proposed Analysis Appropriately Considers the Voluntary
Distinction

A primary benefit of the proposed analysis is that it accords
the voluntary distinction appropriate weight. The
voluntary/involuntary distinction determines whether the formal
custody exception applies. Civilly committed patients fall clearly
within Youngberg and their affirmative rights are clearly
established. 392

For voluntary patients, however, the analysis limits the
importance of the voluntary/involuntary distinction. Consistent
with most circuits, the analysis rejects the idea that all voluntary
patients have Youngberg rights. 393 But the analysis also rejects
the idea that formal commitment procedures are required. 394
Instead, the analysis directs courts to apply the functional
custody and state-created danger exceptions. Circuits cultivated
these doctrines in the years since DeShaney, and there is no
reason that courts cannot apply these tests in state hospital
contexts. 395

391. See infra Parts VI.B.1–3 (discussing each benefit in turn).
392. See supra note 67 and accompanying text (explaining Youngberg’s rule).
393. See, e.g., Torisky v. Schweiker, 446 F.3d 438, 446 (3d Cir. 2006) (“[A]
custodial relationship created merely by an individual’s voluntary submission to
state custody is not a ‘deprivation of liberty’ sufficient to trigger the protections
of Youngberg.”).
394. See id. (rejecting the argument that “a court commitment to state
custody is a necessary characteristic of a deprivation of liberty sufficient to
trigger Youngberg[.]”).
395. See Walton v. Alexander, 44 F.3d 1297, 1309 (5th Cir. 1995) (Parker, J.,
concurring) (“Rather than simply asking whether a person entered state custody
‘voluntarily,’ we should examine the nature of the custodial relationship that
existed between the State and the plaintiff.”); Higgs v. Latham, No. 91-5273,
1991 WL 216464, at *6 (6th Cir. Oct. 24, 1991) (per curiam) (Suhrheinrich, J.,
concurring) (arguing that courts should “go beyond asking whether [the patient]
was a voluntary admittee” and consider whether the state-created danger or
functional custody exceptions apply).
The proposed analysis protects voluntary patients’ rights. Psychiatric patients—both voluntary and involuntary—face barriers to vindicating their legal rights. And the patients labeled “voluntary” may be even more vulnerable, given that their admissions may be tainted by coercion or flawed consent. The analysis recognizes this vulnerability and allows courts some discretion in ruling on a particular set of facts. This discretion is fitting. Substantive due process formulates a concept less rigid and more fluid than those envisaged in other constitutional provisions. Its application is less a matter of rule. Asserted denial is to be tested by an appraisal of the totality of facts in a given case. That which may, in one setting, constitute a denial of fundamental fairness, shocking to the universal sense of justice, may, in other circumstances, and in the light of other considerations, fall short of such denial.

Tests relying on the voluntary distinction constrain courts by obstructing their ability to find affirmative duties in specific cases.

Giving courts more flexibility to find affirmative duties to voluntary patients furthers the goals of § 1983. Section 1983 should not “supplant traditional tort law,” but neither should § 1983 claims be rendered useless. As the Supreme Court explained in *Monroe v. Pape*, § 1983 claims should

---

396. See *Wyatt v. Aderholt*, 503 F.2d 1305, 1316 (5th Cir. 1974).

Mental patients are particularly unlikely to be aware of their legal rights. They are likely to have especially limited access to legal assistance. Individual suits may be protracted and expensive, and individual mental patients may therefore be deterred from bringing them. And individual suits may produce distortive therapeutic effects within an institution, since a staff may tend to give especially good—or especially harsh—treatment to patients the staff expects or knows to be litigious.

397. See *supra* Part V (discussing how issues related to competence and coercion may render a formal voluntary admission status misleading).


supplement state law claims. If a voluntary patient makes out sufficient facts to convince the court that the state owed him affirmative duties, he must be able to enforce this right.

Moreover, § 1983’s remedies—the injunction and litigation costs, specifically—are powerful and important tools for prompting institutional reform in state hospitals. As state hospitals evolve by closing down beds, admitting more criminal defendants, and treating more dangerous and more seriously ill patients, § 1983 should retain its power as a sword. The Supreme Court recently indicated that courts cannot limit the injunction’s potency by denying this relief to a broad class of people. The analysis offers voluntary patients the chance to persuade a court to use this sword.

2. The Proposed Analysis Retains Doctrinal Integrity

The proposed analysis is also faithful to existing law. First, with regard to substantive due process, doctrinal integrity is imperative because substantive due process is vulnerable to

401. See id. at 171 (concluding that Congress intended for the Act to “give a remedy to parties deprived of constitutional rights”). After reviewing legislative history, the Court found that Congress planned for the Act to serve three purposes: to override invidious state laws, to provide remedy when state law insufficiently protected rights, and to provide additional remedies when state remedies were inadequate or impractical. Id. at 173–74.

402. See Marbury v. Madison, 5 U.S. 137, 163 (1803) (“The very essence of civil liberty certainly consists in the right of every individual to claim the protection of the laws.”).

403. See, e.g., United States v. Tennessee, 615 F.3d 646, 648 (6th Cir. 2010) (refusing to vacate outstanding orders stemming from a district court’s original 1993 injunction against a state hospital that was violating patients’ substantive due process rights); Thomas v. Flaherty, 902 F.2d 250, 251–52 (4th Cir. 1990) (upholding a district court’s order for injunctive relief for a class of patients whose constitutional rights were violated by the “deficient care” in state hospitals).

404. See supra Part III.B (discussing changes in state hospitals since the 1950s).

confusion and misapplication.\footnote{406} Courts frequently approach claims purporting to potentially expand individual rights with “wariness and even embarrassment.”\footnote{407} By deriving its analysis directly from \textit{DeShaney}, the analysis avoids arbitrarily expanding substantive due process’s protections. After focusing courts’ attention through \textit{DeShaney}’s lens, the analysis points to specific factors relevant to state hospitals. For example, questions about coercion and competence raise significant concerns in state hospitals, but are probably irrelevant in contexts lacking any potential for the victim to “consent to” the state’s control. The analysis captures a clear picture of a patient’s relationship to the state, rather than a mere glimpse of voluntary status.

Second, the proposed analysis is more consistent with legal doctrine related to custody than are the status-driven approaches some circuits currently apply. While the Supreme Court has not addressed whether voluntary patients are “in custody,” it has offered guidance on this question in the Fourth Amendment context.\footnote{408} Custody assessments—at least for Fourth Amendment purposes—require courts to approach the situation from the individual’s point of view, not the state’s.\footnote{409} How do these rules relate to custody for substantive due process purposes? The Sixth Circuit argued that in the context of voluntary patients, custody claims under the Fourth and Fourteenth Amendments should be treated identically.\footnote{410} The Third Circuit similarly ruled in \textit{Torisky} that affirmative duties must depend on whether the patient is free to leave.\footnote{411} Because custody turns on the individual’s objective perspective, in state hospitals, a patient’s legal


\footnote{407. \textit{Id}.}

\footnote{408. See \textit{Stansbury v. California}, 511 U.S. 318, 323 (1994) (per curiam) (stating that whether or not a person is “in custody” depends on the perspective of a reasonable person in the suspect’s position).}

\footnote{409. \textit{Id}.}

\footnote{410. See \textit{Lanman v. Hinson}, 529 F.3d 673, 683 (6th Cir. 2008) (“Differentiating Fourteenth Amendment cases from . . . Fourth Amendment [cases] based on the voluntary or involuntary nature of the state’s custody would lead to arguably inconsistent results.”).}

\footnote{411. See \textit{supra} note 337 (giving \textit{Torisky}’s ruling).}
voluntary status should not be conflated with his *reasonably perceived* status. If a voluntary patient reasonably believes that he has no right to leave, his “voluntary” label means little.

Third, the proposed analysis offers courts a coherent framework for analysis. This point is highlighted by *Kennedy*. In that case, the Eighth Circuit found affirmative duties by determining that Kathleen, while formally a voluntary patient, may have been constructively involuntary. By stretching the category of “involuntary” patients in this way, the decision unnecessarily muddies the law. The court’s rationale that “[f]acts change, and legal status follows facts,” would suggest that if the facts of *Kennedy* remained the same, but Kathleen had not killed herself, she nevertheless would have been involuntary—but without any of the due process protections afforded by the civil commitment process. *Zinermon* directly prohibits such a result. The *Kennedy* court’s analysis comports more squarely within the proposed analysis, allowing it to neatly find affirmative duties arising from state-created danger, rather than stretching the voluntary/involuntary distinction beyond its fibers.

3. The Proposed Analysis Will Not Unduly Expand Affirmative Duties

Allowing some voluntary patients *Youngberg* rights is unlikely to produce intolerable policy results. States frequently assert that imposing liability will force unsustainably weighty burdens on them and cause a disastrous fallout. These concerns should be viewed with a critical eye. For example, some advocates charge that states will stop admitting voluntary

---

413. *Id.* at 296.
patients altogether because the risk of liability “is too great to jeopardize” state resources.\textsuperscript{416} Aware that current “voluntary” patients might be relabeled “involuntary” by some court in the future, the argument goes, hospital staff will be left unsure whether they owe affirmative duties.\textsuperscript{417}

To be sure, increasing potential liability in state hospitals will produce corollary burdens. Underfunded state mental health systems may be forced to spend their tight budgets on legal fees and judgments.\textsuperscript{418} State hospital workers would also face increased exposure to personal liability, and may be unable to obtain insurance coverage for constitutional violations.\textsuperscript{419}

When considered in light of the current state hospital system,\textsuperscript{420} however, these concerns appear overblown. First, it seems implausible that states will stop admitting and treating voluntary patients. History belies such an argument. \textit{Youngberg} explicitly granted affirmative duties to every involuntary patient, but states did not react by refusing to treat involuntary patients, a decision that would have been within states’ discretion.\textsuperscript{421} Why would a different result follow if some voluntary patients are granted affirmative rights? Hypothetically, if states did react by shutting their doors to voluntary patients, then they would correspondingly open their doors to more involuntary patients.\textsuperscript{422} The result would be that every patient would possess affirmative \textit{Youngberg} rights.\textsuperscript{423} Allowing voluntary patients to establish

\begin{itemize}
\item \textsuperscript{416} Id.
\item \textsuperscript{417} Id. at *16–17.
\item \textsuperscript{418} See id. at *17–19 (arguing that state budgets will suffer potentially untenable consequences if voluntary patients are found to have affirmative rights).
\item \textsuperscript{419} See \textsc{Appelbaum & Gutheil}, supra note 46, at 138 (discussing due process claims against individual clinicians).
\item \textsuperscript{420} See supra Part III.B (discussing how state hospitals evolved in recent decades).
\item \textsuperscript{421} See supra Part II.B.1 (discussing \textit{Youngberg}); see also supra note 183 and accompanying text (noting that state hospitals are trending toward involuntary patients).
\item \textsuperscript{422} See SAMHSA \textsc{Funding Report}, supra note 175, at 69–70 (stating that many states currently face a shortage of state hospital beds).
\item \textsuperscript{423} See Youngberg v. Romeo, 457 U.S. 307, 316–17 (1982) (ruling that states owe affirmative duties under substantive due process to involuntarily committed individuals).
\end{itemize}
affirmative constitutional rights, then, should not be precluded out of concerns for state budgets.

