ARTICLES

An Antitrust Common Law for the Twenty-First Century
Thomas A. Piraino, Jr.

Allocating Influence
Heidi Reamer Anderson

Collateral Damage and Securities Litigation
Bradford Cornell & James C. Rutten

From Identity Politics to Ideology Politics
Jessica Knouse

Extending the Shadow of the Law: Using Hybrid Mechanisms to Develop Constitutional Norms in Socioeconomic Rights Cases
Brian Ray

Disability-Selective Abortion and the Americans with Disabilities Act
Dov Fox & Christopher L. Griffin, Jr.

Height Discrimination in Employment
Isaac B. Rosenberg

Divided Loyalties: How the Metlife v. Glenn Standard Discounts ERISA Fiduciaries’ Conflicts of Interest
Beverly Cohen

Web Site Proprietorship and Online Harassment
Nancy S. Kim

Good Faith, Fiduciary Duties, and the Business Judgment Rule in Delaware
Clark W. Furlow
The *Utah Law Review* was founded in 1948 with the purpose of serving the interests of students, the bench, and the bars of the state and surrounding areas. Since then, its scope has expanded to include legal issues of importance both domestically and internationally. The *Utah Law Review* is a student-run organization, with all editorial and organizational decisions made by student-editors enrolled at the S.J. Quinney College of Law at the University of Utah. The *Utah Law Review* publishes four issues per year, with each issue containing approximately 250 pages of legal scholarship.

The *Utah Law Review* makes every effort to conform to the following editing sources: *The Bluebook: A Uniform System of Citation* (18th ed. 2005) and *The Chicago Manual of Style* (15th ed. 2003). The *Utah Law Review* Board of Editors strives to maintain the text of published articles in its original format by editing only for accuracy and, where necessary, clarity. Views expressed herein are to be attributed to their authors and not to the *Utah Law Review*, its editors, the S.J. Quinney College of Law, or the University of Utah.

COPYRIGHT: © 2009 by the *Utah Law Review* Society. Printed Fall 2009 (ISSN 0042-1448). In addition to the rights to copy that are granted by law, the owner of the copyright of each article published in this issue, except as expressly otherwise noted, grants permission for copies of that article to be made and used for educational purposes by nonprofit educational institutions, provided that the author and journal are identified, that proper notice of copyright is affixed to each copy of the article, and that copies are distributed at or below cost.

ACKNOWLEDGEMENTS: The editors of the *Utah Law Review* wish to express their gratitude to the following: The University of Utah S.J. Quinney College of Law, the Journal Alumni Association and the University of Utah Publications Council for their support; Professor Amy J. Wildermuth, faculty advisor; Angela Turnbow, editorial assistant; and the S.J. Quinney College of Law faculty and staff.
SUBMISSIONS: The Utah Law Review receives submissions in March and April. Submissions should include a cover letter, providing the title of the work and author’s contact information (phone number and e-mail address). Submissions should be double-spaced, in Times 11-point or larger font, printed on one side of the page only, contain footnotes rather than endnotes, and have numbered pages. Microsoft Word format is preferred, but other formats also will be accepted. Electronic submissions can be sent to the Utah Law Review Expresso account or via e-mail at utahlawreview_submit@law.utah.edu, or a hard copy can be sent to: Executive Articles Editor, Utah Law Review, S.J. Quinney College of Law, University of Utah, 332 S 1400 E Rm. 101, Salt Lake City, UT 84112-0730.

SUBSCRIPTIONS: The Utah Law Review is published quarterly. All subscriptions begin January 1 and expire on December 31. For subscriptions, renewals, or cancelations of the Utah Law Review, please send notice and/or payment to: Utah Law Review, Attn: Angela Turnbow, S.J. Quinney College of Law, University of Utah, 332 S 1400 E Rm. 101, Salt Lake City, UT 84112-0730.

PRICES: $40.00 per volume per year; United States and possessions $45.00 per volume per year; foreign (postage included)

To subscribe to more than one copy of the Utah Law Review, send notice of the number of copies requested along with payment. Please note that renewal payment must be received in advance before any issues are sent out to a subscriber. If you work with a subscription agency, send notice and payment to that agency and it will contact the Utah Law Review regarding the status of your subscription.

The Utah Law Review can now be accessed online at: http://epubs.utah.edu/index.php/ulr.

ISSUES AND REPRINTS: Single issues and back issues are available from William S. Hein & Co., Inc., 1285 Main Street, Buffalo, New York 14209-1987; telephone (800) 828-7571; e-mail: wsheinco@class.org. Reprints are available from Joe Christensen, Inc., 1540 Adams Street, Lincoln, NE 68521; telephone (800) 228-5030; e-mail: sales@christensen.com.
S. J. QUINNEY COLLEGE OF LAW AT THE UNIVERSITY OF UTAH

FACULTY & ADMINISTRATION 2009–2010

BROWN, TENEILLE (2009), Associate Professor of Law. B.A., 2000, Univ. of Pennsylvania; J.D., 2004, Univ. of Michigan.
CHAHINE, KENNETH GREGORY (2009), Visiting Assistant Professor of Law. Ph.D., 1992, Univ. of Michigan; J.D., 1996, Univ. of Utah.


GUIORA, AMOS N. (2007), Professor of Law. B.A., 1979, Kenyon College; J.D., 1985, Case Western Reserve Univ.


LOCKHART, WILLIAM J. (1964), Professor of Law. B.A., 1955, J.D., 1961, Univ. of Minnesota.


McLAUGHLIN, NANCY ASSAF (2000), Associate Dean for Faculty Research and Development and Robert S. Swenson Professor of Law. B.S., 1987, Univ. of Massachusetts, Amherst; J.D., 1990, Univ. of Virginia.


MORRIS, JOHN K. (1979), General Counsel, University of Utah and Professor of Law. B.A., 1966, Univ. of California at Los Angeles; J.D., 1969, Univ. of California at Berkeley.


POULTER, SUSAN R. (1990), Professor of Law Emerita. B.S., 1965, Ph.D., 1969, Univ. of California at Berkeley; J.D., 1983, Univ. of Utah.


ROSKY, CLIFFORD J. (2008), Associate Professor of Law. B.A., 1996, Amherst College; J.D., 2001, Yale Univ.


S.J. QUINNEY LAW LIBRARY FACULTY

DARAIS, SUZANNE (1990), Head of Information Technical Services and Adjunct Professor of Law. B.S., 1983, Univ. of Utah; J.D., 1989, Brigham Young Univ.; M.L.S., 1990, Univ. of Washington.

FULLER, RONALD (2006), Assistant Librarian and Adjunct Professor of Law. B.A., 2001, J.D., 2005, Brigham Young Univ.

MCPHAIL, ROSS (2007), Assistant Librarian and Adjunct Professor of Law. B.A., 1990, Quincy College; J.D., 1994, Hamline Univ.

STEPHENSON, LINDA (1986), Head of Information Access Services and Adjunct Professor of Law. M.A., 1983, Univ. of Missouri at Columbia; J.D., 1987, Univ. of Missouri-Kansas City.

BOARD OF EDITORS
2009–2010

Editor-in-Chief
ADAM E. WEINACKER

Managing Editor
MIRIAH SORENSON ELLIOTT

Executive Articles Editor
TYLER HAWKINS

Note & Comment Editors
RACHEL JACQUES
ADAM KAAS
NATE OLSEN

Senior Staff
BROOKS BENSON
JACOB LLOYD

Executive Text Editor
SAMUEL HOOD

Executive Footnote Editor
STEVEN JOFFE

Symposium Editor
TESIA STANLEY

Articles Editors
JACOB FONNESBECK
SARAH GOLDBERG
BEN LEAR
MARY ANN CALL MAY
GREG NEWMAN
ADAM REISER

Associate Editors
DAVID CALL
NICHOLAS CRAFT
ASHTON HYDE
RAZVAN UNGUREANU
RUKE WANG

Staff Members
BETH RANSCHAU
JON REED
DALLAS ROSEVEAR
JEFFRY ROSS
ELIZABETH SILVESTRINI
JAMES SWAN
RACHEL WERTHEIMER

Advisor Committee
PATRICIA M. LEITH
CLAYTON J. PARR

Editorial Assistant
ANGELA TURNBOW

Computing Services
DUC PHUONG
# Table of Contents

## Articles

<table>
<thead>
<tr>
<th>Title</th>
<th>Author(s)</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>An Antitrust Common Law for the Twenty-First Century</td>
<td>Thomas A. Piraino, Jr.</td>
<td>635</td>
</tr>
<tr>
<td>Allocating Influence</td>
<td>Heidi Reamer Anderson</td>
<td>683</td>
</tr>
<tr>
<td>Collateral Damage and Securities Litigation</td>
<td>Bradford Cornell &amp; James C. Rutten</td>
<td>717</td>
</tr>
<tr>
<td>From Identity Politics to Ideology Politics</td>
<td>Jessica Knouse</td>
<td>749</td>
</tr>
<tr>
<td>Extending the Shadow of the Law: Using Hybrid Mechanisms to Develop Constitutional Norms in Socioeconomic Rights Cases</td>
<td>Brian Ray</td>
<td>797</td>
</tr>
<tr>
<td>Disability-Selective Abortion and the Americans with Disabilities Act</td>
<td>Dov Fox &amp; Christopher L. Griffin, Jr.</td>
<td>845</td>
</tr>
<tr>
<td>Height Discrimination in Employment</td>
<td>Isaac B. Rosenberg</td>
<td>907</td>
</tr>
<tr>
<td>Divided Loyalties: How the Metlife v. Glenn Standard Discounts ERISA Fiduciaries’ Conflicts of Interest</td>
<td>Beverly Cohen</td>
<td>955</td>
</tr>
<tr>
<td>Web Site Proprietorship and Online Harassment</td>
<td>Nancy S. Kim</td>
<td>993</td>
</tr>
<tr>
<td>Good Faith, Fiduciary Duties, and the Business Judgment Rule in Delaware</td>
<td>Clark W. Furlow</td>
<td>1061</td>
</tr>
</tbody>
</table>
AN ANTITRUST COMMON LAW FOR THE TWENTY-FIRST CENTURY

Thomas A. Piraino, Jr.*

TABLE OF CONTENTS

I. INTRODUCTION ................................................................. 636
   A. Political Support for a New Antitrust Common Law ................. 636
   B. The Courts’ Varying Interpretation of the Antitrust Laws ........ 637
   C. The Foundations of a New Antitrust Common Law ................. 639
   D. A Proposal for a New Antitrust Common Law ..................... 640

II. THE EVOLUTION FROM A FIXED TO A CONTINUUM-BASED APPROACH TO ANTITRUST ......................... 641

III. THE COURTS’ APPROACH TO FIRST-PORTION CONDUCT ................................................................. 644
   A. Horizontal Restraints ....................................................... 645
   B. Vertical Resale Restraints ................................................. 647
      1. The Legality of Restraints Implemented by Manufacturers .... 647
      2. The Illegality of Restraints Induced by Distributors .......... 648
   C. Vertical Mergers and Joint Ventures .................................. 649

IV. THE COURTS’ APPROACH TO SECOND-PORTION CONDUCT .............................................................. 651

V. THE COURTS’ APPROACH TO THIRD-PORTION CONDUCT ................................................................. 653
   A. Protecting Monopolists’ Legitimate Competitive Conduct ......... 653
   B. The Evolution of the Essential Facilities Doctrine ............... 655

VI. AN INTEGRATED ANTITRUST COMMON LAW FOR THE 21ST CENTURY ................................................. 657
   A. First-Portion Conduct ....................................................... 658
      1. Horizontal Restraints of Trade .................................... 658
      2. Vertical Resale Restraints ............................................ 658
      3. Vertical Mergers and Joint Ventures ............................... 659
   B. Second-Portion Conduct .................................................. 659
   C. Third-Portion Conduct ..................................................... 659

VII. APPLYING THE PROPOSED APPROACH TO HORIZONTAL RESTRAINTS OF TRADE .................................. 659

VIII. APPLYING THE PROPOSED APPROACH TO VERTICAL RESALE RESTRAINTS ........................................ 662

IX. APPLYING THE PROPOSED APPROACH TO VERTICAL MERGERS AND JOINT VENTURES ....................... 664

X. APPLYING THE PROPOSED APPROACH TO HORIZONTAL MERGERS AND JOINT VENTURES .................. 666

* © 2009 Thomas A. Piraino, Jr., Vice President, General Counsel, and Secretary, Parker Hannifin Corp., Cleveland, Ohio; J.D., Cornell Law School, 1974; B.A., Allegheny College, 1971; Distinguished Adjunct Lecturer, Case Western Reserve University School of Law. The opinions expressed in this Article are personal to the author and do not reflect the opinions of Parker Hannifin Corp.
A. The Competitive Effects of Horizontal Mergers and Joint Ventures .......... 666
B. A Balancing Test for Horizontal Mergers and Joint Ventures ................... 666
C. Applying the Proposed Approach to Particular Mergers .......................... 669
   1. Automotive Mergers ............................................................................... 669
   2. Airline Mergers ..................................................................................... 670
   3. Hospital Mergers ................................................................................... 671
D. Applying the Proposed Approach to Particular Joint Ventures ................ 672
   1. Joint Ventures That Facilitate Entry ..................................................... 672
   2. Upstream Joint Ventures ...................................................................... 673
   3. Downstream Joint Ventures .................................................................. 674
XI. APPLYING THE PROPOSED APPROACH TO MONOPOLY CONDUCT .......... 676
   A. The Plaintiff’s Burden of Proof ............................................................. 676
   B. Access Restrictions ............................................................................... 677
   C. Applying the Approach to High Technology Networks ........................ 678
XII. CONCLUSION .......................................................................................... 681

I. INTRODUCTION

A. Political Support for a New Antitrust Common Law

The economic problems facing the United States at the end of the first decade of the twenty-first century resemble those that prompted the passage of the first federal antitrust laws at the turn of the nineteenth century. Those laws were a response to the public’s concern about the misuse of economic and political power by trusts such as Standard Oil and United States Steel. Teddy Roosevelt called the owners of such trusts “malefactors of great wealth.” A large number of Americans have blamed the economic crisis of 2008 on the “outright recklessness and greed” of the CEOs of commercial banks, investment banks and hedge funds who subjected their institutions to undue risk for their own gain. As a result, there is now a political consensus in favor of increased regulation of the conduct of American business.

Antitrust enforcement is now at one of its historic inflection points, poised between the laissez-faire policies of the Bush administration and the more activist

3 See Gerald F. Seib, In Crisis, Opportunity for Obama, WALL ST. J., Nov. 21, 2008, at A2 (“[C]ertainly the broad dissatisfaction with the way financial markets were regulated will make it easier to rebuild regulatory structures.”).
regulatory approach desired by the American public and the Obama administration. The current economic crisis provides an opportunity—and a risk—for President Obama’s antitrust regulators. They could develop policies that more effectively protect consumers from anticompetitive conduct, or they could over-reach by precluding American firms from competing aggressively to provide the lower prices or innovative products desired by American consumers.

Fortunately, the antitrust regulators of the twenty-first century can base their course upon a rich tradition of antitrust decision-making dating back to 1890. This Article explains how the federal courts and antitrust enforcement agencies can integrate the enduring values of previous antitrust eras into a comprehensive antitrust common law capable of regulating the principal types of competitive conduct in which American firms are likely to engage in the twenty-first century. The new common law proposed in this Article can provide the foundation for a centrist approach to antitrust enforcement. This approach would be consistent with the desire of the Obama administration and the American public for a bipartisan solution to economic problems. Since it comports with the most discerning antitrust precedent of the last 119 years, the proposed approach should be supported across the political spectrum.