Second, state actors are protected by the deferential professional judgment standard for imposing liability.424 The professional judgment standard is more deferential than the “reasonable care” standard in malpractice claims.425 The reasonable care standard imposes liability if a defendant’s conduct deviates at all from what a reasonable professional would do; the professional judgment standard imposes liability only for a substantial deviation.426 The Supreme Court chose this standard in Youngberg to avoid overly burdening states.427 The Court reasoned that the professional judgment standard protects hospitals and individual clinicians by relieving them of a burden “to make each decision in the shadow of an action for damages.”428 With this standard’s protections, state actors will be insulated from an uncontrolled new wave of liability.

VII. Conclusions

The government owes us nothing. DeShaney highlighted this general principle with its unrelentingly formal analysis of a shockingly tragic case.429 And DeShaney’s message has been received in lower courts.430

424. See id. at 324 (“In determining whether the State has met its obligations . . . decisions made by the appropriate professional are entitled to a presumption of correctness.”).

425. See Parry, supra note 351, at 652 (discussing both standards).

426. Id.

427. See Youngberg, 457 U.S. at 324 (“Such a presumption is necessary to enable institutions of this type—often, overcrowded and understaffed—to continue to function.”).

428. Id. at 325.

429. See supra Parts II.A, II.B.2 (discussing the no-duty rule and DeShaney).

430. See Schwartz, supra note 5, § 3.09[B] (“DeShaney has generated an unusually large volume of important lower federal court rulings.”); see also Doe v. Milwaukee Cnty., 712 F. Supp. 1370, 1371 (W.D. Wis. 1989) (“Joshua DeShaney will never know it, unfortunately, but he has had a dramatic impact on constitutional law. The case that grew out of a tragedy . . . is already affecting law suits across the country.”).
But amidst this post-DeShaney storm, courts must not overlook the exceptional cases. Sometimes the government does owe a duty to protect or care for individuals.431 State hospitals present a complicated context for evaluating when affirmative duties apply, and state-hospital patients are particularly vulnerable.432 A patient’s relationship with the state is more complicated than his voluntary status. Courts falter when they conflate a patient’s voluntary status—either formal or de facto—with a sound legal test for affirmative duties.

Voluntary status should not preclude affirmative duties. Courts should have some modicum of discretion to provide a just ruling in the case at hand.433 The proposed analysis suggested in this Note offers a flexible standard without sacrificing doctrinal integrity.434

Stepping back and reflecting on DeShaney and its wake, it’s almost surprising to remember that some of the most important facts remain a mystery. Wisconsin never explained why it chose not to intervene to protect little Joshua, even as the social worker dutifully chronicled her suspicions “in detail that seems almost eerie.”435 No explanation was required once the Court found that no duty existed.436

But DeShaney explicitly demanded that sometimes an explanation is needed. Sometimes state actors should be called upon to explain their decisions. In cases involving state-hospital patients, courts should not gloss over the special issues that may

431. See supra Parts II.B.1, II.C.1–2 (discussing Youngberg, the functional custody exception, and the state-created danger exception).

432. See supra Parts III, V (reviewing mental health law and the nature of state hospitals).

433. See, e.g., BENJAMIN CARDOZO, THE NATURE OF THE JUDICIAL PROCESS 66 (1921)

There is an old legend that on one occasion God prayed, and his prayer was “Be it my will that my justice be ruled by my mercy.” That is a prayer which we all need to utter at times when the demon of formalism tempts the intellect with the lure of scientific order.

434. See supra Part VI (containing the proposed analysis and arguing its strengths).


436. Id.
arise. Courts should ask whether the state created a duty, whether by civilly committing a patient, by exercising functional custody, or by creating a danger. When the government takes those actions against us, then it does owe us something.
Joshua’s Children: Constitutional Responsibility for Institutionalized Persons After DeShaney v. Winnebago County

Susan Stefan*

Table of Contents

I. Introduction ................................................................. 794

II. DeShaney and the Real World of Institutionalized Psychiatric Patients ......................................................... 797

III. Chaos in the Circuits: The Interpretation of DeShaney’s Application to Institutionalized Patients ........................ 802

A. The DeShaney Decision is not Concerned with Institutionalized Individuals ............................................ 803

B. The Patient’s Legal Status on Admission as Voluntary or Involuntary is not Determinative of Constitutional Rights ............................................. 805

C. Functional Custody: An Alternative to Formal Legal Status in Determining the Affirmative Constitutional Rights of Institutionalized Patients . 814

D. State-Created Harm and State Increase of Danger in the Context of Institutionalized Patients ............. 816

1. Actions by State Actors vs. Private Individuals ............................................................................. 816

2. State-Created or -Exacerbated Harm in the Context of Suicidal Patients .......................... 817

* Visiting Professor, University of Miami School of Law. Dedicated to my clients who never made it out of institutions, especially Bob C., and to my former students who devote their lives to ensure justice for the people who remain institutionalized, especially Beth Mitchell and Michael Recco.
I. Introduction

“Poor Joshua!”¹

I cried when I read those words. It isn’t often that a Justice of the Supreme Court expresses understanding that real people are affected by the Court’s rulings, let alone grieves for them. It was an extraordinary moment in a Supreme Court dissent, a heartfelt response to a majority opinion in which the heart played no part.

But, as Claire Hagan’s Note² makes clear, it is not only in matters of the heart that the majority opinion in DeShaney v. Winnebago County³ falls short. The Supreme Court used DeShaney to set out a framework limiting the power of citizens to claim affirmative constitutional rights under the Constitution,⁴ but its chosen framework is far less workable than the majority would have us believe. While federal constitutional protections are obviously not boundless, the lines drawn by the Supreme Court have led lower courts to chaotic and inconsistent interpretations and outcomes in cases involving institutionalized individuals.

Ms. Hagan presents this case in a detailed and well-argued piece. She persuasively contends that in deciding DeShaney, the Supreme Court set the stage for the inevitable confusion and division among the circuits that continues to this day.

Ms. Hagan begins by describing the DeShaney decision, pointing out errors in the assumptions on which the majority

---

3. DeShaney, 489 U.S.
4. See id. at 199–201 (stating that affirmative constitutional rights arise only when the state, through an affirmative action, restrains a person’s liberty, prevents the person from acting on his own behalf, or creates the danger at issue).
based its holding. Then, linking *DeShaney* with the Court’s decisions in *Youngberg v. Romeo* and *Zinermon v. Burch*, she surveys the intersection of these cases in six different circuits, summarizing the current divergent approaches to determining the affirmative rights of institutionalized individuals. She gives the reader a sense of how people in psychiatric institutions actually experience their admission and institutional treatment, and she exposes the pretextuality of the concept of voluntariness in the institutional setting. Finally, she proposes a clear test that could be effectively applied by courts in determining whether an institutionalized individual may claim affirmative constitutional rights.

This is an excellent Note, and there is little to disagree with in its approach. My Comment will begin by summarizing what I have learned from almost thirty years of experience about what actually happens in institutional settings. Ms. Hagan accurately sets forth both the coercion and the loss of even the most elementary forms of choice and control represented by life in an institution. These everyday realities of state institutional settings are not commonly known to the public, and while much has improved since I began this work in a law school clinic in 1984, the need to protect patients against abuses unique to the culture of the institutional environment still remains strong in many state facilities.

Although the facilities have improved, the legal barriers to vindicating constitutional rights for institutionalized individuals have actually increased since the early 1980s. Increasingly conservative courts have imposed more and more roadblocks,

---

8. See Hagan, *supra* note 2, at Part IV (reviewing the split among the First, Second, Third, Fifth, Sixth, and Eighth Circuits regarding how to analyze whether states owe affirmative duties to voluntary patients).
9. See id. at Part V (describing how problems with competency and coercion affect voluntariness).
10. See id. at Part VI (setting forth a proposed analysis).
11. *Infra* Part II.
both substantive and procedural, to litigation on behalf of disenfranchised and vulnerable populations. The barrier in some circuits to affirmative constitutional rights for legally voluntary institutionalized patients is one major example of this trend.

In the second part of the Comment, I will do what public interest lawyers in my field have always done: translate the real-life experiences of patients in institutional settings into the formal structure of the law. Ms. Hagan has done an excellent job of surveying the tangled, complex, and contradictory reactions of the circuits to *DeShaney* by describing the leading cases relating to the affirmative constitutional rights of institutionalized psychiatric patients. I will complement her work by examining additional federal and state court cases involving these rights, as well as the most recent caselaw specifically interpreting *DeShaney*. Not surprisingly, the messages from the circuits continue to be at odds with each other, and some circuits do not even have an internally consistent approach. The conflicts in the circuits are so numerous that it’s surprising there has not been a Supreme Court follow-up clarifying the interpretation of *DeShaney* in institutional settings.

Finally, I will evaluate Ms. Hagan’s proposed standard in light of the standards currently used in the circuits and those that have been recommended in the scholarly literature. Her proposed standard is among the best, if not the best, of the proposals that have been formulated since the *DeShaney* decision to clarify and determine the affirmative constitutional rights of institutionalized individuals.

13. *Infra Part III.*
14. Except for sexual offenders and people in the criminal justice system, the Supreme Court has not had a single case on the constitutional rights of people who are civilly committed or voluntary patients in institutions since *Heller v. Doe*, 509 U.S. 312 (1993), and *Zinermon v. Burch*, 494 U.S. 113 (1990), and both were cases involving procedural due process rather than substantive due process. The Supreme Court has not had a substantive due process case involving the rights of institutionalized persons since *Youngberg v. Romeo*, 457 U.S. 307 (1982).
15. *Infra Part IV.*
II. DeShaney and the Real World of Institutionalized Psychiatric Patients

First, like Justice Blackmun, I will speak from the heart. I spent almost thirty years as a public interest lawyer representing people with psychiatric disabilities, many of whom were institutionalized. I perennially negotiated the tension between understanding that courts must craft workable rules providing predictability and guidance in myriad situations, and frustration at the artificiality and inadequacy of those rules when applied to the people I represented and the situations in which they found themselves.

Public interest lawyers serve as the intermediaries between their clients’ often horrific experiences of injustice and the judges in whose worlds those experiences are impossible to imagine. So we take true stories from a world that is fundamentally alien to a skeptical judiciary, and repackage them as legal claims, omitting the nuances to create a narrative that is meaningful in the judges’ language and world, all the while trying to be as faithful as possible to our clients’ original experiences. In doing this, we must accept and work with legal doctrines based on assumptions that fly in the face of what we know to be true.

Nowhere was this cognitive dissonance more stark to me than in the doctrinal abyss that some courts have interpreted DeShaney v. Winnebago County to require, sharply dividing the residents of institutions. Under these decisions, involuntarily

16. See, e.g., Shelton v. Ark. Dep’t of Human Servs., 677 F.3d 837, 840 (8th Cir. 2012) (stating that, under DeShaney, a patient’s status as voluntary or involuntary determines whether affirmative duties are owed, though recognizing that voluntary status may change during admission); Torisky v. Schweiker, 446 F.3d 438, 444–46 (3d Cir. 2006) (recognizing that under DeShaney, states generally do not owe patients affirmative duties, but clarifying that DeShaney does not preclude affirmative duties when voluntary patients are not free to leave); Walton v. Alexander, 44 F.3d 1297, 1304 (5th Cir. 1995) (applying DeShaney strictly and concluding that states do not owe patients affirmative duties unless “the person is involuntarily taken into state custody and held against his will through the affirmative power of the state”); Monahan v. Dorchester Counseling Ctr., 961 F.2d 987, 991 (1st Cir. 1992) (applying DeShaney to mean that a state does not owe affirmative duties unless it restrains a patient’s liberty by involuntarily committing him); Higgs v. Latham, 946 F.2d 895 (table), No. 91-5273, 1991 WL 216464, at *4 (6th Cir. Oct. 24, 1991) (per curiam) (applying DeShaney as precluding affirmative duties to
committed individuals have affirmative constitutional rights to safety, freedom from unreasonable bodily restraint, and minimally adequate treatment necessary to realize those rights, while legally voluntary patients have no affirmative constitutional rights at all. The rationale for the distinction appears to be the assumption that legally voluntary patients can leave the institution if they are dissatisfied.

I have spent a considerable amount of time inside state institutions in fourteen states and the District of Columbia, and in my experience, there is no difference between the loss of liberty suffered by civilly committed and voluntary patients in state institutions. Nor is there any difference in the control exercised by those institutions over the lives of civilly committed and voluntary patients. Voluntary patients are detained, restrained, contained, secluded, locked up, assaulted, and denied ground “privileges” and visiting passes in exactly the same way as civilly committed patients. Their lives are equally subject to control from the moment they arise in the morning at a preordained time, through meals and cigarette breaks and “groups,” to who is permitted to visit and when, to the long stretches of empty hours in the dayhall.

voluntary patients). But see Walton v. Alexander, 44 F.3d 1297, 1306 (5th Cir. 1995) (Parker, J., concurring) (“The Court’s holding [that a bright line separates voluntary and involuntary patients] is based on an erroneous reading of the Supreme Court’s guidance in DeShaney, and draws an arbitrary, illogical, and formalistic line between persons who are entitled to constitutional protection.”); Higgs v. Latham, 946 F.2d 895 (table), No. 91-5273, 1991 WL 216464, at *6 (6th Cir. Oct. 24, 1991) (Suhrheinrich, J., concurring) (“I do not read DeShaney to control the outcome of this case... [E]ven DeShaney compels us to go beyond asking whether [the patient] was a voluntary admittee.” (emphasis added)).

17. See cases cited supra note 16 (consistently recognizing that involuntary patients are owed the affirmative duties of care and protection recognized in Youngberg v. Romeo); Youngberg v. Romeo, 457 U.S. 307, 324 (1982) (establishing that the affirmative duties applicable to institutionalized patients include “conditions of reasonable care and safety, reasonably nonrestrictive confinement conditions, and such training as may be required by these interests”).

18. See, e.g., Monahan v. Dorchester, 961 F.2d 987, 992 (1st Cir. 1992) (“Monahan’s complaint did not allege that he would have been barred from leaving [the state facility] upon request.”).

19. Like Ms. Hagan, I am excluding the ever-expanding number of forensic patients in state psychiatric facilities from my comments.
Voluntary patients can't leave, and they know they can't leave. Or at least they know they can't leave once they ask to leave. If they ask verbally, they are often told they cannot be released. Sometimes they are given papers to fill out officially requesting release. If they file formal notifications of their intention to leave, they are either “persuaded” to withdraw those notifications or involuntary commitment proceedings are filed against them.

The fact that state commitment statutes operate differently in different states may inform the formal legal analysis, e.g., whether staff is legally obligated to offer patients the opportunity to change their status to voluntary. But on the ground, the picture always looks the same: however a patient gets to a state mental facility, that patient cannot leave until either the institution or a court decides it’s time to go.

In addition to being places of total constraint on liberty, which patients cannot leave at will, many institutions are places where use of force and threat of force is rampant and underreported, and where acts of violence that would be crimes in the outside world are covered up or characterized as “patient abuse.” Many of the female patients have histories of childhood sexual abuse and trauma that make institutional practices like being put in restraints by male aides particularly excruciating. Patients with histories of sexual abuse are asked to endure pat

downs, strip searches, and body cavity searches.\textsuperscript{21} Even bed checks, a practice involving staff coming into patients’ rooms at night and shining flashlights on their beds to ensure that they are there, can be difficult to endure.

Sexual activity in institutional settings is more common than outsiders might imagine, and runs that gamut from mutual and supportive relationships between patients through exploitation, coercion, and rape by other patients and staff.\textsuperscript{22}

\begin{itemize}
\item \textsuperscript{21} See Serna v. Goodno, 567 F.3d 944, 947 (8th Cir. 2008), \textit{cert. denied} 130 S. Ct. 465 (2009) (noting that 150 patients institutionalized in Minnesota’s Sex Offender Program were subjected to visual body-cavity inspections after a contraband cellphone was found); Aiken v. Nixon, 236 F. Supp. 2d 211, 218, 236 (N.D.N.Y. 2002) (noting that a voluntary psychiatric patient was subjected to a strip search upon admission pursuant to a “standing order,” rather than to a reasonable belief that the patient possessed drugs or contraband); Anne Donahue, \textit{Strip Searches: Going Too Far for Safety Needs?}, COUNTERPOINT, July 2009, at 1 (describing policies implemented by some Vermont hospitals to perform mandatory full body searches on patients being admitted to psychiatric units).
\item \textsuperscript{22} See Ammons v. State Dep’t of Soc. & Health Servs., 648 F.3d 1020, 1032 (9th Cir. 2011), \textit{cert. denied} 132 S. Ct. 2379 (2012) (holding that a teenage patient’s sexual relationship with a staff member may lead to the facility director’s constitutional liability for failure to protect); Beck v. Wilson, 377 F.3d 884, 886–87, 892 (8th Cir. 2004) (concluding that qualified immunity protected state officials in a suit brought by an involuntarily committed raped woman who was the only woman on a ward full of men and had reported being afraid she would be sexually assaulted); Neely v. Feinstein, 50 F.3d 1502, 1505 (9th Cir. 1995) (denying qualified immunity for failure to supervise a staff person who repeatedly sexually assaulted patients); Walton v. Alexander, 20 F.3d 1350, 1353 (5th Cir. 1995) (involving a student at a state-operated residential school who was sexually assaulted by another student on multiple occasions); Rodgers v. Horsley, 39 F.3d 308, 311 (11th Cir. 1994) (per curiam) (concluding that an involuntarily committed pregnant woman raped while on observation cannot recover because of insufficient evidence of past rapes at facility, even though the state was aware of sexual activity at facility); Higgs v. Latham, 946 F.2d 885 (table), No. 91-5273, 1991 WL 216464, at *1–2 (6th Cir. Oct. 24, 1991) (per curiam) (involving a claim by a psychiatric patient who was sexually assaulted by another patient); Davis v. Holly, 835 F.2d 1175, 1177 (6th Cir. 1987) (involving a claim brought by mental patient who had sex with a staff member and gave birth to a child); Elizabeth M. v. Ross, No. 8:02CV5585, 2005 U.S. Dist LEXIS 45107, at *4, *7–8 (D. Neb. May 11, 2005) (certifying a class of women subject to rape, sexual assault, sexual harassment, sexual exploitation, and physical assault at three state institutions), \textit{vacated by} Elizabeth M. v. Montenez, 458 F.3d 779, 787–88 (8th Cir. 2006) (finding a lack of typicality in that some women alleged being raped and sexually assaulted by staff and others by other patients, that most of the women were from one institution, and that most had already been discharged); Caroline C. v. Johnson, 174 F.R.D. 452, 455–57 (D. Neb. 1996) (certifying a class of raped and sexually assaulted women
\end{itemize}
The use of restraints, which varies widely from facility to facility, can result in death or serious injury. In some facilities, patients are tied to beds with “five-point restraints” (arms, legs, and across the chest). In other facilities, they are tied to chairs. Due to a tremendous effort by the Joint Commission, the National Association of State Mental Health Program Directors, and extraordinary dedication in states like Pennsylvania and Massachusetts, some state facilities are virtually restraint-free. But in others, restraint remains commonly used, including daily use on some patients.

at institution); see also Winiviere Sy, The Right of Institutionalized Disabled Patients to Engage in Consensual Sexual Activity, 23 WHITTIER L. REV. 545, 546–48 (2001) (discussing policies regarding consensual sexual activities at various California hospitals).


In sum, Ms. Hagan is correct that institutionalized people are vulnerable and in need of protection; they are also isolated from the oversight of the larger community and the norms of community culture. Institutions develop cultures of their own. Sometimes these are healing; more often they are not. The move to true community integration has begun, but in the meantime, those who are left behind in institutional settings deserve better than to be told that the state is not responsible for their safety because they chose to be in an environment that, in all truth, they cannot escape.

III. Chaos in the Circuits: The Interpretation of DeShaney’s Application to Institutionalized Patients

Ms. Hagan’s Note accurately summarizes, for the most part, the differences in approaches taken by the circuits in applying DeShaney to the affirmative constitutional rights of individuals in state institutions. If anything, she is too tactful to assert just how irrational and counterintuitive these results have been, including conflicting results within the same circuits. This is in part because the DeShaney Court was resolving a factual and legal situation far different from the situation of individuals in state institutions. The Court did not specifically address the nuances of voluntary and involuntary institutionalization because those issues were not before it. And yet courts in the various circuits confidently cite DeShaney as supporting their divergent approaches to affirmative constitutional rights within institutional settings.