B. The Courts’ Varying Interpretation of the Antitrust Laws

When Congress enacted the principal federal antitrust statutes, it allowed the courts considerable leeway to devise a federal common law of competition. Each of these statutes is phrased in general terms, leaving to the federal courts the task of defining how those terms should apply to specific types of competitive conduct. The broad authority delegated by Congress has allowed the federal judiciary, over the last 119 years, to create a flexible and ever-evolving set of competitive rules for American business. The federal courts’ ability to incorporate

---

5 See id. (“As President, I will direct my administration to reinvigorate antitrust enforcement.”).
6 See John Harwood, Change is Landing in Old Hands, N.Y. TIMES, Nov. 23, 2008, at WK 1, WK 4 (“Mr. Obama’s pledge of bipartisan comity foreshadows centrist compromise on national problems that have long appeared intractable.”).
8 Section 1 of the Sherman Act of 1890 prohibits “[e]very contract, combination . . . or conspiracy, in restraint of trade . . . .” 15 U.S.C. § 1 (2006). Section 2 of the Sherman Act states it is illegal for firms to “monopolize, or attempt to monopolize” trade in any relevant market. Id. § 2. Section 7 of the Clayton Act of 1918 prohibits mergers or acquisitions whose effect “may be substantially to lessen competition, or to tend to create a monopoly.” Id. § 18.
9 Thomas Sullivan has stated, “It is axiomatic that the framers of the Sherman Act intended to enact a statute of general principles that afforded the federal judiciary wide discretion in interpretation.” E. Thomas Sullivan, On Nonprice Competition: An Economic and Marketing Analysis, 45 U. Pitt. L. Rev. 771, 772 (1984).
the latest political and economic theories into their decisions has ensured the continued relevance of the antitrust laws, but it has also created conflicting standards of competitive conduct.\footnote{As John J. Flynn has observed, “Delegating to the courts a common law authority... enabled the courts to take account of... the changing insights from political science, history, economics, and related disciplines.” John J. Flynn, The Role of Rules in Antitrust Analysis, 2006 UTAH L. REV. 605, 621. Herbert Hovenkamp has argued that “almost every political generation has abandoned the [antitrust] policy of its predecessors in favor of something new.” Herbert Hovenkamp, Antitrust Policy After Chicago, 84 MICH. L. REV. 213, 213 (1985).}

The history of antitrust decision-making since 1890 can be divided into three eras, based upon the degree to which the courts and enforcement agencies tended toward an interventionist or laissez-faire approach to antitrust regulation: the “Early Era,” from the passage of the Sherman Act in 1890 to the 1930s; the “Harvard Era,” from the 1930s until the late 1960s; and the “Chicago Era,” which has continued from the late 1970s to the present time.

The principal cases of the Early Era involved a titanic struggle between the federal government and the steel, oil, and railroad trusts to determine whether the Sherman Act’s prohibition of restraints of trade and monopolies would have any real effect on American business. By the end of the Early Era, the federal courts had made it clear that they would exercise their congressionally delegated authority to regulate American companies’ competitive conduct.\footnote{See, e.g., United States v. Trans-Missouri Freight Ass’n, 166 U.S. 290, 292–97, 299, 342 (1897) (illegality of price fixing among 18 railroad companies); N. Sec. Co. v. United States, 193 U.S. 197, 325–26 (1904) (illegality of merger between Great Northern and Northern Pacific railroads).}

The Harvard Era was characterized by an aggressive interventionist approach to antitrust regulation. The Harvard Era took its name from a group of scholars and economists at Harvard University who argued that, when Congress enacted the antitrust laws, it was concerned with protecting individual competitors from the market power wielded by large firms.\footnote{See Herbert Hovenkamp, Distributive Justice and the Antitrust Laws, 51 GEO. WASH. L. REV. 1, 16–17 (1982) (stating that Congress was concerned with distributing economic and political power more widely).} From the early 1930s through the 1960s, the federal courts used Harvard rationale to preclude a wide range of competitive conduct thought to be harmful to small business.\footnote{See, e.g., United States v. Arnold, Schwinn & Co., 388 U.S. 365, 379, 382 (1967), overruled by Cont’l T.V., Inc. v. GTE Sylvania Inc., 433 U.S. 36, 37 (1977) (declaring per se illegality of non-price vertical restrictions imposed by manufacturers on distributors); United States v. Von’s Grocery Co., 384 U.S. 270, 281, 301 (1966) (declaring illegality of a merger between supermarkets with a combined market share of only 8.9 percent); United States v. Aluminum Co. of Am., 148 F.2d 416, 431 (2d Cir. 1945) (finding that Alcoa violated the Sherman Act simply by adding to its capacity in anticipation of increased demand for its products).}

The principal cases of the Early Era involved a titanic struggle between the federal government and the steel, oil, and railroad trusts to determine whether the Sherman Act’s prohibition of restraints of trade and monopolies would have any real effect on American business. By the end of the Early Era, the federal courts had made it clear that they would exercise their congressionally delegated authority to regulate American companies’ competitive conduct.\footnote{See, e.g., United States v. Trans-Missouri Freight Ass’n, 166 U.S. 290, 292–97, 299, 342 (1897) (illegality of price fixing among 18 railroad companies); N. Sec. Co. v. United States, 193 U.S. 197, 325–26 (1904) (illegality of merger between Great Northern and Northern Pacific railroads).}

The Harvard Era was characterized by an aggressive interventionist approach to antitrust regulation. The Harvard Era took its name from a group of scholars and economists at Harvard University who argued that, when Congress enacted the antitrust laws, it was concerned with protecting individual competitors from the market power wielded by large firms.\footnote{See Herbert Hovenkamp, Distributive Justice and the Antitrust Laws, 51 GEO. WASH. L. REV. 1, 16–17 (1982) (stating that Congress was concerned with distributing economic and political power more widely).} From the early 1930s through the 1960s, the federal courts used Harvard rationale to preclude a wide range of competitive conduct thought to be harmful to small business.\footnote{See, e.g., United States v. Arnold, Schwinn & Co., 388 U.S. 365, 379, 382 (1967), overruled by Cont’l T.V., Inc. v. GTE Sylvania Inc., 433 U.S. 36, 37 (1977) (declaring per se illegality of non-price vertical restrictions imposed by manufacturers on distributors); United States v. Von’s Grocery Co., 384 U.S. 270, 281, 301 (1966) (declaring illegality of a merger between supermarkets with a combined market share of only 8.9 percent); United States v. Aluminum Co. of Am., 148 F.2d 416, 431 (2d Cir. 1945) (finding that Alcoa violated the Sherman Act simply by adding to its capacity in anticipation of increased demand for its products).}
approach of the Harvard School. For example, in a 1966 law review article, Robert Bork argued that the antitrust laws were designed not to protect individual competitors against large firms’ exercise of market power, but simply to increase the efficiency of the American economy.\(^{14}\) Bork defined economic efficiency in terms of conditions that maximized wealth, and he equated wealth enhancement with “consumer welfare,” meaning lower costs, reduced prices, and increased output of goods and services desired by consumers.\(^{15}\) During the Chicago Era, the federal courts dismantled many of the Harvard Era’s proscriptions, giving firms wide latitude to engage in aggressive competitive conduct perceived to benefit consumers.\(^{16}\)

**C. The Foundations of a New Antitrust Common Law**

The history of antitrust decision-making, like the history of most political thought in the United States, has been to move and change with the times, but also, over its long course, to build progressively upon lasting values. Although courts’ approach to antitrust has varied since the passage of the Sherman Act in 1890, each of the prior eras of antitrust decision-making has bequeathed certain enduring values to the present day. Because the consensus on these values has been hard-won, it is likely to be long-lasting.\(^{17}\)

A new school of antitrust, just beginning to gain prominence, is attempting to identify those values around which a new consensus on antitrust regulation can form. The need for a synthesis between the opposing approaches of Harvard and Chicago became apparent in the closing years of the twentieth century as certain commentators, the Clinton administration’s antitrust enforcement agencies, and some federal courts began to adopt a post-Chicago approach to antitrust regulation.\(^{18}\) More aggressive than the Chicago School but less interventionist than the Harvard School, the post-Chicago approach recognizes that markets must be respected, but it also is aware of the appropriate circumstances under which courts


\(^{15}\) *Id.*

\(^{16}\) See, e.g., Bus. Electronics Corp. v. Sharp Electronics Corp., 485 U.S. 717, 721, 735–36 (1988) (finding that a conspiracy between a manufacturer and dealer to terminate another price-cutting dealer should not be per se illegal); Cont’l T.V., 433 U.S. at 37 (overruling *Schwinn* by holding that non-price vertical restrictions should not be per se illegal).

\(^{17}\) As Professor Hovenkamp has explained, “Each of these [antitrust] schools left an impression that affected antitrust policy indefinitely . . . .” Hovenkamp, *supra* note 10, at 214.

should intervene to prevent abuses of the competitive process harmful to consumers. The post-Chicago movement toward a middle ground holds the promise of reconciling the conflicting decisions of past eras and establishing an integrated common law of competitive conduct.

Post-Chicago decisions have accelerated what had been a slow evolution in the courts’ general approach to antitrust analysis. Section II of this Article describes this 119-year evolution. During most of that period, the courts followed a formalistic approach, assuming they could employ only two opposing and discrete methods of analyzing competitive conduct. However, a 1999 Supreme Court decision adopted a more sophisticated, continuum-based approach that will allow courts to precisely tailor their analysis to the likely economic effects of the conduct at issue. In California Dental Association v. FTC (“California Dental”), the Supreme Court, citing the work of this author and other post-Chicago commentators, rejected the fixed antitrust categories of the past and held that fact finders should engage in whatever degree of analysis they deem appropriate to confirm the impact of the relevant conduct upon consumers. Relying on California Dental, courts should recognize that their degree of inquiry in particular cases should depend upon how confident they can be about how the relevant conduct affects the goals most important to U.S. consumers: i.e., reasonable price levels, an abundance of alternative products and services, and (now more than ever) the continued viability of American companies.

A discerning look at the 119 years of antitrust decision-making reveals that courts have been surprisingly consistent in their attempts to protect these core desires of American consumers. Sections III, IV, and V of this Article explain how courts have analyzed the consumer welfare effects of the principal types of competitive conduct in which American firms have engaged since the passage of the Sherman Act in 1890.

D. A Proposal for a New Antitrust Common Law

Section VI proposes a new antitrust common law that integrates the enduring standards of the past by placing all competitive conduct on a continuum according to its likely effect on consumers. Courts’ experience over the last 119 years should allow them to classify the principal types of conduct in which American firms are most likely to engage into one of three categories on the continuum of antitrust analysis, progressing from conduct requiring the least amount of inquiry to conduct requiring the most.

Courts have now become familiar enough with the consumer welfare effects of horizontal restraints of trade, vertical resale restraints, and vertical mergers and joint ventures to place them in the first portion of the antitrust continuum, where they can apply presumptions of legality or illegality to such conduct. To emphasize

---

20 See infra note 38 and accompanying text.
21 526 U.S. at 781.
the place of such conduct on the continuum, horizontal restraints of trade, vertical resale restraints, and vertical mergers and joint ventures will hereinafter be referred to as “First-Portion Conduct.” Sections VII, VIII, and IX explain how courts can use the new common law described in this Article to more effectively analyze the types of First-Portion Conduct in which American firms are most likely to engage in the twenty-first century.

The courts’ 119-year experience with horizontal mergers and joint ventures has demonstrated that such transactions are as likely to benefit consumers as to harm them. Thus, courts should place such transactions in the second portion of the antitrust continuum, where they can balance their potential beneficial and adverse effects before ruling on their legality. Horizontal mergers and joint ventures will henceforth be referred collectively as “Second-Portion Conduct.” Section X explains how the courts can prioritize the relevant factors necessary to confirm the legality of the types of horizontal mergers and joint ventures in which American firms are most likely to participate in the near future.

Monopoly conduct should require the most detailed inquiry of all. Firms may obtain and retain monopoly power either by unfairly excluding their rivals from the relevant market or by successfully delivering the most innovative products to consumers at the lowest possible prices. In many cases, these two types of behavior are indistinguishable on their face. Without engaging in a more detailed inquiry, courts could deter legitimate competitive conduct. Thus, a court should not hold a monopolist liable until after it has confirmed that the monopolist had an anticompetitive intent to engage in the relevant conduct. Monopoly conduct will henceforth be called “Third-Portion Conduct.” Section XI explains how courts can distinguish between permissible and illegal conduct undertaken by present-day monopolists such as Google and Microsoft.

II. THE EVOLUTION FROM A FIXED TO A CONTINUUM-BASED APPROACH TO ANTITRUST

The courts’ general approach to antitrust regulation has become more sophisticated in recent years. In earlier cases, courts adopted a formalistic approach, under which they mechanically classified most competitive conduct into one of two opposing categories. More recently, however, courts have recognized that all antitrust inquiry has the same objective—to determine the impact of the relevant conduct on consumers—and that courts can engage in several different degrees of inquiry depending upon how confident they can be of the consumer welfare effects of the relevant conduct.

During most of the last 119 years, federal courts assumed there were only two means of analyzing restraints of trade. Under the “per se rule,” courts conclusively presumed the illegality of the conduct at issue, affording no opportunity to defendants to rebut the presumption by proving a justification for the conduct or
showing market factors limiting its potential adverse effect.\textsuperscript{22} Under the “rule of reason,” however, courts allowed no presumptions to benefit plaintiffs, and they refused to find a restraint of trade illegal until a plaintiff introduced specific proof of a defendant’s market power and other relevant indications of a restraint’s anticompetitive effects.\textsuperscript{23}

During the Early Era, the Supreme Court began to distinguish between conduct appropriate for per se condemnation and conduct that could be judged under a more lenient rule of reason.\textsuperscript{24} In the Harvard Era, the Supreme Court and lower federal courts increasingly used a per se approach to preclude restraints of trade.\textsuperscript{25} The expansion of the per se rule created a clear dichotomy between the large number of restraints of trade that could be classified as illegal ab initio and the smaller number of restraints that could only be precluded after a case-by-case analysis of their particular effects on competition. Courts in the Chicago Era took an opposite approach to the per se/rule of reason divide. The rule of reason, with its requirement of empirical proof of economic effects, was a perfect weapon to implement Chicago Era assumptions about the natural efficiency of markets. By the late 1970s, federal courts— influenced by the scholarship of Chicago Era academics—began to rethink the application of the per se rule and to broaden the scope of the rule of reason.\textsuperscript{26}

As the Chicago Era reached its zenith in the late 1980s, the per se rule and the rule of reason were so divergent that a court’s choice of one analysis over the other

\textsuperscript{22} Herbert Hovenkamp has concluded that conduct is considered per se illegal “because no one has made a plausible argument that the action is competitive, and its anticompetitive potential seems fairly obvious.” Herbert Hovenkamp, Antitrust Policy, Restricted Distribution, and the Market for Exclusionary Rights, 71 MINN. L. REV. 1293, 1294–95 n.8 (1987).


\textsuperscript{24} See Standard Oil Co. of N.J. v. United States, 221 U.S. 1, 63–70 (1911) (acknowledging that, despite the Sherman Act’s broad condemnation of all “restraints of trade,” the Act precluded only “unreasonable” restraints); United States v. Trans-Missouri Freight, Inc., 166 U.S. 290, 342 (1897) (applying a per se approach to a horizontal price-fixing arrangement).

\textsuperscript{25} See, e.g., United States v. Arnold, Schwinn & Co., 388 U.S. 365, 379 (1967), overruled by Cont’l T.V., Inc., 433 U.S. at 37 (applying the per se rule to non-price vertical restraints); N. Pac. R.R. Co. v. United States, 356 U.S. 1, 5–6 (1958) (applying the per se rule to a tying arrangement); United States v. Socony-Vacuum Oil Co., 310 U.S. 150, 217–20 (1940) (applying the per se rule to a price-fixing agreement).