28. Hagan, supra note 2, at Part IV.
30. See supra note 16 (collecting cases that apply DeShaney to cases involving state hospital patients).
A. The DeShaney Decision Is Not Concerned with Institutionalized Individuals

Although the circuit courts have taken *DeShaney* as a touchstone for determining the affirmative constitutional rights of institutionalized individuals, their interpretations of its direction have diverged considerably. This is not surprising: the Court’s comments on institutionalized individuals are delphic and ambiguous because the facts in *DeShaney* did not raise any of the issues involved in institutionalization.

Joshua DeShaney lived at home in the community. He was in the custody of a hospital for less than a week because his injuries prompted review of his father’s parental rights. The state’s interactions with Joshua amounted to a visit, once a month, by a social worker. It is true that during each of these monthly visits, the social worker could see much that should have been greatly troubling, and that in the last two visits, she did not see Joshua at all. Joshua’s father beat him over most of his life, and continued to beat him without state intervention, until Joshua suffered severe and irreversible brain damage.

As horrific as the circumstances of the *DeShaney* case are, they are not analogous in terms of state responsibility and control to the situation of individuals residing in a facility operated by the state, staffed entirely by state employees, whose actions, schedules, and activities are under supervision twenty-four hours a day, who cannot leave the grounds of the facility without a pass, much less decide to leave the facility itself. The community has little or no interaction, oversight, or power to intervene at state hospitals; family visits are often regulated to a few hours a week and in some cases prohibited altogether. An institutionalized

31. *Supra* note 16.
33. *Id.* at 192.
34. *Id.* at 192–93.
35. *Id.* at 193.
36. *Id.* at 192–93.
37. See, e.g., Doe v. Pub. Health Trust, 696 F.2d 901, 905–06 (11th Cir. 1983) (Hatchett, J., concurring) (describing a hospital’s policy of restricting adolescent patients’ ability to communicate with their parents as part of a
individual is entirely in the hands of the state and at its mercy, a situation that could not be more different than that of Joshua DeShaney.

Therefore, it is not surprising that, in rejecting claims of state responsibility for Joshua DeShaney, the Supreme Court was not completely clear about the boundaries of those responsibilities. As Ms. Hagan points out, the Court variously states that affirmative constitutional rights are afforded to “involuntarily committed mental patients,”38 to “a person [taken] into custody and [held] there against his will,”39 and to a person who “is institutionalized and wholly dependent on the State.”40

The majority is clear about one thing: the basis of any state affirmative responsibility must lie in affirmative state action.41 Neither an individual’s vulnerability to harm nor the state’s knowledge of both that vulnerability and the risk of private violence to the individual suffice to create affirmative responsibility on the part of the state.42 Rather, the state must clearly act to restrain the individual’s liberty.43 Ms. Hagan argues, and I agree, that the Supreme Court never intended to limit the evaluation of state action to restrain an individual’s liberty to an individual’s legal status at admission to a state institution.44 Rather, the analysis requires an examination of the degree to which the state has acted to replace an individual’s

39. Id. at 199–200.
40. Id. at 200.
41. Id.
42. Id. 200–02.
43. The Court reiterates that the focus must be on the state’s action: when the state “by an affirmative exercise of its power so restrains an individual’s liberty that it renders him unable to care for himself” the affirmative duties are created. Id. A few sentences later, the Court underscores that “it is the State’s affirmative act of restraining the individual’s freedom to act on his own behalf—through incarceration, institutionalization, or other similar restraint of personal liberty—which is the ‘deprivation of liberty’ triggering the protections of the Due Process Clause, not its failure to act to protect his liberty interests against harms inflicted by other means.” Id.
44. Hagan, supra note 2, at 786–87.
ordinary control of his or her own life with its own control, regardless of the individual’s own wishes and desires. Ms. Hagan’s question to judge the state’s control over an individual—can the individual leave state custody, or be taken from state custody by caring family and friends, or would the state prevent or prohibit those actions?—is the right one, with the right results.

B. The Patient’s Legal Status on Admission as Voluntary or Involuntary as Determinative of Constitutional Rights

As Ms. Hagan notes, some circuits have chosen to rely primarily on a patient’s formal legal status at the time of admission to the institution to determine whether the individual is entitled to assert constitutional rights to safety and bodily security, although even these circuits recognize the state-created danger, “special relationship,” and other exceptions, and many of the circuits also have intracircuit conflicts.

Those circuits that rely primarily on the legal status of the patient may initially appear to have chosen a framework that at least allows for clarity and certainty. But even if that was the case, the test is not rational and leads to unjust results.

It is not rational for a number of reasons. First, it bears no relationship to the Supreme Court’s underlying explication of its own standard in DeShaney. According to the majority, the reason that the state owes affirmative obligations when it deprives an individual of liberty against his will and places the individual in an institution is that the state has deprived the individual of the ability to provide for his or her own needs. This, as both Ms. Hagan and Justice Brennan’s DeShaney dissent underscore, is a

45. Id. at Part VI.
46. Id. at Part VI.A.
47. It actually is not so simple. See infra notes 90–101 and accompanying text.
48. See DeShaney v. Winnebago Cnty. Dep’t of Soc. Servs., 489 U.S. 189, 200 (1989) (“[W]hen the State by the affirmative exercise of its power so restrains an individual’s liberty that it renders him unable to care for himself, and at the same time fails to provide for his basic human needs . . . it transgresses the substantive limits on state action.”).
whopping legal fiction. Nicholas Romeo, the plaintiff in *Youngberg v. Romeo*, was described as being “profoundly [mentally] retarded” and was no more able to fend for himself without help than Joshua DeShaney, regardless of his legal status.

Second, Ms. Hagan points out that some voluntary patients are not actually competent to provide informed consent to hospitalization. In fact, she notes that the Supreme Court itself has held that states should foresee the possibility that voluntary patients may be incompetent to execute valid consents to hospital admission and treatment. Circuits have similarly considered cases in which the patients appear incompetent to consent to voluntary admission, especially in the area of people institutionalized with developmental disabilities. There is a substantial research literature going back over thirty years, but increased after the Supreme Court’s decision in *Zinermon v. Burch*, underscoring the alarming percentage of voluntary

---

49. *Id.* at 206 (Brennan, J., dissenting).
51. *Hagan, supra* note 2, at Part V.A.
53. In addition to the material cited by Ms. Hagan, *supra* note 2, at Part V.A, see *Wilson v. Formigoni*, 42 F.3d 1060, 1062–63 (7th Cir. 1994) (reviewing hospital documentation, where staff persuaded patient to sign in as voluntary even though her intake assessment noted that she “lacked decision making abilities, as well as appearing to be delusional” and “was only partially oriented” and “her thought process was fragmented”); *Rennie v. Davis*, 997 F. Supp. 137, 138–39 (D. Mass. 1998), aff’d sub. nom. *Davis v. Rennie*, 264 F.3d 86 (1st Cir. 2001), cert. denied 535 U.S. 1053 (2002) (finding that the patient’s medical record evidenced that he was mentally incapacitated at the time hospital staff allowed him to sign a voluntary admission form); *Butler v. Comm’r of Mental Health*, 463 F. Supp. 806, 807 (E.D. Tenn. 1978) (describing plaintiff’s allegations that she was taken to a hospital against her will and never knowingly signed any voluntary admission forms although the hospital had her listed as a voluntary patient).
54. *See, e.g.*, *United States v. Tennessee*, 615 F.3d 646, 648, 650 (6th Cir. 2010) (describing Tennessee laws allowing parental consent to substitute for a patient’s consent to voluntary treatment for institution serving patients with mental retardation); *Doe v. Austin*, 848 F.2d 1386, 1388–89 (6th Cir. 1988) (describing Kentucky’s comprehensive statutory regime designed to safeguard patients with mental retardation from involuntary commitment, but that parental consent was used to obtain voluntary admission for “virtually all” patients).
patients who have little or no understanding of the documents they have signed.\footnote{55} Some state statutes explicitly contemplate noncompetent “voluntary” commitments.\footnote{56}

A related, but distinct, issue is whether even competent patients exercise informed consent when they agree to voluntary hospitalization. An individual may be competent but not have sufficient information to make a decision. For example, Massachusetts law requires that a patient seeking voluntary status be given the opportunity to consult an attorney on the benefits and drawbacks of such a decision,\footnote{57} while Wisconsin requires that a guardian ad litem visit some voluntary patients within the first three days of admission.\footnote{58} In New Jersey, concerns about competence in the context of voluntary admissions led to a statutory requirement that people who want to be voluntary patients must have hearings, with legal counsel, to ensure their decisions are competently made.\footnote{59}


\footnotetext{56}{Wisconsin, for example, has a category of commitment called “non-protesting voluntary admission” that “usually... involves a catatonic or paranoid person who is unwilling to sign an admission form but who does not object to the admission[, or] an elderly confused person with mental illness who needs to be transferred from a nursing home to a psychiatric unit.” Diane Greenley, \textit{Civil Commitment and Voluntary Treatment}, in \textit{RIGHTS & REALITIES II} 352 (Wis. Disability Rights ed. 2008), http://www.disabilityrightswi.org/wp-content/uploads/2008/09/civil-commitment-voluntary-treatment.PDF (citing Wis. Stat. § 51.10(4m)(b) (2012) (providing that a non-protesting “voluntary” admission is only good for seven days)).}


\footnotetext{58}{Wis. Stat. 51.10(4m)(c) (2012) (requiring that a guardian ad litem visit any patient admitted as a non-protesting voluntary patient within seventy-two hours).}

\footnotetext{59}{N.J. Ct. R. 4:74-7 (2012); see also Matter of Commitment of G.M., 526 A.2d 744, 745 (Ch. Div. 1987) (“What good is all our effort in preventing patients from being ‘lost in the cracks’ if such a hapless category as the so-called ‘volunteers’ is subject to being lost?”).}
Ms. Hagan also accurately describes the coercion that many hospitals employ to pressure patients into voluntary status. In many states, patients have a statutory right to elect voluntary hospitalization rather than be involuntarily committed; in discussing this right, courts have continually noted that although technically voluntary, the patients can be prevented from leaving if they give any indication that they want to be discharged. Doctors also sometimes improperly condition granting voluntary status on patient concessions such as agreeing to take medications.

Fourth, a substantial number of patients are voluntary only as a legal technicality, including minors, individuals under guardianship, and individuals in the custody of state agencies, who can be voluntarily admitted by their parents or guardians, or even their health care proxies. This was the case in Kennedy v. Schafer, described at length by Ms. Hagan, and in numerous other cases.

However, the most important reason that the voluntary–involuntary distinction has no traction is that, both as a matter of fact and as a matter of law in every state, admission status makes no difference in a patient’s liberty and ability to leave once institutionalized. Even if they are completely able to provide for

---

60. See Hagan, supra note 2, at Part V.B (discussing coercive forces in inpatient psychiatric care).
their own needs, so-called voluntary patients are no more free to simply walk out the door and provide for those needs than involuntarily committed patients. A district court in Illinois cogently explained the distinction between voluntary legal status and the kind of liberty that the Supreme Court incorrectly assumed that legal status conferred:

When an individual signs into a mental health facility run by the state, that patient may be consenting to treatment, but the patient is also being committed under the control of the state. The state may restrict the patient’s freedom to act, although the patient volunteered for this treatment. The nature of the discretion afforded state officials may be extensive because of the patient’s voluntary act, but the court is not convinced that the discretion of state officials is wide enough to sanction the deliberate indifference to the patient’s medical needs, or the patient’s right to safe conditions, while the patient is incapacitated or restrained in the mental health facility.67

In all states, voluntary patients who decide to leave are subject to involuntary holds while institutional personnel attempt to persuade the patient to change his or her mind, and failing that, to decide whether to file for involuntary commitment. In virtually all cases where a voluntary patient files a notice of intention to leave, the hospital pressures the patient to reconsider, and as caselaw confirms, the hospital is frequently successful.68 If a patient continues to insist on leaving, the hospital counters by filing a petition for involuntary commitment. This is a reality of which patients are well aware. The DeShaney Court’s rationale that voluntary patients freely choose to avail themselves of the state’s services, and can choose to leave at any time, is fundamentally undermined by state laws permitting the involuntary detention of any patient—regardless of legal status—who expresses a desire to leave, until an involuntary commitment

68. Wilson v. Formigoni, 42 F.3d 1060 (7th Cir. 1994); Higgs v. Latham, 946 F.2d 895 (table case), No. 91-5273, 1991 WL 216464, at *6 n.1 (6th Cir. Oct. 24, 1991) (Suhrheinrich, J., concurring) (stating that a voluntary patient repeatedly requested to be discharged and was refused); Rennie v. Davis, 997 F. Supp. 137, 139 (D. Mass. 1998) (describing the numerous times the patient signed a form indicating his desire to leave and was pressured to withdraw it by staff).
petition can be filed. As the Sixth Circuit pointed out in *Lanman v. Hinson*:

69

In the present case even though Lanman was technically voluntarily committed, under Michigan law, once he gave the hospital notice of his intent to leave, the hospital could retain him against his will for up to three days. . . . [A]pplying the district court’s reasoning, if Lanman had decided to leave the hospital, and been retained involuntarily under § 330.1419(1), any § 1983 claims arising in those three days of involuntary confinement would fall under the Fourteenth Amendment. But immediately prior to his decision to leave, while his confinement was technically voluntary, the Fourth Amendment would apply to any § 1983 claims. Under such a system, while Lanman’s relationship and dependence on the state would not have changed, his constitutional protection would have. We do not believe such a distinction is warranted. 70

*Lanman* is typical of a number of cases in which the court puzzles over the constitutional significance of the legal status of “voluntary” applied to a patient whose injury occurs while actually physically restrained or locked in a seclusion room. 71 In fact, the majority of cases involving patients who are voluntary and supposedly free to leave involve plaintiffs who are secluded in locked rooms, injured or killed in restraints, under continual or virtually continual observation, raped or sexually assaulted, or some combination of these.

The district court in *Estate of Cassara* aptly noted that a patient in seclusion or restraints, even one whose legal status is that of a “voluntary” patient, is not actually free to leave:

Moreover, that Cassara was placed by the Mental Health Center’s staff in a room that is used for restraining patients under guidelines established by state statute suggests a state restriction on liberty in a custodial, institutional setting. Although not specifically alleged, it is a reasonable inference to draw that Cassara may not have been free to leave the

70. Id. at 683 (citations omitted).
facility immediately if he had requested to do so. The court finds little practical difference between a locked jail cell and a locked restraint room under the circumstances alleged here. Although Cassara was purportedly free to leave at any time—even while in the restraint/seclusion room—it is not clear that he could merely walk out of the Mental Health Center without the state affirmatively consenting to or allowing his egress. Cassara may have had access figuratively to the key to his cell. The right to leave, however, does not guaranty the power to leave.73

In addition to being unwarranted and leading to irrational results, the distinction between formal involuntary and voluntary commitment status does not even confer the virtue of clarity, as demonstrated by a number of cases.

Ms. Hagan discusses the Sixth Circuit struggling with real world circumstances in Higgs v. Latham.74 In that case, Ms. Higgs was involuntarily detained by court order, but due to clerical errors and omissions related to her transfer, the staff at Western State Hospital thought she was a voluntary patient.75 Despite this, staff repeatedly refused her requests to leave.76 The Sixth Circuit concluded that because Western State did not know of the court’s order to involuntarily detain her, she had no constitutional right to security and safety, and therefore her constitutional rights were not violated when she was raped by another patient.77 It is clear from the Sixth Circuit’s decision that if Western State had understood she was an involuntary patient, she would have had a constitutional right to be protected from rape at the State Hospital.78 And yet Western State would not

73. Id. at 279.
74. Higgs v. Latham, 946 F.2d 895 (table), No. 91-5273, 1991 WL 216464 (6th Cir. Oct. 24, 1991) (per curiam). It should be noted that at the time Higgs was decided, it was denominated as a decision which should not be given precedential value. The Sixth Circuit discussed this issue later in United States v. Tennessee, 615 F.3d 646, 653–54 (6th Cir. 2010), noting that, as an unpublished decision, Higgs did not bind the court in later decisions.
76. Id. at *6 n.1 (Suhrheinrich, J., concurring).
77. Id. at *5 (majority).
78. See id. at *4 (“[T]he hospital was unaware of any commitment order and had been led by a telephone referral to believe that Mrs. Higgs’ admission would be a voluntary one.”).
have treated her differently in any way whether she was legally involuntary or voluntary; she was no more free to leave in either case.

In another case, *Davis v. Rennie*, the court was faced with a plaintiff who had signed an application for voluntary admission. However, the top of the form contained the typed notation: “the patient refused to sign a voluntary saying, ‘I don’t want to be here for a year.’” In addition, as plaintiff pointed out, he had been transferred to the state hospital from a private hospital because he was acutely psychotic, so much so that the transferring psychiatrist testified he could not possibly have been competent to sign a voluntary admission on the day he arrived. The state pointed out that after signing the voluntary admission, Mr. Davis had also executed four forms indicating his intention to leave the hospital, and then retracted them, arguing that those repeated retractions clearly indicated that he was actually a voluntary patient. The court rejected this argument as well because Mr. Davis had not been provided with statutorily mandated protections for individuals retracting their intent to leave the hospital.

Yet in a case decided earlier by the highest state court in Massachusetts, the estate of a patient who had been involuntarily committed to a state facility by a court, and then apparently converted to voluntary status, was found to have no affirmative federal constitutional rights to safety or protection because, even though she had been involuntarily committed, she had been converted to voluntary status prior to her death. No inquiry was made regarding the competence of her decision or whether she

---

80. *Id.* at 138.
81. *Id.*
82. *Id.*
83. *Id.* at 139.
84. *Id.*
85. See *Williams v. Hartman*, 597 N.E.2d 1024, 1028 (Mass. 1992) (“Because the decedent was committed voluntarily to Fuller, she did not possess the Federal constitutional rights that the defendant allegedly violated.”).
was afforded the rights guaranteed to voluntary patients under any Massachusetts statute.86

As is underscored by all of these decisions, a patient’s clinical and legal status can change a number of times during a single hospitalization, from voluntary to involuntary and back to voluntary, just as a patient can be incompetent on admission and regain competence during the admission.87

None of this, however, has any impact on how a patient is treated, the liberty he or she has while in the hospital, or the patient’s freedom to leave. As Ms. Hagan notes, in Higgs v. Latham the hospital’s misunderstanding about the nature of a patient’s legal status made absolutely no difference in the patient’s actual liberty or control over her circumstances.88 Although the hospital thought Ms. Higgs was voluntary, she was still repeatedly told that she couldn’t leave when she asked to be released.89 Yet in Higgs v. Latham, as in numerous other cases, Ms. Higgs was denied the constitutional rights that she would have had if the hospital had known that she was involuntarily committed. Civilly committed patients have a panoply of rights (safety, freedom from unreasonable bodily restraint, and minimally adequate treatment) that are substantially different from the constitutional rights of so-called voluntary patients (nothing).

Well, not exactly nothing. As Ms. Hagan points out, there are numerous divisions and splits both among circuits and arguably within the same circuit, as courts struggle to interpret DeShaney and to create a framework that is both clear and rational to explain the degree of state involvement necessary for individual injuries to trigger constitutional remedies.

86. See id. at 1027–28 (containing the court’s analysis); id. at 1028–29 (O’Connor, J., concurring in part and dissenting in part) (arguing that the question of whether the decedent was a voluntary patient should be remanded, because “[m]any critical factual questions remain unanswered”).


88. Hagan, supra note 2, at 782.

C. Functional Custody: An Alternative to Formal Legal Status in Determining the Affirmative Constitutional Rights of Institutionalized Patients

The Sixth Circuit provides a particularly good illustration of how courts struggle to fit the round peg of reality into the square hole of formal doctrine. Higgs v. Latham, and several previous district court cases involving individuals with intellectual disabilities\(^\text{90}\) strictly followed the voluntary/involuntary distinction to preclude individual claims of damages for injury and death by plaintiffs who were voluntarily admitted to state institutions. Yet both before and after these cases, the Sixth Circuit found—in Doe v. Austin\(^\text{91}\) and in United States v. Tennessee\(^\text{92}\)—that class actions involving individuals with intellectual disabilities voluntarily committed by their guardians did state claims for constitutional violations.\(^\text{93}\)

In both class actions, the Sixth Circuit followed a rationale that would have applied equally well in the individual-patient cases: it looked beyond formal legal status, concluding that severely disabled individuals “voluntarily” hospitalized by their parents or guardians were not, in fact, as a practical matter free to leave, both because of their own circumstances and the requirements of state law, and that they should therefore be considered involuntary for purposes of their affirmative constitutional rights.\(^\text{94}\) This is a classic example of the “functional custody” analysis advocated by Ms. Hagan.\(^\text{95}\)


\(^{91}\) Doe v. Austin, 848 F.2d 1386 (6th Cir. 1988).