\textsuperscript{26} See Nw. Wholesale Stationers, Inc. v. Pac. Stationery & Printing Co., 472 U.S. 284, 295–97 (1985) (refusing to apply the per se rule to a decision by a purchasing cooperative to expel one of its members); NCAA v. Bd. of Regents of the Univ. of Okla., 468 U.S. 85, 100–04 (1984) (applying the rule of reason to limits the NCAA imposed on the number of times its member colleges could appear on television); Broad. Music, Inc. v. Columbia Broad. Sys., Inc. (“BMI”), 441 U.S. 1, 7–10 (1979) (refusing to apply the per se rule to a blanket copyright license implemented by an association of musical composers).
usually determined the outcome of an antitrust case. Because the per se rule constituted an irrebuttable presumption of illegality, plaintiffs almost always won cases decided under the rule. Defendants, on the other hand, nearly always prevailed in rule of reason cases because it was virtually impossible for plaintiffs to meet the demanding standards for proving specific anticompetitive effects.28

Post-Chicago judges and regulators have become dissatisfied with the stark and confining choice between the per se rule and the rule of reason, and they have begun to advocate intermediate means of analyzing section 1 conduct. Courts and enforcement agencies have proposed a more structured approach to the rule of reason, market share safe harbors, and a variation of the rule of reason called the “quick look.”29 This author has argued that federal courts should not view each of these approaches as fixed categories of antitrust analysis, but as points along a continuum. Instead of becoming captive to labels, fact finders should recognize that their degree of analysis for particular competitive conduct should vary according to the conduct’s likely effect upon consumers.30 In California Dental, the Supreme Court, citing this author and other post-Chicago commentators, endorsed a similar continuum-based approach to antitrust analysis.31

California Dental involved advertising restrictions imposed by the California Dental Association. The restrictions precluded dentists from advertising their prices as “low” or of similar effect.32 The Ninth Circuit had held that the restrictions should be analyzed under the quick-look approach because they made it “more difficult for consumers to find a lower price and for dentists to compete on the basis of price.”33 The Supreme Court concluded that a quick-look analysis was not appropriate because it was not intuitively obvious that the advertising

31 See infra note 38 and accompanying text.
33 Cal. Dental Ass’n v. FTC, 128 F.3d 720, 727 (9th Cir. 1997), vacated, 526 U.S. 756 (1999).
restrictions were likely to have anticompetitive effects.\textsuperscript{34} The restrictions might, for example, have promoted competition by eliminating misleading and unverifiable discount advertising.\textsuperscript{35} The Court recognized, however, that the alternative to quick-look analysis need not necessarily be a full rule of reason analysis. Instead of being divided into discrete categories, Sherman Act analysis should be viewed as a continuum under which courts can engage in a variety of inquiries “meet for the case.”\textsuperscript{36} Rejecting the Ninth Circuit’s decision categorizing the advertising restrictions as presumptively illegal, the Court pointed out that “there is always something of a sliding scale in appraising reasonableness,”\textsuperscript{37} and it concluded that in Sherman Act cases “the quality of proof required should vary with the circumstances.”\textsuperscript{38} Turning to the advertising restrictions at issue, the Court held that “a less quick look was required” and remanded the case “for a fuller consideration of the issue.”\textsuperscript{39}

\textit{California Dental} opened the way to a post-Chicago, post-partisan antitrust era. Neither captive to a Harvard nor a Chicago approach, the Court in \textit{California Dental} freed lower federal courts from the confines of the per se/rule of reason dichotomy. Without the restrictions of labels or fixed categories of analysis, the courts will now be able to tailor their approach precisely to the specific consumer welfare effects of the competitive conduct at issue.

The next three sections explain federal courts’ evolving understanding of the consumer welfare effects of the most significant types of competitive conduct in which American firms are likely to engage in the twenty-first century. The courts now have a sufficient understanding of the likely effects of all such conduct to be confident about its appropriate place on the evolving continuum of antitrust analysis.

\section*{III. The Courts’ Approach to First-Portion Conduct}

The history of antitrust decision-making since 1890 reveals that courts should be confident enough of the likely consumer welfare effects of horizontal restraints, vertical resale restraints, and vertical mergers and joint ventures to place such conduct in the first portion of the antitrust continuum, where its legality can be determined after only an abbreviated factual inquiry.

\begin{quote}
\textsuperscript{34} \textit{Cal. Dental}, 526 U.S. at 756, 771–72, 779–81.
\textsuperscript{35} \textit{Id.} at 771–72.
\textsuperscript{36} \textit{Id.} at 781.
\textsuperscript{37} \textit{Id.} at 780 (quoting \textit{AREEDA}, supra note 29, ¶ 1507, at 402).
\textsuperscript{38} \textit{Id.} at 780, 780 n.15 (citing \textit{AREEDA}, supra note 29, ¶ 1507, at 402; William Kolasky, \textit{Counterpoint: The Department of Justice’s “Stepwise” Approach Imposes Too Heavy a Burden on Parties to Horizontal Agreements}, 12 \textit{ANTITRUST} 41, 43 (1998); Thomas A. Piraino, Jr., \textit{Making Sense of the Rule of Reason: A New Standard for Section 1 of the Sherman Act}, 47 \textit{VAND. L. REV.} 1753, 1771 (1994)).
\textsuperscript{39} \textit{Id.} at 781.
\end{quote}
A. Horizontal Restraints

In 1898, in one of the most discerning antitrust decisions ever written, Judge William Howard Taft—who would later become president of the United States, and after that, the chief justice of the Supreme Court—set forth the “ancillary restraint doctrine” as a means of distinguishing between illegal and permissible horizontal restraints of trade. *United States v. Addyston Pipe & Steel Co.* involved an agreement among six manufacturers of cast iron pipe to fix prices and allocate territories. Judge Taft held that the agreement should be condemned as a naked restraint because the pipe manufacturers had not integrated their resources in any way that could enhance their efficiency. Judge Taft noted, however, that a rule of automatic illegality should not apply to restraints necessary to implement the legitimate purposes of an integrated arrangement among competitors. Such ancillary restraints should be upheld when they are “incident to the main purpose” of a cooperative arrangement designed to promote the parties’ efficiency.

Judge Taft’s formulation had the advantages both of simplicity and of profound economic sense. Consider a hypothetical joint venture between General Motors and Toyota for the operation of a new plant to manufacture a hybrid automobile. Assume that General Motors and Toyota agree jointly to determine the number of automobiles to be produced at the plant. If General Motors and Toyota had not entered into a joint venture, but had simply agreed to coordinate their output at their independently-operated plants, such an arrangement would have had no possibility of benefitting consumers by reducing costs or otherwise enhancing the parties’ efficiency, and it would be void under Judge Taft’s ancillary-restraints analysis. The output arrangement in the hypothetical General Motors/Toyota joint venture, however, has efficiency objectives beneficial to consumers. The joint venture would allow the partners to save manufacturing costs and to share best practices on manufacturing techniques. Thus, the venture promotes consumers’ core goals of lowering prices and ensuring the continued viability of American companies. To operate the joint venture effectively, General Motors and Toyota must agree on plans for the number of cars to be produced at the joint venture plant. Indeed, the joint venture for the production of the new hybrid automobile could not exist without an agreement on output between the parties. Under Judge Taft’s approach, the output arrangement would be upheld as a restraint necessary to achieve the legitimate efficiency-enhancing purposes of the production joint venture.

Harvard Era cases cast doubt on the viability of the ancillary-restraints doctrine, finding various horizontal restraints implemented in connection with joint

---

40 85 F. 271 (6th Cir. 1898), aff’d, 175 U.S. 211 (1899).
41 *Addyston Pipe*, 85 F. at 272–74.
42 *Id.* at 283.
43 *Id.* at 280.
44 In 1984, the FTC approved a production joint venture between General Motors and Toyota that included a similar output arrangement between the parties. *See In re Gen. Motors Corp.*, 103 F.T.C. 374, 374 (1984).
ventures to be per se illegal. However, a series of well-reasoned circuit court decisions during the Chicago Era reinvigorated—after an absence of more than 90 years—the ancillary-restraints analysis set forth in *Addyston Pipe*. Finally, in 2006, in *Texaco, Inc. v. Dagher* ("Dagher"), the Supreme Court acknowledged for the first time the validity of the ancillary-restraints doctrine described by Judge Taft in 1898. At issue in *Dagher* was whether the rule of reason or the per se rule should apply to the pricing activities of a joint venture to which Shell and Texaco had contributed all of their gasoline refining and marketing operations in the western United States. The Ninth Circuit had concluded that the arrangement was per se illegal under the ancillary-restraints doctrine because it was not necessary to promote the legitimate objectives of the marketing joint venture between Shell and Texaco. The Supreme Court, however, concluded that the rule of reason should apply because the joint venture had already eliminated any competition between Shell and Texaco in pricing gasoline in the relevant market. In the course of its decision, the Court left no doubt about the propriety of using the ancillary-restraints doctrine to determine the legality of horizontal restraints:

[The ancillary restraints] doctrine governs the validity of restrictions imposed by a legitimate business collaboration, such as a business association or joint venture, on nonventure activities. Under the doctrine, courts must determine whether the nonventure restriction is a naked restraint on trade, and thus invalid, or one that is ancillary to the legitimate and competitive purposes of the business association, and thus valid. See United States v. Topco Assocs., 405 U.S. 596, 608–09 (1972) (applying per se rule to territorial limitations among an association of regional supermarket chains); United States v. Sealy, Inc., 388 U.S. 350, 357–58 (1967) (applying the per se rule to territorial restrictions implemented among the members of a joint venture among a mattress manufacturer and its distributors).


See Nat’l Bancard Corp. v. VISA U.S.A., 779 F.2d 592, 603–05 (11th Cir. 1986) (allowing an interchange fee among the members of the VISA credit card system); Rothery Storage & Van Co. v. Atlas Van Lines, 792 F.2d 210, 224–27 (D.C. Cir. 1986) (finding that in BMI, the Supreme Court had effectively over-ruled Sealy and Topco and that *Addyston Pipe’s* ancillary restraints analysis “fits . . . exactly” the case in which a van line required its agents to deal exclusively with it); Polk Bros. v. Forest City Enters., 776 F.2d 185, 191 (7th Cir. 1985) (upholding a non-competition covenant between two adjoining stores in a single building).

Id. at 3.


Id. at 7 (citation omitted).
B. Vertical Resale Restraints

The courts’ evolving approach has been to recognize that vertical resale restraints independently implemented by manufacturers should be upheld because they promote consumer welfare, while vertical restraints induced by firms at the resale level are harmful to consumers and therefore should be precluded.

1. The Legality of Restraints Implemented by Manufacturers

During the Early Era and the Harvard Era, the Court found both price and non-price resale restraints implemented by manufacturers to be per se illegal. In 1911, in *Dr. Miles v. John D. Park & Sons*, the Court applied the per se rule to an agreement between a medicine manufacturer and a group of wholesalers to fix the retail prices of certain medicines. In 1967, in *United States v. Arnold, Schwinn & Co.*, the Court extended the per se rule to non-price vertical restrictions imposed by a supplier on its distributors.

During the Chicago Era, however, the Court overruled its two decisions applying the per se rule to vertical restraints and held that both price and non-price restraints should be judged under the rule of reason. In *Continental T.V., Inc. v. GTE Sylvania, Inc.*, the Court reversed its decision in *Schwinn* on the per se illegality of non-price vertical restrictions. Sylvania had required its distributors to sell its televisions only from the locations at which they were franchised. The Court pointed out that, while Sylvania’s location clause limited “intrabrand” competition among Sylvania distributors, it also promoted interbrand competition between Sylvania and other television manufacturers by inducing distributors to provide consumers with additional services. The Court concluded that Sylvania’s territorial restrictions should be judged under the rule of reason rather than the per se rule.

*Sylvania* opened the way to a series of decisions by the lower federal courts upholding resale restraints imposed by manufacturers on their distributors.
In 2007, in *Leegin v. PSKS, Inc.*, the Court overruled its Early Era decision in *Dr. Miles* and held that resale price fixing was not per se illegal. The Court concluded that minimum pricing requirements imposed by manufacturers on retailers should be judged under the rule of reason because they had the potential to benefit consumers. Such requirements could, for example, ensure retailers a sufficient profit margin to afford services desired by consumers. The Court rejected the plaintiff’s argument that the *Dr. Miles* per se approach should be retained under the principle of stare decisis, pointing out that “[f]rom the beginning, the Court has treated the Sherman Act as a common-law statute” and that “[j]ust as the common law adapts to modern understanding and greater experience, so too does the Sherman Act’s prohibition on ‘restraints of trade’ evolve to meet the dynamics of present economic conditions.”

2. The Illegality of Restraints Induced by Distributors

The Supreme Court has made it clear that vertical resale restraints induced by rivals of a distributor adversely affected by such restraints should not be permitted because they are likely to harm consumers. Unlike resale restraints imposed independently by a manufacturer, distributor-induced restraints are not designed to promote interbrand competition, but only to restrict intrabrand competition, either by excluding a rival entirely from the resale market or by making it more difficult for the rival to compete in that market. Such conduct is contrary to consumers’ interests because it is likely to result in higher resale prices and limits on the availability of alternative products and services in the resale market. During the Harvard Era, the Court referred to such conduct as a “group boycott” and consistently found it illegal under the per se rule. In 1959, in *Klor’s Inc. v. Broadway-Hale Stores, Inc.*, a retailer induced its suppliers to cease doing business with a competing retailer, and the Court held the agreement between the retailer and its suppliers to be per se illegal. In 1966, in *United States v. General Motors Corp.*, a group of automobile dealers sought to avoid competition with dealers that were discounting General Motors automobiles, and the group convinced General Motors to prohibit all its dealers from selling automobiles to the discounters. Pointing out that General Motors had merely acquiesced to the anticompetitive demands of the dealers, the Court found that General Motors and its dealers were engaged in a “classic conspiracy” to prevent sales by discounters and held that the conspiracy was per se illegal.

---

60 Id. at 2707–08.
61 Id. at 2714–16.
62 Id. at 2720.
65 Id. at 145–47.
In the Chicago Era case of *Monsanto Co. v. Spray-Rite Service Corp.*, the Supreme Court adopted dicta which appeared to limit the circumstances in which fact finders could infer group boycotts from circumstantial evidence. After receiving complaints from several of its distributors about another distributor’s low prices, Monsanto decided to terminate the price-cutting distributor. The Court stated that, to protect Monsanto from liability for legitimate independent conduct, it should not infer a conspiracy merely from the fact that Monsanto had received the complaints. Although the Court ultimately found Monsanto liable for terminating the price-cutting distributor based on Monsanto’s direct communications with other distributors, several lower federal courts have interpreted *Monsanto*’s dicta on conspiracy as a rationale for dismissing group boycott claims when the evidence of conspiracy is evenly balanced.

The Seventh Circuit’s 2000 decision in *Toys “R” Us, Inc. v. FTC* confirmed that, even after *Monsanto*, fact finders should be able to infer the existence of an illegal group boycott from circumstantial evidence. Toys “R” Us had complained to a group of toy manufacturers about price cutting by warehouse clubs competing with Toys “R” Us. After the complaints, the manufacturers had discontinued selling their toys to the warehouse clubs. The Seventh Circuit affirmed the Federal Trade Commission’s finding of an illegal group boycott among Toys “R” Us and the toy manufacturers, pointing out that it was reasonable to infer a Sherman Act violation from the manufacturers’ response to the complaints.

**C. Vertical Mergers and Joint Ventures**

The federal courts have increasingly recognized that, in most cases, vertical mergers and joint ventures should be permitted because they pose no threat to consumer welfare. During the Harvard Era, federal courts precluded several vertical mergers among companies with a customer-supplier relationship on the grounds that the transaction would give the merged entity both the incentive and the ability to foreclose its competitors from access to a product or service...

---

67 Id. at 757.
68 The Court concluded that “something more than evidence of complaints is needed. There must be evidence that tends to exclude the possibility that the manufacturer and non-terminated distributors were acting independently.” Id. at 764.
69 Id. at 765.
71 221 F.3d 928, 934–35 (7th Cir. 2000).
72 Id. at 936.
73 Id. at 935 (stating that the case was “a modern equivalent of the old Interstate Circuit decision, in which the Supreme Court had inferred a Sherman Act violation from circumstantial evidence of a conspiracy involving actors at two levels of the distribution chain”).
controlled by the merged entity. For example in *Ford Motor Co. v. United States*,\(^{74}\) the Supreme Court precluded Ford’s acquisition of Autolite’s spark plug business on the grounds that Ford, as a potential competitor in the spark plug manufacturing market, had an incentive to foreclose other spark plug manufacturers from access to the original equipment market for automobiles.\(^{75}\) During the Chicago Era, however, the government failed to bring any enforcement actions against vertical mergers, and the courts generally upheld joint ventures among parties that were not actual or potential competitors.\(^{76}\)

The enforcement agencies during the Clinton administration established a middle ground between the federal courts’ hostility to vertical mergers in the Harvard Era and their permissive approach to vertical mergers and joint ventures during the Chicago Era. The Clinton agencies recognized that vertical mergers and joint ventures should be permitted in most cases but that, in special circumstances, they can be detrimental to consumers. They only attacked vertical mergers or joint ventures when the parties had a customer-supplier relationship that gave the relevant entity the incentive and the ability to foreclose rivals from access to a resource essential for effective competition in a particular market. In such cases, the parties to the relevant transactions could exclude firms completely from the relevant market, thereby limiting the availability of alternative products and services and allowing the remaining firms to raise prices. Even in such cases, however, agencies recognized that they did not need to preclude a vertical merger or joint venture entirely, but could allow the transaction to proceed subject to consent decrees requiring the relevant entity to give its competitors non-discriminatory access to the essential product or service.\(^{77}\)

\(^{74}\) 405 U.S. 562, 575 (1972).

\(^{75}\) *Id.* at 568, 574. *See also* Brown Shoe Co. v. United States, 370 U.S. 294, 331–32 (1962) (prohibiting a shoe manufacturer from acquiring a shoe retailer based on the manufacturer’s incentive and ability to foreclose competing manufacturers from access to the acquired retailer’s outlets).

\(^{76}\) *See, e.g.*, United States v. FCC, 652 F.2d 72, 96–101 (D.C. Cir. 1980) (en banc) (holding joint venture to construct a communications satellite network does not violate antitrust laws if the parents were not actual or potential competitors in the relevant market). Chicago Era antitrust commentators concluded that joint ventures among non-competitors pose no antitrust risk at all. *See, e.g.*, Robert Pitofsky, *A Framework for Antitrust Analysis of Joint Ventures*, 54 ANTITRUST L.J. 893, 897–99 (1985) (“It is possible to challenge a joint venture on the basis of loss of potential competition, but the plaintiff’s burden is a formidable one.”).

\(^{77}\) *See* Proposed Final Judgment and Competitive Impact Statement, United States v. MCI Commc’ns Corp., 59 Fed. Reg. 33,009, 33,012, 33,018 (June 27, 1994) (providing notice of a consent decree for the joint venture between MCI Communications Corp. and British Telecommunications, which required the venture to give MCI’s competitors open access to its monopoly telephone network in the United Kingdom); Proposed Final Judgment and Competitive Impact Statement, United States v. Tele-Comm’ns, Inc., 59 Fed. Reg. 24,723, 24,727 (May 12, 1994) (providing notice of a consent decree allowing the merger of two companies, each of which held substantial shares of the video
IV. THE COURTS’ APPROACH TO SECOND-PARTY CONDUCT

During the last few decades, the federal courts have become willing to consider an ever-increasing number of factors bearing on the likely adverse and beneficial effects of horizontal mergers and joint ventures. With so many factors in play, the analysis of horizontal mergers and joint ventures has become more complex. Unfortunately, the courts and enforcement agencies have not yet been able to assign priorities and evidentiary weights to the relevant factors necessary to balance the adverse and beneficial effects of horizontal mergers and joint ventures. As a consequence, American firms have become less certain of the types of transactions in which they can engage.

During the Harvard Era, the Supreme Court focused primarily on the market power of horizontal merger partners. The Court applied a presumption of illegality to mergers among competitors with collective market shares as low as 8 percent, and it refused to recognize efficiencies or factors mitigating the defendants’ market power as legitimate defenses in such cases. This combination was deadly for Harvard Era defendants in horizontal merger cases. As Justice Potter Stewart pointed out in 1966, “the sole consistency” in the Supreme Court’s approach to such cases was that “the Government always wins.”

By the mid-1970s, Chicago School scholars began to criticize the Harvard Era’s low market share presumption of illegality for horizontal mergers. The presumption, these scholars asserted, precluded many efficiency-enhancing transactions. Responding to the concerns of the Chicago School, the Supreme Court, in the mid-1970s, began to open up merger analysis to consideration of factors other than market share concentration statistics. In 1974, in 
United States v. General Dynamics Corp.
, the Court, in a marked departure from its Harvard Era antitrust jurisprudence, held that a merger could not be deemed illegal simply because the defendants held high market shares. The Court expressed a willingness to consider conditions that might mitigate the future market power of the merging

programming and cable markets, and requiring the merged entity not to discriminate against other providers of programming or other operators of cable systems).


79 See FTC v. Proctor & Gamble Co., 386 U.S. 568, 580 (1967) (“Possible economies cannot be used as a defense to illegality” in merger cases.); United States v. Phila. Nat’l Bank, 374 U.S. 321, 371 (1963) (“[A] merger the effect of which ‘may be substantially to lessen competition’ is not saved because, on some ultimate reckoning of social or economic debits or credits, it may be deemed beneficial.”).

80 Von’s Grocery Co., 384 U.S. at 301.

81 See, e.g., RICHARD A. POSNER, ANTITRUST LAW: AN ECONOMIC PERSPECTIVE 112 (1976) (arguing that there should be no presumption of illegality against a merger when the combined share of the four largest firms in the relevant market is less than 60 percent).

parties. General Dynamics opened the way to a more sophisticated approach to horizontal merger and joint venture analysis that could consider factors other than the parties’ market shares.

Economists have recognized that horizontal mergers and joint ventures may achieve economies of scale, synergies, cost savings, or other efficiencies that will allow parties to offer lower prices or more innovative products to consumers after the completion of the transaction. Despite the opening provided by General Dynamics, however, many courts have been unwilling to consider such efficiencies as a defense to an otherwise illegal merger or joint venture. Some commentators have ascribed the reluctance to recognize an efficiencies defense to the difficulty of confirming the legitimacy of defendants’ efficiency claims.

Post-Chicago antitrust regulators have been more willing to recognize efficiency defenses in horizontal merger and joint venture cases. The antitrust enforcement agencies have come a long way since 1962, when they argued in a Supreme Court case that efficiencies should count against a merger. As a commissioner of the Federal Trade Commission has explained, the agencies’ Merger Guidelines “have become progressively more hospitable to efficiency defenses.” In 1997, under the Clinton administration, the FTC and Department of Justice substantially revised the efficiencies section of the 1992 Merger Guidelines, reflecting FTC Chairman Robert Pitofsky’s concern that the previous standards were too restrictive. The 1997 revisions state that “[t]he Agency will not challenge a merger if cognizable efficiencies are of a character and magnitude such

83 Id. at 501–02 (stating that, in a merger between two coal companies, their past market shares did not give a clear picture of their future market shares because the companies’ long-term coal supply contracts might prevent them from competing in the future against other companies without such restraints).
84 See FTC v. H.J. Heinz Co., 246 F.3d 708, 720 (D.C. Cir. 2001) (citing 4A PHILLIP E. AREEDA et al., ANTITRUST LAW ¶ 970, at 22–25 (rev. ed. 1998) (stating that commentators have recognized that “a merger’s primary benefit to the economy is its potential to generate efficiencies”)).
85 As a former DOJ official has stated, “Although the Merger Guidelines recognize [the efficiencies defense], . . . that recognition has not been translated into decisions by either the agencies or the [lower federal] courts to accept an efficiency claim as a counterweight to the anticompetitive effects of a merger.” Joseph Kattan, The Role of Efficiency Considerations in the Federal Trade Commission’s Antitrust Analysis, 64 ANTITRUST L.J. 613, 616 (1996) (footnote omitted).
86 See, e.g., Robert Pitofsky, Antitrust Policy in a Clinton Administration, 62 ANTITRUST L.J. 217, 221 (1993) (“If long-term U.S. economic interests turn on improvements in productivity and innovation, it may be time to treat assertions of efficiency in defense of a transaction in a less grudging way.”).
87 Timothy J. Muris, The Government and Merger Efficiencies: Still Hostile After All These Years, 7 GEO. MASON L. REV. 729, 730 (1999) (“This argument [against merger efficiencies] was made at the Federal Trade Commission as late as the mid-1970’s.”).
that the merger is not likely to be anti-competitive in any relevant market.”^89 The agencies will consider whether the efficiencies “likely would be sufficient to reverse the merger’s potential to harm consumers in the relevant market.”^90

In 1984, the FTC used an efficiency argument to uphold a joint venture between Toyota and General Motors for the production of a compact automobile at a plant in Fremont, California. The FTC recognized that such a joint venture between what were then the largest and the third largest automobile manufacturers in the world could have an adverse effect on competition.91 Nevertheless, the FTC concluded that the adverse competitive effects of the venture were outweighed by General Motors’ opportunity to learn more efficient Japanese manufacturing techniques.92

Now that courts and enforcement agencies have sanctioned the consideration of so many factors in horizontal merger and joint venture cases, they must develop a practical means of prioritizing and weighing such factors. Section X, infra, proposes an approach that would effectively distinguish between those horizontal mergers and joint ventures that should be permitted and those that should be precluded.

V. THE COURTS’ APPROACH TO THIRD-PARTY CONDUCT

A. Protecting Monopolists’ Legitimate Competitive Conduct

Section 2 of the Sherman Act, which prohibits monopolization and attempts to monopolize, has been called “perhaps the most amorphous aspect of antitrust law.”^93 During the entire history of antitrust enforcement, there have been “significant differences of opinion . . . about the appropriate approach to Section 2.”^94 Indeed, “how to evaluate single-firm conduct under Section 2 poses among the most difficult questions in antitrust law.”^95 The confusion in section 2 cases stems from courts’ difficulty in distinguishing a firm’s abuse of its monopoly power from legitimate competitive conduct beneficial to consumers.

Harvard Era cases took an unduly harsh approach in monopoly cases, punishing firms for aggressive competitive conduct that arguably benefitted consumers. Indeed, once a Harvard Era defendant was found to possess monopoly power, any conduct having the purpose or effect of protecting or increasing that

^90 Id.
^92 Id.
^94 Id. at 43.
^95 Id.
power was deemed illegal. It was sufficient that a defendant evidenced a general intent to attain or maintain its monopoly power, i.e., that it chose to do the acts that led to the establishment or perpetuation of its monopoly. 96 Commentators have pointed out that such a standard “is hardly a requirement at all . . . . It is almost inconceivable that [a monopolist] can possess [monopoly] power . . . without taking some volitional act that may fairly be characterized as an exercise of its power . . . .”97 One commentator has argued that the Harvard Era cases instituted a regime of ‘soft competition,’ which made it dangerous to be too successful and for the successful to compete aggressively. General Motors, for example, with a 50 per cent share of the automobile market, did not dare lower its prices, as that would increase its market share and surely invite a government antitrust suit.98

During the Chicago Era, the courts began to question the Harvard Era’s inference of a section 2 violation merely from a monopolist’s general intent to maintain its market power. The Chicago School assumed that firms usually acquire monopoly power because of their ability to provide consumers with superior products at low prices.99 Thus, punishing monopolists simply for exercising their market power would discourage firms from competing aggressively to grow in ways that meet consumers’ needs.100 In 1966, in United States v. Grinnell Corp.,101 the Supreme Court stated that section 2 of the Sherman Act should only prohibit “the willful acquisition or maintenance of . . . [monopoly] power as distinguished from growth or development as a consequence of a superior product, business acumen or historical accident.”102 Following Grinnell, the lower federal courts generally concluded that firms should not be punished simply for acquiring


99 See Hovenkamp, supra note 10, at 246 (“An antitrust policy of promoting efficiency would at least passively encourage firms to become large, perhaps by permitting mergers or internal growth that achieved production economies. . . .”).

100 For example, Chicago School commentators argued that it should not be illegal for monopolists to refuse to deal with their actual or potential rivals. See, e.g., POSNER, supra note 81, at 242–44 (arguing that unconditioned refusals to deal by dominant firms should never be violations of section 2 of the Sherman Act).


102 Id. at 570–71.
monopoly power, or even for exercising it, but only for using their monopoly power in an abusive way. No case has yet articulated an objective standard for confirming which types of monopoly power should be permitted or precluded. Section XI, infra, proposes a standard that would encourage monopolists to compete aggressively while protecting consumers from conduct that unduly limits competition.

B. The Evolution of the Essential Facilities Doctrine

In one area of monopoly conduct, the courts have sent a clear and consistent message on the type of behavior they will not tolerate. In all three antitrust eras, the courts have precluded monopolists from perpetuating their market power in a monopolized market or extending that power to an adjacent market by denying a competitor access to a product, facility, or other resource that is essential for effective competition in either of those markets.

In the Early Era, the Supreme Court established the basis for the “essential facilities doctrine,” under which monopolists could be required to grant their competitors access to a resource essential to effective competition in the relevant market. During both the Harvard and Chicago Eras, the Supreme Court continued to find monopolists liable under section 2 of the Sherman Act for refusing to make such resources available to all qualified parties on reasonable terms. In Aspen Skiing Co. v. Aspen Highlands Skiing Corp., the Supreme Court concluded that a ski resort violated section 2 when it refused to continue to cooperate with a competing resort in marketing a multi-mountain ticket that allowed skiers to use either resort’s slopes, and in Eastman Kodak Co. v. Image Technical Services Inc., the Court concluded that Kodak, which manufactured photocopiers and related service parts, violated section 2 when it refused to continue to sell replacement parts to independent service organizations servicing its photocopiers.

---

103 See Foremost Pro Color, Inc. v. Eastman Kodak Co., 703 F.2d 534, 541–43 (9th Cir. 1983) (allowing Kodak to bundle camera, film, and developing process); Cal. Computer Prods., Inc. v. IBM Corp., 613 F.2d 727, 743 (9th Cir. 1979) (finding that IBM did not engage in predatory pricing when it sold products at a level above marginal cost).
104 In United States v. Terminal Railroad Association, the Court required an association of railroad companies that controlled the only railroad bridges crossing the Mississippi River at St. Louis to provide equal access to the bridges to all other railroads. 224 U.S. 383, 411–12 (1912).
106 Id. at 603, 611.
108 Id. at 483. For other essential facilities cases, see MCI Commc’ns Corp. v. AT&T Co., 708 F.2d 1081, 1131–33 (7th Cir. 1983) (requiring AT&T, which at the time of the suit still owned the local Bell telephone systems, to allow MCI, its competitor in the long-distance market, to interconnect its long-distance lines with AT&T’s local lines); Otter Tail Power Co. v. United States, 410 U.S. 366, 377–82 (1973) (requiring an electric utility to
One Chicago Era Supreme Court case, however, limited the scope of the essential facilities doctrine to cases in which a monopolist had already made the relevant product or service available to third parties. In Verizon Communications Inc. v. Trinko, the Court concluded that Verizon, the local telephone company serving New York State, did not violate section 2 when it refused to allow its potential rivals to use its local network. The Court distinguished its earlier essential facilities decisions on the grounds that in each of those cases, the defendant refused to continue to provide a service that it had previously made available. In Trinko, however, the plaintiffs sought to compel the defendant to provide a service it had not previously provided to rivals.

In 1998, in United States v. Microsoft Corp., the District of Columbia Circuit Court, after finding that Microsoft held a monopoly in the market for operating systems used in personal computers (“ PCs”), held that Microsoft violated section 2 by imposing license restrictions on PC manufacturers that were designed to protect that monopoly. The court found that the license restrictions were designed to keep “rival [Internet] browsers [such as Netscape’s “Navigator”] from gaining the critical mass of users necessary to attract developer attention away from Windows as the platform for software development.” Microsoft’s licenses prohibited the PC manufacturers from altering the Windows desktop to remove icons for Microsoft’s Internet browser. This restriction prevented PC manufacturers from preinstalling other Internet browsers on Windows because preinstalling more than one browser would lead to confusion among users, significantly increasing the manufacturers’ support costs. In addition, Microsoft precluded PC manufacturers from adding icons or folders to the desktop that would identify or promote Internet browsers other than Microsoft’s. In effect, by

provide wholesale power to cities that had established their own power companies); Lorain Journal Co. v. United States, 342 U.S. 143, 152–57 (1951) (requiring the only newspaper in Lorain, Ohio, to sell advertising to the patrons of a radio station that competed with the newspaper in the local media market).