\(^{92}\) United States v. Tennessee, 615 F.3d 646 (6th Cir. 2010).

\(^{93}\) Austin, 848 F.2d at 1395; Tennessee, 615 F.3d at 655.

\(^{94}\) See Austin, 848 F.2d at 1388 (“[T]he commitment of mentally retarded adults by the Commonwealth upon application by a parent or guardian is to be considered voluntary.”); Tennessee, 615 F.3d at 651–52 (affirming that the circuit’s law had not changed from a prior ruling in this class action litigation, where the circuit held that voluntary and involuntary patients in a facility treating developmental disabilities enjoy identical rights (citing United States v. Tennessee, 798 F. Supp. 483, 486 (W.D. Tenn. 1992))).

\(^{95}\) Hagan, supra note 2, at Part VI.A.
The only distinction among these cases appears to be procedural: individual claims versus class actions. However, in another class action, after *Doe* but prior to *Tennessee*, another district court felt bound by *Higgs v. Latham* to exclude constitutional claims of institutionalized individuals with mental retardation and psychiatric diagnoses who were voluntarily placed in the institution by their guardians.\(^{96}\)

The First Circuit provides another example of confusion within the circuits regarding the boundaries of functional custody. Ms. Hagan describes *Monahan v. Dorchester*,\(^{97}\) presaged by an even stricter reliance on the voluntary–involuntary distinction by the state’s highest court.\(^{98}\) Yet the First Circuit later affirmed a substantial damage judgment to a patient, admitted pursuant to a voluntary admission form, against staff members who had beaten him after he was restrained, and against their supervisor who did nothing to stop them.\(^{99}\) Even though the patient was technically voluntary, the district court had refused to dismiss the case because the patient had not been afforded the rights provided under a Massachusetts statute, which allows voluntary patients to consult with an attorney prior to withdrawing a petition to leave the facility.\(^{100}\) This finding was not pursued on appeal.\(^{101}\)

---

98. Williams v. Hartman, 413 Mass. 398, 403 (1992) (“[A] patient who is voluntarily committed to a State mental health facility does not possess the same Federal constitutional rights as an involuntarily committed patient.”).
99. See Davis v. Rennie, 264 F.3d 86, 91 (1st Cir. 2001) (affirming a judgment awarding $100,000 in compensatory damages plus $1,550,000 in punitive damages), *cert. denied* 535 U.S. 1053 (2002).
100. *Id.* at 92–96.
101. *Id.* at 91 (stating the grounds for appeal).
D. State-Created Harm and State Increase of Danger in the Context of Institutionalized Patients

1. Actions by State Actors versus Private Individuals

One crucial limitation of *DeShaney* is that it applies only to the state’s responsibility to protect a citizen from the violence or harm of a private citizen. State actors who directly harm a citizen (as opposed to failing to protect him or her) or deprive the citizen of liberty are not shielded by *DeShaney*: *DeShaney* is simply not relevant to cases claiming violence or deprivation of liberty by state actors. The Constitution’s “charter of negative liberties” is explicitly aimed at protecting citizens from arbitrary denials of life, liberty, and property at the hands of state agents.102

This has been a key distinction in many cases. For example, the panel in *Lanman v. Hinson* found that the central distinction with *Higgs* was that in the latter case, Ms. Higgs was raped by another patient—a third party—while in *Lanman*, state agents were charged with having killed Mr. Lanman during a restraint.103 This distinction—whether the party causing the harm is a state agent or a private party—has been cited by other courts to find that plaintiffs have stated a constitutional claim.104

The distinction is, in fact, useful in cases like *DeShaney* or in cases involving school children that take place in the free world. The state should be responsible for the constitutional violations of its own actors, regardless of the legal status of the institutionalized patient or even whether an individual is institutionalized at all. It is state responsibility for protecting against private violence that must be limited in some way. The Supreme Court has chosen to limit it by confining liability to

102. See *DeShaney v. Winnebago Cnty. Dep’t of Soc. Servs.*, 489 U.S. 189, 196 (1989) (stating that the purpose of the Due Process Clause “was to protect the people from the State, not to ensure that the State protected them from each other”).


cases in which the state created or exacerbated the danger, or has limited or restrained the liberty of the individual in ways that preclude the individual or others from protecting himself or herself.105

Both of these exceptions are at work in the closed setting of an institution. The distinction between private parties and state agents, so useful in the world at large, breaks down when the boundaries of the patient’s world—a world from which no exit is possible without state consent—are completely within state-owned, state-staffed, state-supervised, and state-operated property. Unlike the outside world, other patients are not unconstrained, unsupervised free agents. Other patients are there precisely because a mental health professional, and often more than one, has made a judgment that they are not safe to live unsupervised and free lives. Other patients are there to be assessed, receive treatment, and supervision. Staff is generally well aware of the proclivities and risks posed by various patients: that is the purpose of confinement in a state institution. In such close quarters, where neither voluntary nor involuntary patients are in a position to escape the aggression of others, the responsibility of the state to protect may depend, as it does in prison settings,106 on the degree to which the staff is aware of the inclinations and characteristics of particular patients. If a particularly vulnerable patient is deliberately or recklessly placed in harm’s way, the state may be seen to have created or increased the danger that ultimately befalls the patient.

2. State-Created or -Exacerbated Harm in the Context of Suicidal Patients

A number of cases that raise questions about the state’s affirmative duties to protect individuals arise in the context of suicide, either in the community after or during a police

---

105. *DeShaney*, 489 U.S. at 201 & 201 n.9.
106. See *Farmer v. Brennan*, 511 U.S. 825, 837 (1994) (defining the standard for holding prison officials liable for private harm as requiring that “the official must both be aware of facts from which the inference could be drawn that a substantial risk of serious harm exists, and he must also draw the inference”).
encounter, or in institutional settings. The police cases usually raise the question of whether police intervention escalated or increased the danger of suicide; courts’ analyses of these cases—which tend to be sympathetic to police actions under circumstances perceived to be both exigent and complex—often focus heavily on the fact that the individual was not in state custody.107

When the individual is in state custody, the duty to protect raises different issues. When Deborah Shelton was cut down after being found hanging, she was still alive, but state hospital nurses refused to perform mouth to mouth resuscitation without protective masks, which were unavailable at the time, a refusal which was permitted by hospital policy.108 The state did not have to respond to this issue, because the Eighth Circuit decided that Deborah Shelton, as a “voluntary” patient, had no rights.109 Of course, Deborah Shelton was no more free to leave the hospital than any involuntarily committed patient.

Although I am a strong advocate of limiting the liability of mental health professionals for the suicide of their patients,110 like many courts, I draw a distinction between institutionalized patients and patients who live in the community.111 Even in institutions, the legal issues around suicide revolve not so much around the difference between voluntary and involuntary patients as around the difficulty in determining which patients are truly at risk of suicide, given the small proportion of people with suicidal ideation who actually commit suicide, and the drawbacks of the preventive methods that are implemented to

---

109. Id. at 842.
prevent suicide or self-harm, such as involuntary commitment and physical restraints.\footnote{112}

\textbf{IV. Ms. Hagan’s Proposed Standard of Analysis}

Ms. Hagan ends her Note with a proposed standard analyzing whether an institutionalized person has stated a constitutional claim that is clear and rational. She proposes that courts examine a claim in three steps that parallel the language of \textit{DeShaney} itself:

1. Is the person involuntarily in the custody of the state as a formal legal matter?
2. Is the person involuntarily in the custody of the state as a functional matter?
3. Has the state created or exacerbated the danger or harm that the plaintiff experienced?\footnote{113}

Besides its clarity, Ms. Hagan’s standard makes sense. Her standard revolves around the state’s own actions in restraining the liberty of individuals or creating or exacerbating a dangerous condition,\footnote{114} not around the mental condition or disability of the plaintiff, over which the state has no control.

In determining whether the state exercises functional custody over the individual, Ms. Hagan proposes the standard used in determining whether an individual is in custody for purposes of Fourth Amendment doctrine: would a reasonable individual believe that he or she was free to leave?\footnote{115} This test is supported by the language of case law, including \textit{DeShaney} itself: the majority expressed the exception to the absence of any affirmative duty under the Constitution as existing “when the State takes a person into its custody and holds him there against

\begin{itemize}
\item Hagan, \textit{supra} note 2, at Part VI.
\item Ye v. United States, 484 F.3d 634, 637 n.1 (3d Cir. 2007) (explaining that the state’s physical custody of the individual is a key factor).
\item Hagan, \textit{supra} note 2, at 787–88.
\end{itemize}
his will.” Subsequent cases frequently equate the question of whether a person is in custody with whether a reasonable person in his or her situation would feel free to leave. The Supreme Court has also equated custody with freedom to leave. Even if a person was too disabled to leave by himself or herself, the question could just as easily be asked: could a caring family member or friend simply drive up and remove him or her from state care? If the answer is yes, then there is no state obligation.

This standard is appropriately limited. Under her proposed standard, as she notes, the plaintiff in Monahan v. Dorchester would not have prevailed, while Joyce Higgs and Deborah Shelton would have been allowed to proceed with their claims. Psychiatrically or developmentally disabled individuals in community settings can virtually always move out; no state has created a statutory opportunity to detain and commit an individual who wishes to leave a community residential setting. But in all states a voluntary patient who wishes to leave a state institution can be legally detained and prevented from doing so.

Although there are a number of articles criticizing the analytical chaos created by DeShaney, few of these focus on individuals in psychiatric institutions, and, of those, only one

117. Ewolski v. City of Brunswick, 287 F.3d 492, 506 (6th Cir. 2002) (stating that a man in a hostage situation whose house was surrounded by police would not feel free to leave).
118. United States v. Drayton, 536 U.S. 194, 204 (2002) (“Nothing . . . would suggest to a reasonable person that he or she was barred from leaving the bus or otherwise terminating the encounter.”); California v. Hodari D., 499 U.S. 621, 627–28 (1991) (“[A] person has been ‘seized’ within the meaning of the Fourth Amendment only if, in view of all the circumstances surrounding the incident, a reasonable person would have believed that he was not free to leave.” (citations omitted)).
119. Hagan, supra note 2, at 782.
121. See, e.g., Teresa Cannistraro, A Call for Minds: The Unknown Extent of Societal Influence on the Legal Rights of Involuntarily and Voluntarily Committed Mental Health Patients, 19 ANNALS HEALTH L. 425 (2010).
other author has proposed a standard for interpreting *DeShaney* in this context. Sarah Kellogg endorses the standard suggested by the concurrence in the Fifth Circuit case of *Walton v. Alexander*, which called for a replacement of the voluntary–involuntary analysis with a “special relationship” standard, measured by:

1. the authority and discretion state actors have to control the environment and the behavior of the individuals in their custody;
2. the responsibilities assumed by the State,
3. the extent to which an individual in state custody must rely on the State to provide for his or her basic needs, and
4. the degree of control actually exercised by the state in a given situation.