The Court found that the Telecommunications Act of 1996 already imposed duties upon Verizon with respect to sharing its local telephone network with potential competitors and that the “Act’s extensive provision for access makes it unnecessary to impose a judicial doctrine of forced access.” Id. at 405, 411.

In those cases, the defendants’ refusal to deal as they had in the past suggested an anticompetitive purpose for their conduct. Id. at 409–10.

The Court found that, under such circumstances, “the defendant’s prior conduct sheds no light upon the motivation of its refusal to deal—upon whether its [practices] were prompted not by competitive zeal but by anticompetitive malice.” Id. at 409.

253 F.3d 34 (D.C. Cir. 2001).

Id. at 51.

Id. at 60.

Id.

Id. at 61.

Id.

Id. at 62.
imposing these license restrictions, Microsoft was denying competing browsers access to an essential resource—the Windows operating system—without which other browser manufacturers could not compete in the operating system market.

VI. AN INTEGRATED ANTITRUST COMMON LAW FOR THE TWENTY-FIRST CENTURY

The Supreme Court’s decision in *California Dental* will allow the federal courts to adopt an integrated common law of antitrust appropriate for regulating twenty-first century competitive conduct. Instead of confining their approach to fixed categories of analysis such as the rule of reason or the per se rule, courts and agencies can now adopt an inquiry “meet for [each] case,” as described in *California Dental*. ¹²⁰ The federal courts should adopt a unified approach that arrays the major types of competitive conduct along a continuum according to the amount of inquiry necessary to determine their likely effect on consumers. The courts could then apply presumptions of legality or illegality in accordance with the location of the relevant conduct on the continuum. The courts and agencies should simply presume that the most likely competitive effect did, in fact, occur and shift the burden of proof to the party attempting to demonstrate that the actual outcome was contrary to normal expectations. ¹²¹ Rather than viewing the various types of competitive conduct in isolation, this continuum-based approach would group together conduct, the legality of which could be determined in a similar manner.

Classifying particular competitive conduct in the appropriate place on the continuum of antitrust analysis should not be difficult, as the courts can draw upon their 119-year history of analyzing the likely consumer welfare effects of particular types of conduct. When the courts can be confident that the conduct at issue almost always adversely affects consumers, they can indulge in a strong presumption of illegality, which a defendant could only rebut by demonstrating special circumstances that negate the conduct’s likely adverse effect. Conversely, in cases where the long history of antitrust decision making has revealed that the relevant conduct usually benefits consumers, the courts can indulge in a strong presumption of legality, which the plaintiff could only rebut by proving special circumstances that negate the conduct’s likely beneficial effect.

By using presumptions, courts and agencies can achieve clarity without sacrificing their search for the right economic answer. Judges and juries lack the sophisticated judgment and experience necessary to decide complex economic issues. As Justice Breyer pointed out in the *Leegin* case, “One cannot fairly expect judges and juries . . . to apply complex economic criteria without making a


¹²¹ Eleanor Fox has argued that courts should “restore . . . use of presumptions when it makes economic or administrative sense and the party with knowledge of all of the facts gets a full and fair chance to rebut.” *Perspectives on the Future Direction of Antitrust*, 22 *Antitrust* 21, 23 (2008).
considerable number of mistakes.” Instead of requiring fact finders to delve into economic issues de novo in individual cases, the proposed approach will allow them to make certain assumptions based upon the federal judiciary’s history in considering the consumer welfare effects of various types of competitive conduct. Under such an approach, the courts will only have to make judgments about the economic effects of defendants’ competitive conduct in the small number of cases in which the likely effect of the relevant conduct on consumers, even after 119 years of antitrust decision-making, remains unclear.

Such an approach will conserve judicial resources, leaving a minimal role for the vagaries of economic fact finding in the large number of cases in which the likely effects of the relevant conduct on consumers are obvious. The approach will also clarify the rules of competitive conduct for the types of behavior with the greatest impact on consumers, thus encouraging American firms to pursue activities that benefit consumers and deterring them from engaging in conduct that harms consumers.

Under such an approach, the courts could construct the following continuum of antitrust analysis.

A. First-Portion Conduct

1. Horizontal Restraints of Trade

Horizontal agreements among competitors to fix prices, output, customers, or territories would be presumptively illegal because they harm consumers without the possibility of any offsetting efficiency benefits. A defendant would only be able to rebut the presumption of illegality by demonstrating that a horizontal restraint was necessary to ensure the effectiveness of a legitimate joint venture. In such cases, a horizontal restraint on price, output, customers, or territories could be justified by cost savings or other efficiencies beneficial to consumers.

2. Vertical Resale Restraints

Vertical resale restraints imposed by manufacturers on their distributors would be presumptively legal because they almost always benefit consumers by promoting competition among different brands. A plaintiff would only be able to rebut the presumption of legality for a vertical resale restraint by proving that the restraint was not imposed independently by a manufacturer, but was simply designed to placate rival distributors’ concerns about competition from a more efficient rival. In such cases, the restraint could harm consumers by excluding firms from a distribution market or by deterring them from aggressively competing in that market.

---

3. Vertical Mergers and Joint Ventures

Vertical mergers and joint ventures do not usually limit competition and thus would be presumptively legal. A plaintiff would only be able to rebut the presumption of legality for a vertical merger or joint venture by proving that the transaction gave the parties to the merger or joint venture the ability to exclude rivals from a resource essential to effective competition in a particular market. In such cases, the transaction could harm consumers by excluding firms completely from the relevant market.

B. Second-Portion Conduct

Courts and agencies would balance the likely anticompetitive effects of horizontal mergers and joint ventures against their potential efficiencies. In order to simplify their balancing analysis, courts and agencies would require a plaintiff to prove in its prima facie case that the parties’ collective market share exceeded a particular threshold. In cases in which plaintiffs met this burden, a defendant could rebut by demonstrating that ease of entry in the relevant market mitigated the parties’ market power or that certain efficiencies of the transaction, such as cost savings, were sufficient to outweigh any anticompetitive effects.

C. Third-Portion Conduct

To protect monopolists from liability for aggressive competitive conduct beneficial to consumers, a plaintiff would have the initial burden of proving not only a defendant’s monopoly power, but also its misuse of that power in a manner adverse to consumers. A plaintiff could meet that burden by demonstrating that a monopolist’s conduct made no economic sense other than as an attempt to perpetuate or extend its market power beyond the period in which it would have been overthrown by the natural workings of the marketplace.

VII. APPLYING THE PROPOSED APPROACH TO HORIZONTAL RESTRAINTS OF TRADE

Since 1898, the Supreme Court and lower federal courts have consistently recognized that naked horizontal agreements among competitors to fix prices or allocate customers or territories almost always restrict competition without any offsetting efficiency benefit to consumers.123 Such cartels are incapable of producing any economic benefit because they involve no integration of the parties’

---

123 Robert Pitofsky has pointed out: “It has long been a consensus that so-called hardcore cartel behavior involving direct rivals—agreements to fix price, curtail output, rig or rotate bids, or divide product or geographic markets—seriously injures consumers and the competitive process and cannot be justified by alleged redeeming virtues.” Pitofsky, supra note 18, at 172.
resources. Horizontal cartels raise prices, reduce output, and limit consumers’ range of choices. Indeed, such restrictions on interfirm rivalry strike at the heart of the open competitive system that the antitrust laws were designed to protect.124

Since horizontal restraints almost always harm consumers, courts should presume their illegality and place the burden on defendants to demonstrate their legitimacy. The ancillary-restraints doctrine, first set forth by Judge Taft in 1898 in Addyston Pipe,125 revitalized by the lower federal courts during the Chicago Era,126 and finally explicitly recognized by the Supreme Court in its 2006 Dagher decision,127 provides an effective means of identifying a defendant’s rebuttal burden in horizontal restraint cases. As the Court pointed out in Dagher, naked restraints of trade unrelated to a legitimate business collaboration should be invalid, whereas restraints “ancillary to the legitimate and competitive purposes” of a business association should be permitted.128 Thus a defendant should only be able to rebut the presumption of illegality for horizontal restraints by proving that a restraint was ancillary to a legitimate efficiency-enhancing joint venture. Only in such cases could a horizontal restraint benefit consumers by reducing prices, expanding alternative products and services, or promoting the continued viability of American companies.

Placing such a rebuttal burden on a defendant is fair because it is the party with access to the evidence necessary to prove that a horizontal restraint was reasonably necessary to promote the efficiency objectives of a joint venture. Such a high rebuttal burden should simplify the courts’ analysis of horizontal restraints and deter competitors from conspiring to limit competition. Only in special circumstances where competitors have entered into a joint venture beneficial to consumers will they be able to argue that a court should uphold a horizontal restraint affecting prices, customers, or territories.

This author argued in a series of articles beginning in 1991 that courts should require a defendant to prove that a particular restraint was “reasonably necessary” to promote the legitimate purposes of a joint venture.129 Restraints among the parties to joint ventures that are necessary for the effectiveness of a legitimate joint venture should not have any anticompetitive effect beyond those naturally resulting

124 Although the federal courts have taken a less aggressive approach in many antitrust areas during the last three decades, they have never wavered in their condemnation of cartels. Justice Scalia (rarely an advocate of aggressive antitrust enforcement) has described cartels as “the supreme evil of antitrust.” Verizon Commc’ns. Inc. v. Trinko, 540 U.S. 398, 408 (2004).
126 See supra note 46 and accompanying text.
128 Id.
from the joint venture. If the joint venture is legal, any restraints necessary for its existence should be no less permissible. On the other hand, restraints unrelated to a joint venture’s objectives or that are broader than necessary to accomplish those objectives should be precluded because they have no legitimate purpose or effect other than to limit competition.

Under such a standard, it should not be difficult for fact finders to distinguish between naked and ancillary horizontal restraints. The legality of a particular restraint should be readily apparent from the objectives of a joint venture. Consider a hypothetical research and development joint venture between Ford and General Motors for a new battery system that allows an electric automobile to travel 500 miles without having to be recharged. Assume that Ford and General Motors agree to fix the price at which they will sell the new battery system to third parties. It would not be difficult for a fact finder to confirm the illegality of such an arrangement under the proposed approach. As a horizontal restraint, the price-fixing arrangement would be presumptively illegal, and Ford and General Motors would have the burden of proving that the arrangement was reasonably necessary to promote the legitimate objectives of their joint venture. It should be readily apparent to any fact finder that the parties would be unable to meet their rebuttal burden. To develop the new battery system, the parties would have no need to limit competition in their eventual sale of the system. The only effect of the price-fixing arrangement between Ford and General Motors would be to raise the prices of the new battery system developed by the joint venture, and a court could appropriately find the arrangement illegal after only a minimal inquiry.

Also consider certain hypothetical restraints implemented by Continental Airlines and United Airlines in connection with a marketing joint venture. Assume that the alliance would allow the airlines to sell tickets on each other’s flights, combine their frequent flier programs, and reduce costs by coordinating flight schedules and routes. Assume further that, in connection with the marketing alliance, Continental and United agree to eliminate competing flights on the Chicago to New York route. Consumers in Chicago and New York file a class action against Continental and United, alleging that the flight reductions constitute an illegal horizontal restraint under section 1 of the Sherman Act.

Under the proposed approach, the plaintiffs would be entitled to a presumption that, as horizontal output restraints, the route reductions were illegal. However, United and Continental should be able to rebut the presumption by proving that the flight reductions were ancillary to the legitimate objectives of their marketing alliance. The alliance was designed, among other things, to allow United

---

130 In June, 2008, Continental and United agreed to apply for antitrust immunity for a similar alliance that would operate outside the United States and would allow them to share revenue, coordinate ticket prices and jointly operate flights. See Susan Carey, UAL Pursues Continental Alliance, WALL ST. J., May 31–June 1, 2008, at A4. Within the United States, however, the parties proposed a more limited “code sharing” joint venture that would allow them to combine their frequent flier programs and to route customers over each other’s flights but would not allow them to share revenue or to coordinate fares or flight schedules. See id.
and Continental to reduce their costs by coordinating flight schedules and routes. Thus it should be readily apparent to a fact finder that the parties’ decision to eliminate overlapping flights between Chicago and New York was necessary to implement the joint venture’s legitimate efficiency objectives.

VIII. APPLYING THE PROPOSED APPROACH TO VERTICAL RESALE RESTRAINTS

The likely beneficial effects of vertical restraints are just as clear as the likely adverse effects of horizontal restraints. Vertical resale restraints are usually designed for a procompetitive purpose and have only a minimal anticompetitive effect. As the Supreme Court recognized in its Chicago Era decisions in *Sylvania* and *Leegin*, vertical resale restraints only affect “intrabrand competition,” that is, competition among dealers in the resale of the manufacturer’s own product.\textsuperscript{131} Such restraints may increase resale prices for the manufacturer’s product, but they do not adversely affect interbrand competition. Indeed, their only impact on interbrand competition is beneficial. The *Sylvania* Court emphasized, for example, that non-price vertical restraints promote interbrand competition by inducing “retailers to engage in promotional activities or to provide service and repair facilities” desired by consumers, and the *Leegin* Court pointed out that minimum resale pricing requirements ensure retailers a sufficient profit margin to afford customer services.\textsuperscript{132}

Because vertical resale restraints almost always benefit consumers, they should be afforded a strong presumption of legality. To rebut the presumption of legality for a vertical restraint, a plaintiff should have the burden of demonstrating certain special circumstances in which an ostensibly vertical restraint has been converted to horizontal conduct detrimental to consumers. If the impetus for a vertical restraint comes not from a manufacturer, but from one or more distributors attempting to restrict competition at the resale level, the effect of the restraint will be horizontal rather than vertical, and consumers will be harmed because a distributor can be excluded from the market, or its price cutting restrained, simply due to a rival’s fear of its aggressive competition.

A court’s goal in cases of induced discrimination should therefore be to determine whether the discrimination was initiated independently by a supplier or induced by one or more buyers in order to disadvantage a rival. As the Supreme Court recognized in *Monsanto*, discriminatory actions against buyers are more likely to be initiated by suppliers for legitimate reasons than to be induced by rival retailers for anticompetitive purposes.\textsuperscript{133} Competing firms have a natural incentive to enter into horizontal conspiracies in restraint of trade because such arrangements are likely to raise the parties’ prices and increase their profit margins. Vertical conspiracies between a manufacturer and its distributors, on the other hand, are

\textsuperscript{132} See *Sylvania*, 433 U.S. at 54–55; *Leegin*, 127 S. Ct. 2705 at 2710–11.
less likely to occur because a manufacturer has a natural incentive to avoid any vertical restraints that may antagonize its own resellers. It is thus appropriate for the courts to place a heavy burden on plaintiffs in vertical restraint cases to prove that the most likely event (i.e., independent conduct by the manufacturer) did not in fact occur.

Thus, to rebut the presumption of legality for a vertical restraint, a plaintiff should be required to prove that, but for the anticompetitive demands of one or more buyers, the seller would not have imposed the relevant restraint on the plaintiff. The but-for test would focus a fact finder’s attention on how a manufacturer would have acted in the absence of complaints from its distributors about a rival’s competitive conduct. If the manufacturer would have implemented the relevant vertical restraint on the rival in any event, it should not be liable under the Sherman Act.

Such a high standard of proof would be consistent with Monsanto’s recognition of the importance of protecting manufacturers engaged in legitimate independent conduct. Only when a plaintiff meets the but-for standard could a fact finder infer that a supplier and one or more resellers have entered into an illegal conspiracy to restrain competition at the resale level. Under such circumstances, the manufacturer would have joined the dealers in a group boycott of their rivals. Group boycotts, like cartels, almost always have no purpose other than to restrict competition. Indeed, such restraints often have the potential to completely exclude a disadvantaged distributor from the relevant market.