I don’t disagree with this formulation: like Ms. Hagan’s proposal, it focuses on the situation of the person once he or she is in custody or in the institution, rather than focusing entirely on the process by which he or she ended up in state custody. A person can call a police officer for help, and end up under arrest; the fact that the individual initiated the encounter does not dictate whether he or she is free to leave. No more does the fact that a person may have initially sought help from the state have any bearing on whether that institutionalized individual has any prospect of leaving the hospital through his or her volition alone. While both Ms. Kellogg’s standard and Ms. Hagan’s standard address this crucial problem with interpreting *DeShaney*, Ms. Hagan’s proposal is more clear and provides more guidance.

V. Conclusion

There is no sign that the federal appellate courts are closer to congruence and harmony in their interpretations of *DeShaney*.

---

123. *Id.* at 358.
and its instructions on limiting affirmative constitutional rights. Ms. Hagan’s Note is timely and helpful in proposing an interpretive structure that would be more just and certainly clearer than the current caselaw. She is to be congratulated on a thoughtful and well-written Note.
Coercion, Consent, Compassion

John D. King∗

Law, fundamentally, is about people. Everything a lawyer does, every case a judge decides, every piece of scholarship a legal academic writes, has the capacity to increase or to decrease human suffering. I want to thank the Law Review for inviting me to comment on Claire Hagan’s excellent Note, Sheltering Psychiatric Patients from the DeShaney Storm: A Proposed Analysis for Determining Affirmative Duties to Voluntary Patients.1 I also want to thank Ms. Hagan for taking on such an important and complex set of issues. President Clarke kicked off this Notes Colloquium by saying that good legal scholarship covers topics that matter to people. That should seem obvious, of course, but those of you who have spent the last few months up in the Law Review office sifting through articles may wish that more of the legal academy took that advice to heart. Ms. Hagan’s effort to explore the limits of protection offered to those who are “voluntarily” committed to psychiatric facilities is a critical topic that matters greatly to a lot of people.

Attempts to distinguish “voluntary” actions and valid “consent” to an action in the legal sense often fail to comport with the average citizen’s understanding of these terms. In the world I inhabit, criminal law, issues of voluntariness and consent play crucial roles in a defendant’s journey through the justice system. The smooth operation of the criminal justice system is only achieved by the willingness of the vast majority of defendants to give up their constitutional rights. One context in which the

∗ Associate Clinical Professor of Law and Director of the Criminal Justice Clinic, Washington and Lee School of Law. LL.M., Georgetown University Law Center, 2005; J.D., University of Michigan Law School, 1996; B.A., Brown University, 1992. I am grateful for the research assistance of Mary Nerino. These comments are based on remarks made at the September 26, 2012 Washington and Lee Law Review Notes Colloquium.

issues of voluntariness and consent are particularly important is the waiver of the right to trial. Popular culture celebrates the criminal trial, the apex of our system, in which the truth is supposed to spring full grown from the structured clash of two well-armed adversaries. As we know, however, fewer than 5% of criminal cases ever reach that point. The other 95% of cases are resolved through plea bargaining and depend on an individual's consensual and voluntary waiver of her right to trial. To enter into a plea bargain, of course, the accused must voluntarily consent to a conviction. This raises complex issues concerning the meaning of voluntariness within a particular set of constraints.

Another familiar criminal context is the circumstance surrounding interrogations and the right to remain silent. We all know that a statement made while in state custody is only admissible in court if it is voluntarily made. This designation of "voluntary" turns on whether the police gave Miranda warnings and whether the individual then voluntarily waived her right to remain silent. It is an unfortunate truth that the police can obtain this consent by any manner of intimidation, trickery, isolation, or any number of levels of coercion up to a point. And courts consistently uphold confessions given under coercive

2. See Fed. R. Crim. P. 11(b) (requiring federal courts to ensure that a defendant entering a guilty plea understands the nature and consequences of the plea, and that the defendant is voluntarily choosing to plead guilty).

3. See United States v. Booker, 543 U.S. 220, 273 (2005) (Stevens, J., dissenting) ("[O]ver 95% of all federal criminal prosecutions are terminated by plea bargain.").

4. Id.; see also Brady v. United States, 397 U.S. 742, 747 (1970) ("[G]uilty pleas are valid if both 'voluntary' and 'intelligent.'" (citation omitted)).

5. See Brady, 397 U.S. at 749–55 (discussing the standard for a voluntary plea and determining that a state's use of coercive pressures to obtain a guilty plea does not negate voluntariness); Fed. R. Crim. P. 11(b)(2) (requiring the court, before accepting a guilty plea, to "address the defendant personally in open court and determine that the plea is voluntary and did not result from force, threats, or promises (other than promises in a plea agreement)").


7. Id.

circumstances, as long as these waivers meet certain formalistic requirements of *Miranda.* To a large extent, the formalism of the “*Miranda* warnings” has swallowed the true meaning of what it means to voluntarily consent to waive a right.10

As in the criminal law context, voluntariness and consent play an important role in the mental-health law issue of civil commitment.11 As Ms. Hagan explains in her Note, individuals committed to state mental hospitals have traditionally been labeled either “voluntary” or “involuntary” upon entrance into the hospital system.12 This label then dictates, for the most part, the state’s relationship to the patient and the duty of care the state owes, if any, to the patient should harm befall the patient during her hospitalization.13 Ms. Hagan quite correctly points out the two major flaws in such a framework: (1) that the labels “voluntary” and “involuntary” may be misleading when the

---

9. For example, the Supreme Court has found that a suspect who invoked his right to remain silent, then later asked a police officer, “Well, what is going to happen to me now?” had effectively waived his right. Oregon v. Bradshaw, 462 U.S. 1039, 1042, 1046 (1983). Subsequent statements were considered a voluntary confession. Id. at 1043–46.

10. See, e.g., Nirej Sekhon, *Willing Suspects and Docile Defendants: The Contradictory Role of Consent in Criminal Procedure,* 46 Harv. C.R.-C.L. L. Rev. 103, 124–26 (2011) (arguing that, after *Miranda,* courts generally do not ask whether a confession was voluntary, but generally admit confessions if *Miranda* warnings were properly administered, even though *Miranda* warnings probably do not alleviate coercive forces); Robert E. Toone, *The Incoherence of Defendant Autonomy,* 83 N.C. L. Rev. 621, 646 (2005) (discussing how bureaucratization, professionalism, and an increased focus on efficiency altered the conception of due process protections).


12. Hagan, supra note 1, at Part III.A.

patient’s individual circumstances are considered; and (2) that such a distinction may prevent courts from finding that a state has an affirmative duty to a voluntary patient, even when the patient fits within the existing functional custody or state-created danger exceptions.\(^\text{14}\) Ms. Hagan’s willingness to point out the legal fiction that underlies the formal distinction between “voluntary” and “involuntary” psychiatric patients has broad ramifications beyond the mental health context into a broader understanding of voluntariness and consent.

Limiting courts to a formalistic approach to patients’ rights and a rigid definition of voluntariness and consent is problematic. Ms. Hagan describes the term voluntary as “an artificial signifier,”\(^\text{15}\) and she is right to underscore the importance of allowing courts the discretion to delve into the background of each case in order to determine whether true, meaningful, and voluntary consent was given. She notes the importance of questioning whether a voluntarily committed patient was even competent to give informed consent and questioning whether consent was “tainted by coercion.”\(^\text{16}\) She poses questions with troubling answers in this same vein, such as “What if state law allows a facility to hold a voluntary patient for seventy-two hours after the patient decides he wants to leave?” and “What if the hospital is aware that a particular patient is at high risk of being harmed by another patient?”\(^\text{17}\) What is truly compelling, however, is the manner in which Ms. Hagan connects each of these problems within the current system not just to a case, but to the compelling story behind the case.

In his wonderful Article *Storytelling for Oppositionists and Others*, Richard Delgado years ago described the power of storytelling in legal scholarship:

> Stories, parables, and narratives are powerful means for destroying mindset—the bundle of presuppositions, received wisdoms, and shared understandings against a background of which legal and political discourse takes place . . . Ideology—

---

15. *Id.* at 780.
16. *Id.* at 785.
17. *Id.* at 729.
the received wisdom—makes current social arrangements seem fair and natural. Those in power sleep well at night—their conduct does not seem to them like oppression.18

“The cure,” Delgado said, “is storytelling . . . . [S]tories can shatter complacency and challenge the status quo . . . . They can help us understand when it is time to reallocate power.”19 One of the great accomplishments of Ms. Hagan’s Note is the way that she delves into the stories behind judicial decisions to point out the problems with the current two-dimensional approach used by courts. She argues that “[a] patient’s relationship with the state is more complicated than his voluntary status,”20 and the stories in her Note thoroughly back up this proposition.21 It is by telling stories about some of the most vulnerable members of our society that Ms. Hagan truly makes clear the urgency and importance of adopting a framework that will be able to take an individual’s treatment by the state into consideration when deciding whether patients are owed affirmative duties.