Two examples illustrate how effective the proposed approach would be in identifying cases in which buyers illegally induce suppliers to discriminate against their rivals. First consider a case in which “BabyTech,” a maker of strollers, baby carriers and other infant goods, is supplying such goods to a Wal-Mart store located in the suburbs of Toledo, Ohio. BabyTech sells to “BabyBuys,” a small retailer of goods for infants located in downtown Toledo. BabyBuys sells its infant goods below BabyTech’s suggested retail price and undercuts Wal-Mart, which is observing the suggested price. Concerned about BabyBuys’ low retail prices, Wal-Mart informs BabyTech that, unless it stops selling to BabyBuys within 90 days, Wal-Mart will seek other suppliers. BabyTech acquiesces to Wal-Mart’s demands and ends its relationship with BabyBuys.134

Under the proposed approach, BabyBuys should be successful in rebutting the presumption of legality for BabyTech’s vertical conduct. BabyBuys should be able to demonstrate that, but for Wal-Mart’s threat to cease doing business with BabyTech, BabyTech would not have terminated its relationship with BabyBuys. Although it was against BabyTech’s independent business interest to forego its

---

134 The facts of the hypothetical are based on a recent case in which a federal district court denied a motion to dismiss a complaint alleging an illegal conspiracy among Babies “R” Us, the largest U.S. baby goods retailer, and the seven manufacturers of infant goods to enforce minimum retail prices on other retailers who were undercutting the prices charged by Babies “R” Us. See BabyAge.com, Inc. v. Babies “R” Us, 558 F. Supp. 2d 575, 576 (E.D. Pa. 2008).
business with a good customer, BabyTech could not afford to ignore the demands of a customer as large as Wal-Mart. As one commentator has explained, “those who want to have a good relationship with Wal-Mart have to be diplomatic; with $256 billion in sales . . . , the company has unprecedented power and offers a chance for big sales that manufacturers cannot ignore.” 135 Under the proposed approach, a court would require BabyTech to continue to deal with BabyBuys, thereby ensuring consumers’ access to the lowest price outlet for infants’ goods in Toledo.

Second, consider a General Motors dealer in the Buffalo suburbs (“Al’s Chevrolet”) whose business has been declining for several years due to service problems. A nearby General Motors dealer (“Hank’s Chevrolet”) has been receiving complaints from customers who have decided to stop patronizing Al’s Chevrolet because of its service problems. Hank’s Chevrolet informs General Motors of the customer complaints and suggests that it may wish to consider revoking Al’s Chevrolet’s franchise. General Motors investigates the situation and decides to terminate Al’s Chevrolet as a dealer. Under these circumstances, Al’s Chevrolet should not be able to prevail in a section 1 action against General Motors and Hank’s Chevrolet. Although General Motors and Hank’s Chevrolet may have shared a common commitment to terminate Al’s Chevrolet, Al’s Chevrolet could not meet its burden to prove that but for Hank Chevrolet’s complaints, General Motors would not have proceeded with the termination. If General Motors had received the service complaints directly from customers rather than from Hank’s Chevrolet, it still likely would have terminated Al’s Chevrolet on its own. By removing a poorly performing dealer, General Motors would be able to encourage its remaining dealers to provide consumers with better services, and at the same time enhance its effectiveness in competing against other automobile manufacturers. Thus the plaintiff’s burden of meeting the “but for” standard would protect General Motors against any liability for its own pro-competitive actions.

IX. APPLYING THE PROPOSED APPROACH TO VERTICAL Mergers AND JOINT VENTURES

The analysis of vertical mergers and joint ventures should involve no greater inquiry by fact finders than the analysis of horizontal and vertical restraints of trade. Like vertical restraints of trade, vertical mergers and joint ventures rarely harm consumers and thus should be permitted in most circumstances. Because they do not reduce the number of competing firms in the relevant market, mergers or joint ventures among noncompetitors usually do not give the resulting entity a greater ability to raise prices or otherwise restrict competition to consumers’ detriment. In order to recognize the likely benign effect of vertical mergers and joint ventures, courts and agencies should presume the legality of all such

transactions, regardless of the parties’ market shares. A plaintiff should have the heavy burden to rebut the presumption of legality by proving special circumstances that make a vertical merger or joint venture anticompetitive.

During the Harvard Era, the Supreme Court precluded several vertical mergers among parties with a customer/supplier relationship on the grounds that they would give the merged entity both the incentive and the ability to foreclose competitors from access to a product or service.136 The Chicago Era largely abandoned the antitrust regulation of vertical transactions.137 The Clinton enforcement agencies recognized, however, that vertical mergers or joint ventures should be regulated in the unusual case in which they foreclose rivals from access to a product or service essential in one of the markets covered by the merger or joint venture.138

Thus, the plaintiff should be required to demonstrate that the parties to a vertical merger or joint venture have a customer/supplier relationship that gives the relevant entity an incentive to deny the plaintiff access to an essential product or service, thereby excluding it from the relevant market. Consider a hypothetical merger between Dell Inc. and Microsoft. With its monopoly in the operating systems market, there is no doubt that Microsoft’s Windows software constitutes a product essential to companies attempting to compete in the PC manufacturing market. A Dell/Microsoft entity would have both the incentive and the ability to provide the Windows operating system only for the computers that Dell manufactured and to refuse to deal with competing manufacturers of PCs such as Hewlett-Packard Co. (“HP”), Dell’s primary competitor in the PC market.139 Such a refusal to deal by the Dell/Microsoft entity would deprive consumers of the alternative of purchasing Windows-equipped PCs from other manufacturers, and it thus would be appropriate for a court or enforcement agency to insure that the merged entity did not foreclose HP and other PC manufacturers from licensing Windows software after the transaction was completed.

The Clinton enforcement agencies recognized that, even when a party to a vertical merger or joint venture controls an essential resource, the transaction need not be precluded in its entirety. The agencies allowed such transactions to proceed subject to consent orders that required the parties to grant their rivals non-discriminatory access to the relevant product or service after the completion of the transaction.140 Thus, if a plaintiff rebuts the presumption of legality for a vertical merger or joint venture, the courts and enforcement agencies can still approve the transaction if the parties agree to allow such access. Under the proposed approach, for example, the hypothetical Dell/Microsoft merger described above could be

136 See supra notes 74–75 and accompanying text.
137 See supra note 76 and accompanying text.
138 See supra note 77 and accompanying text.
140 See supra note 77 and accompanying text.
allowed to proceed, provided that the merged entity agreed to license Windows to HP and all other PC manufacturers on equal terms.

X. APPLYING THE PROPOSED APPROACH TO HORIZONTAL MERGERS AND JOINT VENTURES

A. The Competitive Effects of Horizontal Mergers and Joint Ventures

The abbreviated factual inquiry appropriate for horizontal restraints, vertical resale restraints, and vertical mergers and joint ventures should not apply to horizontal mergers and joint ventures. Unlike the competitive conduct described in Sections VII, VIII and IX supra, horizontal mergers and joint ventures affect consumers in ways that cannot be easily predicted. Indeed, horizontal mergers and joint ventures are as likely to benefit consumers as to harm them.

By virtue of their complete integration of previously separate firms, mergers eliminate all pre-existing competition between the parties to the transaction. As a result, post-merger prices may increase and output may decrease, to the detriment of consumers. On the other hand, the complete integration of resources that occurs in a merger can also benefit consumers. Mergers allow the parties to reduce costs by eliminating overlapping functions, achieving economies of scale and effecting synergies by combining complementary resources. These efficiencies usually cannot be achieved in the absence of a merger, and they are often passed on to consumers in the form of lower prices, enhanced quality, or product innovations.141

The adverse and beneficial effects of joint ventures among competitors are also likely to be closely balanced. Such ventures reduce the number of rivals in the market because the partners will refrain, in the natural course, from competing with their own affiliate.142 Horizontal joint ventures, however, also may generate substantial efficiencies beneficial to consumers. Joint ventures, for example, allow competing firms to achieve breakthroughs in research and development that no partner could have achieved on its own. For such ventures, the courts must balance the “tradeoff between efficiency gains . . . and the potential anti-competitive losses.”143

B. A Balancing Test for Horizontal Mergers and Joint Ventures

Since the likely competitive effects of horizontal mergers and joint ventures are not obvious on their face, courts should balance the potential adverse and

---


142 A joint venture “reduces the parents’ incentive for competition because whatever a parent might earn by making individual sales in the joint market is now offset by its lowered profit as a partner in the joint venture.” Joseph F. Brodley, Analyzing Joint Ventures with Foreign Partners, 53 ANTITRUST L.J. 73, 76 (1984).

beneficial effects of such transactions before ruling on their legality. Although the 
courts have begun to identify the factors that should be considered in such a 
balancing test, they have not yet developed an integrated approach that assigns 
evidentiary priorities and weights to each of the relevant factors. Without such a 
roadmap to guide their decision-making, fact finders could be required to make 
economic judgments that are beyond their competence.

The courts can simplify their analysis of horizontal mergers and joint ventures 
by establishing a market share threshold for the legality of such transactions. In its 
prima facie case, a plaintiff should have to prove that the collective market power 
of the parties to a transaction exceeded the relevant threshold. If a plaintiff fails to 
meet its burden of proving the parties’ collective market shares above the 
threshold, a court could dismiss the case without further inquiry. In such a way, the 
courts and agencies could create safe harbors of legality for horizontal mergers or 
joint ventures that create entities with market shares below the applicable threshold.

The Supreme Court in the Harvard Era adopted market share thresholds as 
low as 8 percent for horizontal mergers. The courts and agencies should raise the 
market share threshold for horizontal mergers to 30 percent. As Robert Pitofsky 
has explained, “There is a consensus among economists and antitrust lawyers that 
mergers among competitors should be prohibited [only] when they lead to [substantial increases in] concentration.” Courts and agencies now agree that 
horizontal mergers pose a competitive threat only when they make it easier either 
for the merged firm to exercise market power unilaterally or for the remaining 
firms in the market to coordinate their conduct to limit competition. After the 
completion of a horizontal merger below the 30 percent threshold, it should be no 
more likely that the merged entity acting on its own or in coordination with other 
firms in the market will be able to raise prices or restrict output in a manner 
adverse to consumers.

Courts and agencies should adopt a higher market share threshold for 
“upstream” joint ventures limited to joint buying, research and development, or 
production because the potential anticompetitive effects of such transactions are 
limited. Plaintiffs should be required to prove that the parties to an upstream joint

144 See supra note 78 and accompanying text.
145 Robert Pitofsky, The Political Content of Antitrust, 127 U. PA. L. REV. 1051, 1067 
(1979).
146 See Merger Guidelines, supra note 89, at 20,571 (describing mergers that allow 
individual firms to exercise market power unilaterally or a group of firms to exercise 
market power by coordinating their conduct).
147 Neither unilateral nor coordinated effects should be a concern when the parties to 
a merger collectively control less than 30 percent of the relevant market. By way of 
comparison, the DOJ and FTC 1996 Statements of Antitrust Enforcement Policy in Health 
Care establish an “antitrust safety zone” for joint purchasing arrangements among health 
care providers where “the purchases account for less than 35% of the total sales of the 
purchased product or service in the relevant market.” 4 Trade Reg. Rep. (CCH) ¶ 13,153, at 
20,812 (Sept. 5, 1996).
venture have a collective market share in excess of 40 percent. Upstream joint ventures with smaller market shares should not pose a threat to consumers. Upstream joint ventures are less anticompetitive than mergers because they do not affect the partners’ decisions on pricing. The parties to such ventures will retain their incentive to compete in marketing the relevant product. Ford, General Motors and Chrysler, for example, have entered into several research and development consortia for such products as an electric car, pollution control equipment, safety devices, and lightweight materials to replace steel in automobiles. These consortia have not prevented the Big Three from competing just as fiercely against each other in the sale of automobiles.

Downstream marketing joint ventures, however, have a greater likelihood of raising prices. Indeed, the potential anticompetitive effects of downstream joint ventures are often identical to mergers because such ventures allow the parties jointly to make pricing decisions. Therefore, the 30 percent market share threshold applicable to horizontal mergers should also apply to downstream marketing joint ventures.

If a plaintiff meets its burden of proving a horizontal merger or joint venture above the applicable market share threshold, a court or agency would have to consider the possible counterweights to the transaction’s likely anticompetitive effects. The courts should prioritize the relevant factors to assign them the appropriate place in the balancing analysis. As the Supreme Court recognized in its Chicago Era decision in General Dynamics, fact finders should first consider factors that may mitigate the adverse effects resulting from the relevant entity’s market power, such as ease of entry in the relevant market. Entry conditions have the potential to eliminate entirely the adverse competitive effects of a horizontal merger or joint venture. In markets in which entry is relatively easy, neither coordinated nor unilateral anticompetitive effects are likely to occur after such transactions, even if the parties hold a large collective market share. Neither the relevant entity acting alone nor the remaining competitors acting collectively will be able to raise prices, limit output, or otherwise restrict competition for any meaningful period if other firms are likely to enter the market in response to such anticompetitive conduct.

After considering ease of entry, courts should next consider whether the potential efficiencies resulting from a horizontal merger or joint venture are substantial enough to outweigh its likely anticompetitive effects. Antitrust regulators have struggled for years to determine how to balance the potential efficiencies of a horizontal merger or joint venture against its potential anticompetitive effects. Courts and agencies can clarify the balancing test by

---


149 See Joel B. Eisen, *Antitrust Reform for Production Joint Ventures*, 30 JURIMETRICS J. 253, 263 (1990) (stating that the “potential for price fixing and market domination grow as a company moves closer to the marketplace”).

150 See supra notes 84–92 and accompanying text.
ranking efficiencies in the order of their importance. The critical issue should be whether a horizontal merger or joint venture has the potential to leave consumers better off than they would have been if the parties had never combined. Efficiencies should be weighed in the balancing analysis according to their likelihood of being legitimate, realizable, beneficial to consumers, and unattainable other than through the proposed merger or joint venture. Under these criteria, cost savings should be recognized as the most significant efficiencies of all because, in a competitive market, they will usually be passed on to consumers in the form of lower prices.151

C. Applying the Proposed Approach to Particular Mergers

The proposed approach would simplify courts’ evaluation of horizontal mergers likely to occur in the near future in various domestic markets. Mergers among parties with collective market shares below the 30 percent threshold would be legal on their face. Mergers in which the parties’ collective market shares are slightly above the threshold could be justified when it is relatively easy for new competitors to enter the relevant market or when the parties can achieve highly ranked efficiencies such as cost savings. It would be more difficult for the merger partners to meet their burden of rebutting the presumption of illegality when their collective market shares were substantially above the 30 percent threshold.

1. Automotive Mergers

In late 2008, Ford, Chrysler, and General Motors sought assistance from the federal government to avoid what General Motors’ chief executive officer termed “a catastrophic collapse” of their operations.152 It is possible the companies could consider mergers as a means of eliminating excess capacity, reducing costs, and retooling to meet consumer demand for more fuel-efficient cars. Under the proposed approach, the courts and agencies easily could uphold mergers between any two of the Big Three. Mergers between General Motors and Chrysler, or between Ford and Chrysler, would be legal on their face because the parties’ collective market shares would be well below the 30 percent market share threshold.153 Courts and agencies would have to conduct a balancing analysis to

151 For a more detailed discussion of the relative role of such efficiencies in merger analysis, see generally Piraino, supra note 141, at 822–31.
153 The market shares of the Big 3 are as follows: General Motors, 20.4 percent; Ford, 15.8 percent; and Chrysler, 11.4 percent. See Auto Sales—Market Data Center—WSJ.com, http://online.WSJ.com/mdc/public/page/2_3022-autosales.html (last visited Sept. 1, 2009) [hereinafter Market Data]. However, in the event of a merger between General Motors and Chrysler, it has been estimated that “[t]he elimination of duplicate brands, models and dealers would push a combined market share [of the merged entity] down to 25% or less,” Paul Ingrassia, The Auto Makers Are Already Bankrupt, WALL ST. J., Nov. 21, 2008, at
determine the legality of a merger between Ford and General Motors, whose collective market share of 36.2 percent is just slightly over the 30 percent threshold.\textsuperscript{154} It would be extremely unlikely, given the excess capacity in the domestic automobile market,\textsuperscript{155} that a new company would enter that market after a Ford/General Motors merger. Nevertheless, it should be relatively easy for the courts and agencies to confirm the legality of such a merger. Under the proposed approach, cost savings are ranked the highest of all efficiencies because they have the greatest potential to benefit consumers. Cost savings should be significant in any Ford/General Motors merger due to the elimination of duplicate production facilities, brands, models, and dealers. Such costs savings should be sufficient to outweigh any anticompetitive effects that might result from a merger among parties with a collective market share so close to the 30 percent threshold of legality.

2. Airline Mergers

The airline market experienced its first major consolidation in late 2008, when Delta acquired Northwest, creating the largest domestic airline, with a 17.3 percent market share.\textsuperscript{156} A merger between any two domestic airlines other than Delta and American (the second largest domestic airline) would be permitted under the proposed approach because the parties’ collective market shares would be below the 30 percent threshold necessary to raise a presumption of illegality.\textsuperscript{157} A merger between Delta and American, however, would require additional analysis. Such a merger would create an entity with a 31.9 percent market share,\textsuperscript{158} just enough to create a presumption of illegality under the proposed approach.

Because the parties’ collective market share would be only slightly above the 30 percent threshold, it should be relatively easy for American and Delta to prove...
that their merger would have a net beneficial effect. After the merger, the American/Delta entity should be deterred from raising prices by the threat of entry from new airlines. Indeed, commentators recently pointed out that new entrants such as Southwest and JetBlue already “[have] destroyed the giant carriers’ profits on many routes they once dominated.”\(^ {159} \) The potential efficiencies of the merger should also be compelling. The merger should produce annual cost savings of more than $1 billion.\(^ {160} \) The merged entity should be able to use such cost savings to lower ticket prices and increase services desired by consumers. The potential benefits of such efficiencies, combined with the mitigating effects from the threat of entry by new airlines, should be sufficient to outweigh the adverse effects of a transaction in which the parties’ collective market share is only 1.9 percent above the threshold for a legal merger.

3. Hospital Mergers

In the Cleveland metropolitan area, two hospital systems, the Cleveland Clinic and University Hospitals, currently control 68 percent of hospital beds.\(^ {161} \) A merger between the Cleveland Clinic and University Hospitals clearly would be illegal under the proposed approach. The plaintiff would be able to raise a presumption of illegality by showing that the merged entity would control well more than 30 percent of the hospital health care market in the Cleveland area. The two hospitals would not be able to rebut the presumption of illegality for such a merger. With 68 percent of the relevant market, the merged entity presumably would be able to raise Cleveland-area prices for hospital care substantially.\(^ {162} \) Entry conditions in the Cleveland hospital market also would not constrain the merged entity from raising prices. Indeed, it would be extremely difficult for a new hospital to enter the Cleveland market after the merger. Not only are the capital costs significant, but a new hospital would find it difficult to attract the number of doctors and nurses required to compete effectively with a Cleveland Clinic/University Hospitals entity. Finally, the cost savings resulting from the


\(^ {160} \) The merger between Delta and Northwest is expected to produce $1 billion in annual cost savings. See Richard Anderson & Doug Steenland, Letter to the Editor, Not All Airlines Mergers Make Sense, But This One Does, Wall St. J., June 25, 2008, at A14. With their larger revenue and customer base, American and Delta should be able to generate greater savings if they decide to merge.


\(^ {162} \) Id. The customers of the new entity—insurers and managed-care providers—would not have enough bargaining power to prevent it from raising prices for hospital care. Indeed, the balance of pricing power may even shift in favor of hospitals at market share levels much lower than the 68 percent share of the hypothetical Cleveland Clinic/University Hospitals entity. See id. (“Twenty-two percent [market share for a hospital] is more than enough heft to push around insurers.”).
merger should not be sufficient to justify the higher prices to consumers likely to result from a transaction that would give the Cleveland Clinic/University Hospital entity control over 68 percent of the hospital beds in the Cleveland area.

D. Applying the Proposed Approach to Particular Joint Ventures

1. Joint Ventures That Facilitate Entry

When a joint venture is designed to develop new products that the joint venture partners could not have developed on their own, the market shares of the parties are, by definition, below the threshold for raising a presumption of illegality. In such a case, the joint venture partners would not be actual or potential competitors in the market for the new product, and their share of that market would be zero.\(^\text{163}\) The only effect of such a joint venture would be to introduce a new product or service that would not have existed in the absence of the venture. Since the only effect of such a joint venture would be beneficial to consumers, challenges to any such arrangement should be dismissed at the earliest stage of the litigation.

A joint venture will not limit any preexisting competition between parties if it permits them to produce a new product that was beyond their individual capacities.\(^\text{164}\) Consider, for example, a hypothetical joint venture among the Big Three and the major foreign automobile manufacturers to develop a new battery system for an all-electric “plug-in” vehicle that could operate over long distances without recharging. Assume that the automobile companies do not have the individual financial capacity to make the sustained multibillion-dollar investments necessary for the development of such a revolutionary battery. The joint venture would not have any anticompetitive effects, even if it included all the major automobile manufacturing companies in the world. The companies’ collaboration on the new battery technology would reduce neither their incentive nor their ability to continue to compete in the production and sale of automobiles just as aggressively as in the past. If the joint venture were successful in developing a new battery capable of powering automobiles over long distances, the automobile companies would compete to produce vehicles with the new battery at the lowest possible prices. The net effect of such a joint venture would be to expand the alternatives available to consumers by facilitating the commercialization of a new

\(^{163}\) See Pitofsky, supra note 76, at 898 (“Where neither parent was in the market or a likely entrant into the market served by the joint venture, it is hard to see why those joint ventures should be treated other than legal per se.”).

\(^{164}\) See Joseph Kattan, Antitrust Analysis of Technology Joint Ventures: Allocative Efficiency and the Rewards of Innovation, 61 ANTITRUST L.J. 937, 940 (1993) (describing a joint venture among pharmaceutical companies that gave partners access to patents necessary to produce a multivalent childhood vaccine that could be effective against several diseases at once). This type of joint venture promotes competition by permitting “the introduction of a new competitor that otherwise might never have come into being.” Robert Pitofsky, Joint Ventures Under the Antitrust Laws: Some Reflections on the Significance of Penn-Olin, 82 HARV. L. REV. 1007, 1018 (1969).
technology that none of the companies could have developed on their own. Thus the joint venture should be legal on its face.165

2. Upstream Joint Ventures

Upstream joint ventures should be legal on their face when their partners’ collective market shares are below the 40 percent market share threshold for such arrangements. In recent years, several domestic airlines have entered into “code-sharing” alliances that allow them to combine their frequent flier programs, share airport lounges and use common codes and flight designations that allow customers to travel on connecting flights to their ultimate destinations as if they were flying on one airline.166 These alliances do not have significant anticompetitive effects because they do not affect competition at the marketing stage. Although the parties can route customers on each other’s flights, they remain free to set fares and make capacity decisions on their own.167 Under the proposed approach, no plaintiff could successfully attack a code-sharing alliance among any two of the domestic airlines because the parties’ market shares would be below the 40 percent threshold for upstream joint ventures. If, for example, the two largest domestic carriers, American and Delta, entered into a code-sharing alliance, the joint venture’s collective market share would only be 31.9 percent.168

In 1984, the FTC allowed General Motors and Toyota to enter into a joint venture for the production of a single compact vehicle at a plant in Fremont, California.169 The parties have continued to operate the plant, which is now producing compact cars sold by Toyota as the “Matrix” and by General Motors as the Pontiac “Vibe.”170 The parties’ production joint venture has not prevented them from competing at the marketing stage on pricing and service for these two automobiles.

---

165 In late 2008, fourteen U.S. companies formed a joint venture to produce advanced lithium ion batteries to power electric automobiles. See Rebecca Smith, U.S. Firms are Joining Forces to Build Batteries, WALL ST. J., Dec. 18, 2008, at B1. One of the members of the joint venture acknowledged that, in the absence of the venture, it would be impossible for its partners to manufacture the advanced batteries because of “the overhead costs.” Id. A court would likely uphold the joint venture under the proposed approach because its only effect would be to provide consumers with an alternative product that its partners could not have produced in the absence of the venture.


167 Id.

168 See supra note 156 and accompanying text.


To be successful over the long term, General Motors must not only cut costs; it must also be able to produce the types of fuel-efficient, well-designed vehicles desired by American consumers. \(^{171}\) These new models will require enormous investments in research and development and the retooling of manufacturing plants. General Motors does not have the resources to make these investments on its own, and mergers or joint ventures with Ford or Chrysler, which have their own financial problems, likely cannot provide sufficient capital to produce the types of new vehicles consumers desire.

If the relevant antitrust standards were clarified, General Motors might, however, be able to develop and produce new fuel-efficient vehicles under a broader collaboration with Toyota. Both parties could benefit by entering into a joint venture for the development and production of such vehicles at several plants in the United States. Under such an arrangement, each party could access the other’s fuel-savings technologies, and each could share the costs of designing and producing the new fuel-efficient vehicle. General Motors would benefit in particular by being able to utilize Toyota’s more flexible labor contracts at the joint venture plants where the new vehicle would be produced. \(^{172}\)

The courts should uphold such a joint venture on its face. General Motors currently holds 20.4 percent of the domestic market, and Toyota has a 17.4 percent market share, giving them a combined market share of 37.8 percent, \(^{173}\) which is below the 40 percent threshold of legality for such an upstream research, development, and production joint venture. The broad collaboration between General Motors and Toyota at the research and production phases would benefit consumers without unduly restricting competition. General Motors and Toyota would continue to compete in marketing the new vehicle to consumers. Indeed, it is likely that, in the current competitive domestic automobile market, General Motors and Toyota would pass on a substantial amount of their cost savings to their customers.

3. Downstream Joint Ventures

Downstream joint ventures that are only slightly above the 30 percent market share threshold should be upheld if the parties can prove that they are likely to generate highly ranked efficiencies such as cost savings. Assume that American and Delta decide to enter into a marketing joint venture rather than the merger or the code-sharing arrangement described earlier. \(^{174}\) The joint venture allows the


\(^{172}\) The UAW labor contract negotiated by the workers at the joint venture plant in Fremont has been characterized as “shorter, simpler and more flexible in every respect than the union’s master contract” with the Big 3. See Paul Ingrassia, *The Latest Song of Detroit*, Wall St. J., Dec. 4, 2008, at A17. It could serve as a model for contracts at other plants operated by a General Motors/Toyota joint venture.

\(^{173}\) See Market Data, supra note 153.

\(^{174}\) See discussions supra Part X.C.2, D.2.
parties to sell seats on each other’s flights and to coordinate flight schedules. The venture does not, however, allow the parties to coordinate their pricing. Such a joint venture would be more adverse to consumers than the code-sharing agreement because the parties could jointly reduce their flight schedules. The joint venture, however, would be less anticompetitive than a merger between Delta and American because the parties would continue to compete on ticket pricing. Since American and Delta collectively control 31.9 percent of the airline market, the joint venture would exceed the 30 percent threshold for raising a presumption of illegality for a downstream joint venture. It should be easy, however, for the joint venture partners to rebut the presumption. Just as in the case of the hypothetical merger, a marketing joint venture between Delta and American would generate substantial cost savings due to the elimination of unprofitable routes, and those cost savings could be used by the joint venture partners to reduce ticket prices and increase customer services. Such efficiencies should be sufficient to outweigh the potential anticompetitive effects of a marketing joint venture whose partners’ market shares barely exceed the 30 percent threshold of illegality.

When the market shares of the parties to a downstream joint venture are substantially higher than 30 percent, the joint venture partners should find it more difficult to prove that the likely efficiencies of the venture will outweigh its anticompetitive effects. Consider a marketing joint venture proposed in 2008 by Google and Yahoo, which would have combined the companies’ Internet search businesses. The joint venture “would give the two companies a staggering 90% of the search advertising market . . . .” obviously well over the 30 percent threshold for raising a presumption of illegality for a downstream marketing joint venture. There would be no cost savings or other efficiencies substantial enough to outweigh the anticompetitive effects of a Google/Yahoo entity with a virtual monopoly of the Internet search and advertising market. Not only would the adverse effects of such a downstream joint venture be substantial; any cost savings resulting from the venture likely would be insignificant. The Internet search businesses of Google and Yahoo do not include tangible assets such as production facilities that can be eliminated; nor do such companies have large sales forces that can be combined. Thus it should be clear to a fact finder that the Google/Yahoo marketing joint venture should be prohibited.

---

175 See supra note 156 and accompanying text.
177 Id.
178 The Association of National Advertisers opined, for example, that “advertising prices will be higher as a result of the deal.” Id. (quoting Robert D. Liodice, President of the Association of National Advertisers).
XI. APPLYING THE PROPOSED APPROACH TO MONOPOLY CONDUCT

A. The Plaintiff’s Burden of Proof

The Harvard Era engendered a harsh approach to monopoly conduct, precluding firms with market power from engaging in certain types of aggressive competition that benefitted consumers.\(^\text{179}\) Chicago Era cases rejected the Harvard Era’s animus toward monopolies, emphasizing that firms most often obtain monopoly power simply because they have been more successful than their rivals in delivering to consumers the highest quality, the lowest prices, and the most innovative products.\(^\text{180}\) As a *Wall Street Journal* editorial said of Wal-Mart, “You don’t sell $300 billion a year worth of anything without doing something right.”\(^\text{181}\) Many observers argue that Microsoft gained its monopoly power because its Windows operating system was the most efficient, cost-effective product in the market. One commentator recently opined that “no company anywhere has done more to put high-quality software into more people’s hands.”\(^\text{182}\) A *New York Times* article in mid-2008 pointed out that, in the Internet search market, “Google built its near monopoly in rather exemplary fashion: by building a better mousetrap.”\(^\text{183}\)

As the Chicago School recognized, the mere possession of monopoly power should not suffice for an antitrust violation.\(^\text{184}\) A plaintiff should have the additional burden of proving that a defendant misused its monopoly power to harm consumers. Unfortunately, however, the courts have not yet developed an objective standard for identifying when monopolists misuse their market power. The courts’ current approach to monopoly conduct has confused practitioners and business executives as to the applicable standards of behavior for firms with large market shares.\(^\text{185}\)

It will thus be important for fact finders to adopt a means of identifying illegal monopoly conduct. I have proposed such a standard in previous articles: conduct

\(^{179}\) See supra notes 96–98 and accompanying text.


\(^{183}\) Nocera, * supra* note 176.

\(^{184}\) In the absence of anticompetitive conduct, there is no reason to assume that the mere possession of monopoly power is harmful to consumers. Indeed, some commentators believe that large companies can be more innovative than smaller companies. See, e.g., William J. Holstein, *Novel Thinking as a Survival Tactic*, WALL ST. J., Jan. 1, 2006, at BU8 (“I believe that only large companies like General Electric can solve big, complex problems, which can make a huge difference to humanity.”).

\(^{185}\) See supra notes 93–103 and accompanying text.
should be presumptively illegal under section 2 if a plaintiff can demonstrate that it is contrary to a monopolist’s legitimate self-interest in enhancing its efficiency and “makes no economic sense other than as a means of perpetuating or extending monopoly power.”

This standard would preclude conduct harmful to consumers without deterring monopolists from improving their products and services. When monopolists intentionally incur losses that are contrary to their immediate self-interest, it is reasonable to assume that they anticipate a long-term benefit that will outweigh such losses. Such benefit usually results from the artificial maintenance or extension of a monopolist’s market power beyond the period in which it otherwise would have been overthrown by the natural workings of the marketplace. Consumers are harmed in such cases because monopolists can maintain their control over the relevant market simply by leveraging their market power rather than by providing consumers with lower prices, better service, and more innovative products.