The story at the heart of the problem Ms. Hagan addresses in her Note is that of Joshua DeShaney.22 Mr. DeShaney today is thirty-three years old and he lives in an assisted-living facility outside of Oshkosh, Wisconsin.23 He is partially paralyzed; he lives with profound mental retardation, and severe and irreversible brain damage.24 He was born to very young parents in Cheyenne, Wyoming.25 His father left him and the family before Joshua was one year old.26 When Joshua was fourteen

19. Id. at 2414–15.
21. See id. at Parts II.B, IV.A-B (summarizing the case histories—the stories—behind the leading Supreme Court and circuit court cases addressing affirmative duties to institutionalized people).
24. Id.
26. Id.
months old, his mother decided that she was unable to take care of Joshua. 27 When later asked her reasons, her response was poignant: she felt herself “too young, too alone, and too poor.” 28 And so she sent him back to live with his father. 29 When she was asked later, after all of this happened, why she did that, she said, she wanted for him to have a “nice kid life” that she could not give him. 30

The next couple of years of Joshua’s life were unspeakably sad. It was marked by repeated visits to the emergency room, repeated calls to the police, and officers and state officials of various stripes coming to the house. 31 The first time Joshua went to the emergency room, when he was just two years old, he had injuries on his face, scalp, spine, buttocks, thighs, penis, ankle, and heel. 32 The Wisconsin Department of Social Services opened a file on him that remained open, in some way or the other, for the rest of his childhood. 33 A social worker was assigned to visit the house once a month, and she did. 34 During the last four months prior to the final injury at the hands of his father, that social worker never laid eyes on Joshua. 35 She went to the house one month and was told he was asleep. 36 She took no action to try to see him. 37 She went the next month and nobody was home and she didn’t go back. 38 She went the month after that, March 7th of 1983, and was told again that he was asleep. 39 That time she didn’t try to wake him up, didn’t ask to see him, but stayed in the kitchen with his father decorating a birthday cake for another

27. Id. at 14.
28. Id.
29. Id.
30. Id.
31. See Dodd, supra note 23, at 186–91 (providing detailed information about the state’s involvement with Joshua while he was living in his father’s house).
32. Id. at 186.
33. Id. at 186–89.
34. Id. at 187.
35. Id. at 187–88.
36. Id. at 188.
37. Id.
38. Id.
39. Id.
The next day Joshua showed up at the hospital, in a coma, with irreversible brain damage. Joshua's mother claims that the social worker later said, "I knew one day the phone would ring and Joshua would be dead." Despite the extensive files that were kept on Joshua DeShaney, and despite the mindboggling and infuriating negligence with which the state handled this, the Supreme Court denied relief because of its conclusion that the state had no constitutional duty to protect Joshua.

The DeShaney case was about many issues, but an important part of what it was about was competing visions of society and what duties the state owes to the vulnerable among us. As Ms. Hagan aptly discusses in her Note, it was not an easy case. Thoughtful critics of DeShaney who are sympathetic to a broader reading of state liability recognize that there could be a detrimental, real-world impact of allowing these suits to go forward and that it can be difficult to determine where the slippery slope would end. But the majority opinion in DeShaney, which Ms. Hagan criticizes rightly as "unrelentingly formal," seemed almost to revel in the tragic facts of that case, and to celebrate the Court's unwillingness to yield to the impulse of

---

40. Curry, supra note 25, at 32.
41. Id. at 33, 35–36.
42. Id. at 38.
43. DeShaney v. Winnebago Cnty. Dep't of Soc. Servs., 489 U.S. 189, 202 (1989) (“Because ... the State had no constitutional duty to protect Joshua against his father’s violence, its failure to do so—though calamitous in hindsight—simply does not constitute a violation of the Due Process Clause.”).
44. Id.
45. Compare id. at 196 (emphasizing the importance of protecting individuals against the state’s interference and power), with id. at 208–10 (Brennan, J., dissenting) (emphasizing the state’s systematic and inextricable role in protecting vulnerable citizens), and id. at 212–13 (Blackmun, J., dissenting) (emphasizing “fundamental justice” and compassion for vulnerable individuals).
46. See, e.g., Barbara E. Armacost, Affirmative Duties, Systemic Harms, and the Due Process Clause, 94 Mich. L. Rev. 982, 1003 (1996) (“Permitting individuals to bring failure-to-protect claims [like DeShaney] would require the courts to review resource-allocation decisions and permit judges to mandate a level of protection different from the level determined by the political branches.”).
47. Hagan, supra note 1, at 790.
natural sympathy.\textsuperscript{48} Instead, the Court focused on this illusory and arbitrary distinction between action and inaction,\textsuperscript{49} and the Court created an incentive for agencies, such as Winnebago County DSS, \textit{not} to act.\textsuperscript{50} And not for nothing, as Ms. Hagan pointed out, \textit{DeShaney} has very few supporters among those who have studied or written about it.\textsuperscript{51}

Ms. Hagan also artfully navigates the tension between two well-meaning groups of people who address these issues,\textsuperscript{52} one that I would characterize as rights-oriented, and the other that I would call treatment-oriented. That is a thicket in the mental health world, and one that Ms. Hagan does a nice job of navigating. She does not shy away from bold statements, and she proposes a realistic and nuanced analysis of how courts should analyze the duties that a state may owe a patient with a formal status of voluntary.\textsuperscript{53} She takes seriously Justice Cardozo’s advice, which she cites in a footnote, to resist the temptation “when the demon of formalism tempts the intellect with the lure of scientific order.”\textsuperscript{54}

The most famous part of \textit{DeShaney} is Justice Blackmun’s dissent, which is memorable for the language he uses, for the emotion he conveys, and for the frank manner in which he compares Chief Justice Rehnquist and the majority to those judges who decided \textit{Dred Scott}\textsuperscript{55} in the fugitive slave cases, for

\begin{itemize}
\item \textsuperscript{48} See, \textit{e.g.}, \textit{DeShaney}, 489 U.S. at 202 (majority opinion) (“Judges and lawyers, like other humans, are moved by natural sympathy in a case like this . . . .”).
\item \textsuperscript{49} See \textit{id.} at 200–01 (explaining that states’ affirmative actions trigger corresponding duties); \textit{id.} at 203 (“The most that can be said of the state functionaries in this case is that they stood by and did nothing when suspicious circumstances dictated a more active role for them.”).
\item \textsuperscript{50} See, \textit{e.g.}, \textit{id.} at 212 (Brennan, J., dissenting) (pointing out that the majority’s strict action–inaction analysis incentivizes a state “at the critical moment, to shrug its shoulders and turn away” rather than take protective action, thereby avoiding any potential liability).
\item \textsuperscript{51} See Armacost, \textit{supra} note 46, at Part I (summarizing the overwhelmingly negative reaction to \textit{DeShaney}).
\item \textsuperscript{52} Hagan, \textit{supra} note 1, at 753–54, 778–79.
\item \textsuperscript{53} \textit{id.} at Part VI.
\item \textsuperscript{54} \textit{Benjamin Cardozo, The Nature of the Judicial Process} 66 (1921).
\item \textsuperscript{55} \textit{Dred Scott v. Sandford}, 60 U.S. 393 (1857).
\end{itemize}
their exaltation of formalism over justice.\textsuperscript{56} Blackmun memorably wrote: “Faced with a choice, I would adopt a ‘sympathetic’ reading, one which comports with dictates of fundamental justice, and recognizes that compassion need not be exiled from the province of judging.”\textsuperscript{57} Ms. Hagan’s Note is not only well-researched, powerfully argued, and beautifully written, it is also, at its heart, a compassionate piece of scholarship that allows for interpretation of the law to affect lives for the better.

Ms. Hagan’s Note is timely and helpful in proposing an interpretive structure that would be more just and certainly more clear than the current caselaw. She is to be congratulated on a thoughtful and well-written Note.

\textsuperscript{56} \textit{DeShaney}, 489 U.S. at 212–13 (Blackmun, J., dissenting).

\textsuperscript{57} \textit{Id.} at 213.
Statement of Ownership, Management, and Circulation

1. Publication Title
Washington and Lee Law Review

2. Publication Number
0 0 4 3 - 0 4 6 3

3. Filing Date

4. Issue Frequency
Quarterly

5. Number of Issues Published Annually
Four

6. Annual Subscription Price
$35.00

7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4®)
Washington and Lee Law Review
Washington and Lee Univ. School of Law
Lexington, VA 24450

Contact Person
Lisa Gearheart

Telephone (Include area code)
540-458-8566

8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer)
Same

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank)
Publisher (Name and complete mailing address)
Same

Editor (Name and complete mailing address)
Alexander M. Sugzda, Editor in Chief
Washington and Lee Law Review
Washington and Lee Univ. School of Law, Lexington, VA 24450

Managing Editor (Name and complete mailing address)
Luther R. Ashworth II & Paul Bennett IV, Managing Editors
Washington and Lee Law Review
Washington and Lee Univ. School of Law, Lexington, VA 24450

10. Owner (Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)

Full Name
Washington and Lee University
Complete Mailing Address
Lexington, VA 24450

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box

Full Name

Complete Mailing Address

12. Tax Status (For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one)
The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

PS Form 3526, August 2012 (Page 1 of 3 (Instructions Page 3))  PSN: 7530-01-000-9931  PRIVACY NOTICE: See our privacy policy on www.usps.com.
15. Extent and Nature of Circulation

<table>
<thead>
<tr>
<th>a. Total Number of Copies (Net press run)</th>
<th>Average No. Copies Each Issue During Preceding 12 Months</th>
<th>No. Copies of Single Issue Published Nearest to Filing Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mailed Outside-County Paid Subscriptions Stated on PS Form 3541 (Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)</td>
<td>450</td>
<td>450</td>
</tr>
<tr>
<td>Mailed In-County Paid Subscriptions Stated on PS Form 3541 (Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)</td>
<td>269</td>
<td>269</td>
</tr>
<tr>
<td>Paid Distribution Outside the Mails Including Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Paid Distribution Outside USPS®</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>Paid Distribution by Other Classes of Mail Through the USPS (e.g., First-Class Mail®)</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>c. Total Paid Distribution (Sum of 15b (1), (2), (3), and (4))</td>
<td>344</td>
<td>344</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>d. Free or Nominal Rate Distribution (By Mail and Outside the Mail)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Free or Nominal Rate Outside-County Copies included on PS Form 3541</td>
<td>2</td>
</tr>
<tr>
<td>Free or Nominal Rate In-County Copies Included on PS Form 3541</td>
<td>2</td>
</tr>
<tr>
<td>Free or Nominal Rate Copies Mailed at Other Classes Through the USPS (e.g., First-Class Mail)</td>
<td>75</td>
</tr>
<tr>
<td>Free or Nominal Rate Distribution Outside the Mail (Carriers or other means)</td>
<td></td>
</tr>
</tbody>
</table>

| e. Total Free or Nominal Rate Distribution (Sum of 15d (1), (2), (3) and (4)) | 77 | 77 |

| f. Total Distribution (Sum of 15c and 15e) | 421 | 421 |

| g. Copies not Distributed (See Instructions to Publishers #4 (page #3)) | 29 | 29 |

| h. Total (Sum of 15f and g) | 450 | 450 |

| i. Percent Paid (15c divided by 15f times 100) | 82% | 82% |


17. Publication of Statement of Ownership

☐ If the publication is a general publication, publication of this statement is required. Will be printed in the Winter 2013 issue of this publication.

☐ Publication not required.

18. Signature and Title of Editor, Publisher, Business Manager, or Owner

Alexander M. Sugzda, Editor in Chief

Date

2-11-2013

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including civil penalties).