B. Access Restrictions

Plaintiffs are most likely to meet their initial burden of proof when they can demonstrate that a monopolist has denied its rivals equal access to an essential facility. Such conduct is clearly harmful to consumers because it allows a monopolist completely to exclude actual or potential competitors either from the monopolized market or from a related market to which a monopolist is attempting to extend its market power. In such cases, monopolists will no longer need to fear competition from new entrants, and they will lose any incentive to continue to provide consumers with better prices, products, and services.

Access restrictions, on their face, are contrary to a monopolist’s legitimate self-interest. A monopolist, like any other firm, should be motivated to maximize the return on its assets. In the ordinary course, therefore, a monopolist should want the largest possible number of firms to purchase its products or services. Most access restrictions limit the potential market for a monopolist’s products or services. No firm intentionally engages in conduct that will harm it in the marketplace. There must be an explanation for monopolists’ willingness to incur losses in implementing access restrictions. A firm would not want to forego profits, irritate its customers, or reduce its sales unless it perceived a payback that compensated for such losses. In the case of a monopolist, the payback often comes from its ability to exclude competitors that might threaten its monopoly position in its current market or in a new market to which it is attempting to extend its monopoly power. Thus, when a monopolist adopts a restrictive access policy, it is

---


187 Eleanor Fox has argued that courts should “[t]ake seriously the problem that a dominant firm’s use of leverage to fence out rivals . . . from a significant share of the market can degrade the competition, stave off challengers, and hurt consumers and the market.” Perspectives on the Future Direction of Antitrust, supra note 121, at 23.
more likely than not acting to perpetuate or extend its monopoly power. Under the proposed approach, such conduct would be presumptively illegal.

Such an approach is consistent not only with the Supreme Court’s Harvard Era essential facilities decisions but also with its Chicago Era decision in *Trinko*, which limited the scope of the essential facilities doctrine. The approach would impose a duty to deal in cases such as *Aspen* and *Kodak*, where a monopolist had refused to provide access to a product or service it had previously made available to other parties. In such cases, it would be clear that a monopolist was acting improperly because it would have no legitimate reason suddenly to cease providing a preexisting product or service. At the same time, the approach would recognize *Trinko*’s concern about protecting monopolists engaged in legitimate conduct. Under the proposed approach, a monopolist would not have a duty to deal with a rival when, as in *Trinko*, it did not already provide the relevant product or service to third parties. In such a case, the monopolist would not be acting contrary to its legitimate interests because it would not be foregoing revenue that otherwise would have been available to it. Thus, there would be no reason to infer that the monopolist was acting for an anticompetitive purpose.

After the plaintiff has proven that a monopolist has denied access to an essential facility, the monopolist should have the burden of rebutting by proving that it had a legitimate purpose for limiting access to its products or services. Shifting the burden to the monopolist to prove a benign intent for its conduct is fair because the monopolist is the party with access to the evidence that will prove its legitimate competitive purpose. Thus it is appropriate, after the plaintiff has made its prima facie case, to require monopolists to prove that they were acting in their legitimate self-interest, and not with an intent to exclude potential entrants in the relevant market.

C. Applying the Approach to High Technology Networks

Post-Chicago commentators have recognized the importance of protecting consumers against exclusionary conduct by high technology monopoly networks such as computer operating systems, telecommunications networks, and Internet software. Such networks have a unique ability to perpetuate their monopoly power by excluding potential rivals from their markets. High technology networks “are subject to very significant positive network externalities, which means that they become more valuable to a particular user as the system has a larger number of other users.” Such network effects benefit consumers, but they also make it easier for monopolists artificially to perpetuate their market power. Once a network standard becomes accepted, consumers are not likely to migrate to another network. Microsoft’s Windows operating system for PCs, for example, is attractive to users because it has a large installed base and is compatible with many

---

189 Hovenkamp, supra note 188, at 300.
applications programs. Monopoly networks such as Windows have become critical gateways through which potential entrants must pass to compete in the markets controlled by such networks.

It is therefore important for the courts to prevent the owners of high technology monopoly networks from excluding their rivals from particular markets. The proposed approach would clarify those circumstances in which network conduct should be illegal. Consider how the proposed approach would prevent Microsoft from leveraging its monopoly in the PC operating market into the market for Internet services. President Obama has expressed a particular interest in using antitrust as a tool to ensure open access to the Internet. The proposed approach would ensure open access to the Internet search market by firms attempting to provide an alternative to Microsoft’s web browser.

Under the proposed approach, Microsoft’s Windows operating system would be deemed an essential facility to which Microsoft would be required to grant actual or potential rivals access on nondiscriminatory terms. Programs for applications such as word processing, spreadsheets, databases, and Internet access can only run in conjunction with an operating system. Indeed, one commentator has stated that the “Windows operating system has become almost the sole entry to cyberspace.”

In September 2008, Google announced plans to introduce its own Internet browser (called “Chrome”) to compete with Microsoft’s Internet Explorer. Assume hypothetically that Microsoft, concerned that consumers might prefer Chrome because it “is designed to make it faster to browse the Web,” decides to implement various barriers to Chrome’s entry to the Internet browser market. Specifically, Microsoft prohibits PC manufacturers from pre-installing Chrome on Windows. Since “[m]any people find it easier to use the browser that comes loaded on their computer,” Google’s Internet browser fails to gain a significant toehold in the market.

See United States v. Microsoft Corp., 147 F.3d 935, 938 (D.C. Cir. 1998).

See Barack Obama: Connecting and Empowering All Americans Through Technology and Innovation, at 2, available at http://obama.3cdn.net/780e0e91cbeb6cdbf6e_6udymvin7.pdf. (“Barack Obama strongly supports the principle of network neutrality to preserve the benefits of open competition on the Internet.”).

Alan Murray, Antitrust Isn’t Obsolete in an Era of High-Tech, WALL ST. J., Nov. 10, 1997, at A1. As the president of America Online has pointed out, “In the new digital world, the operating system for computers is similar to the dial tone for telephones. . . . You can’t call anybody without going through the dial tone, and you can’t use software or a service without going through the operating system.” Steve Lohr, Microsoft Defends Its On-Line Plans, N.Y. TIMES, June 10, 1995, at BU1 (quoting Steve Case, President of America Online).


Id.

Id.

Microsoft has already been successful in using similar tactics against Netscape’s internet browser. Microsoft prevented computer manufacturers from pre-installing the
Under the proposed approach, a court or agency should presume that, in denying Chrome full access to Windows, Microsoft was acting against its legitimate self-interest in making Windows a universal operating platform for all related applications. Microsoft should not be successful in rebutting the presumption of illegality by demonstrating that it had a legitimate reason for excluding Chrome from the Windows operating system. The operating system should not be at risk of breakdowns, slower operation, or other deficiencies simply because Google’s browser was preinstalled on PCs. Thus a court could assume that Microsoft’s only purpose was to exclude Google from the Internet search market, thereby facilitating Microsoft’s extension of its own operating system monopoly to the market for Internet services. Under such circumstances, it would be appropriate for a court to order Microsoft to allow PC manufacturers to preinstall Google’s web browser.

A recent experience by a business-to-business Web site illustrates how the proposed approach could prevent Google from misusing its monopoly power to exclude potential entrants from the online advertising market. Since Google’s advertising software holds a 65 percent share of the online advertising market, access to such software is critical for new firms attempting to enter that market. Thus, under the proposed approach, Google’s advertising software would be considered an essential facility. In 2008, Sourcetool.com, a business-to-business Web site for purchasers of industrial products, began to use Google’s advertising software to place targeted advertising on its home page. Sourcetool was paid approximately ten cents every time a user clicked on an advertisement. Sourcetool was able to make a profit on such transactions because it only had to pay Google five to six cents per advertisement for using its program. As Joe Nocera explained in the New York Times, “In the summer of 2006, however, Google . . . , [s]uddenly and without warning, . . . raised Sourcetool’s minimum bid requirement from 5 or 6 cents to $1.00 . . . ,” thus eliminating any possibility for Sourcetool to make a profit on its ads.

Under the proposed approach, Google’s conduct would be presumptively illegal because, by making it impossible for Sourcetool to use its advertising software, Google would be acting against its legitimate independent interest in maximizing its revenue from the software. It would thus be reasonable for a fact finder to infer that the real reason for Google’s conduct was to prevent Sourcetool
from competing with Google in the online advertising market.\footnote{Indeed, the owner of Sourcetool believes that Google “didn’t like his Web directory because it was a search engine itself—though much more narrowly focused than Google’s search engine—and Google found it a competitive threat.”} Google could attempt to rebut the presumption of illegality by arguing that it had a legitimate reason for excluding Sourcetool from its advertising software. Google, in fact, informed Sourcetool that its Web directory did not meet Google’s standards for attributes such as loading speed, user friendliness, relevance, and originality. Google, however, never explained to Sourcetool any problems with its approach and never offered any suggestions for improvements.\footnote{Id.} Under such circumstances, a fact finder would be justified in concluding that Google’s stated reasons for excluding Sourcetool from a resource essential to compete in the Internet advertising market were pretextual. As Joe Nocera pointed out, “How can you adapt your business model to Google’s specs if Google won’t tell you what the specs are?”\footnote{Id.} Thus, under the proposed approach, Google could be ordered to return to its prior pattern of dealing, under which it charged Sourcetool a reasonable fee for using its advertising software.

XII. CONCLUSION

The drafters of the principal federal antitrust statutes gave the federal courts the flexibility to adapt rules of competitive conduct to evolving economic and political conditions. Over 100 years of judicial experience have demonstrated the wisdom of the drafters’ approach. Indeed, antitrust history since the passage of the Sherman Act constitutes one of the best examples of the advantages of the American common law system—an approach so successful that it has been copied by countries around the world.\footnote{See Senator Barack Obama, Statement for the American Antitrust Institute, supra note 4 (“America has been a longtime leader in antitrust, and our antitrust rules and institutions have often served as models for other countries wanting to make capitalism work for consumers.”).} The federal courts have been able to make a system originally designed to regulate the steel and railroad trusts of the nineteenth century equally relevant to Google, Microsoft, and other high technology companies of the twenty-first century.

At the end of the first decade of the twenty-first century, American consumers desire a centrist antitrust policy that precludes American companies from engaging in anticompetitive conduct but also frees them to pursue mergers, joint ventures, and other transactions that can enhance their long-term viability. This Article has explained how courts can fashion a new antitrust common law that will promote such goals. Under this approach, the legality of particular conduct will become clearer, and the antitrust laws will become more effective in deterring conduct harmful to consumers and encouraging conduct that benefits consumers. Once the
courts recognize that this new approach is supported by the best and most enduring precedent of the last 119 years, they can ensure that the federal antitrust laws continue to benefit American consumers through the next century.
ALLOCATING INFLUENCE

Heidi Reamer Anderson *

TABLE OF CONTENTS

I. INTRODUCTION .................................................................684

II. DEFINING AN ALLOCATING INFLUENCE CONFLICT .................688
   A. The Scene of the Conflict: Administrative Law Practice .............688
   B. A Basic Definition ...........................................................................691
   C. Distinguishing Allocating Influence Conflicts from Other Resource Allocation Issues ...............................................................691

III. THE CASE FOR ADDRESSING ALLOCATING INFLUENCE CONFLICTS NOW ....693
   A. Allocating Influence Conflicts Occur Frequently ..........................693
   B. Allocating Influence Conflicts Harm Clients and the Profession .......695

IV. EXISTING ETHICS SOURCES REGARDING ALLOCATING INFLUENCE ..........699
   A. Model Rules Regarding Conflicts of Interest .................................700
   B. State-Level Rules Regarding Conflicts of Interest .........................703
   C. ABA and State Bar Opinions ..............................................................704
   D. Statutory and Executive Lobbying Restrictions ...............................705

V. ADDRESSING THE ALLOCATION OF INFLUENCE THROUGH THE CONFLICT OF INTEREST RULES ........................................................................................................707
   A. The Allocating Influence “Problem” Is a Conflict of Interest That Should Be Addressed in the Conflict of Interest Rules .........................708
   B. Treating Allocating Influence Issues as Conflicts of Interest Is a Natural Extension of Previous Conflict of Interest Scholarship .................709
   C. Treating Allocating Influence Dilemmas as Conflicts of Interest Is Consistent with Existing Guidance Regarding Positional Conflicts ............711
   D. The ABA Should Amend the Comments to Model Rule 1.7 to Address Allocating Influence Conflicts ..................................................................................713

VI. CONCLUSION ..............................................................................716

* © 2009 Heidi Reamer Anderson, Assistant Professor of Law, Florida Coastal School of Law; B.A. 1997, Duke University; J.D. 2001, Washington & Lee University School of Law. I owe many thanks to Thomas Anderson, Susan Daicoff, Sarah Mietz, Nikki White, and Charmaine Yu for their helpful comments, research assistance, and encouragement. All errors are my own.
The herd seek out the great, not for their sake but for their influence; and the great welcome them out of vanity or need.

— Napoleon Bonaparte

I. INTRODUCTION

Influence peddlers, influence seekers, and government officials have a long and fascinating history. Centuries ago, those in need of government action greatly appreciated the value of “an audience with the king” and paid handsomely for such access, to the benefit of both the influence peddler and the influence seeker. Twenty-first century influence peddlers offer influence seekers a similar service, but they often do so subject to restrictions not faced by their medieval predecessors. Many of these restrictions are intended to make the whole influence

1 COLONEL MICHAEL B. COLEGROVE, DISTANT VOICES 35 (2005).
2 Some recent commentators note, only partially in jest, that influence peddling—i.e., lobbying—and not prostitution, is the world’s oldest profession. In their spin on the well-known “Garden of Eden” temptation story from the Bible, the Serpent was the first lobbyist, and his first successful target was Eve, whom the Serpent persuaded to eat the forbidden apple by “portraying knowledge gained from the apple as a virtue rather than [the] vice” that the other side, God, had made it out to be. Nicholas W. Allard, Lobbying Is an Honorable Profession: The Right to Petition and the Competition to be Right, 19 STAN. L. & POL’Y REV. 23, 23 & n.1 (2008) (citing Thomas M. Susman, Lobbying in the 21st Century—Reciprocity and the Need for Reform, 58 ADMIN. L. REV. 737, 738 (2006)). Although lobbyists are viewed by many as villains, less pejorative assessments of lobbyists note that they “provide vital information to lawmakers, help craft carefully balanced legislation, communicate insights from government officials to their clients, and otherwise provide the essential grease that makes the government run (more or less) smoothly.” See Alan B. Morrison, Lobbyists—Saints or Sinners?, 19 STAN. L. & POL’Y REV. 1, 1 (2008); see also Lloyd Hitoshi Mayer, What Is This “Lobbying” That We Are So Worried About?, 26 YALE L. & POL’Y REV. 485, 486 (2008) (noting that “the most successful ‘lobbyists’ include American heroes such as Patrick Henry, Susan B. Anthony, and Martin Luther King, Jr.”).
3 See Susman, supra note 2, at 738 (“Lobbying has surely been around as long as there has been government itself.”).
4 See generally WILLIAM V. LUNEBURG & THOMAS M. SUSMAN, THE LOBBYING MANUAL: A COMPLETE GUIDE TO FEDERAL LAW GOVERNING LAWYERS AND LOBBYISTS, (3d ed. 2005) (outlining regulations applicable to lobbyists); see also discussion infra Part IV.D (describing lobbying disclosure rules). However, there is some evidence that even medieval practitioners recognized and respected conflict of interest principles. See Helen A. Anderson, Legal Doubletalk and the Concern with Positional Conflicts: A “Foolish Consistency”? 111 PENN ST. L. REV. 1, 11–12 n.44 (2006) (“As early as 1280, a London Ordinance forbade attorneys from representing adverse parties in the same action and from dropping one client to represent another in the same case.”) (citing Jonathan Rose, The Ambidextrous Lawyer: Conflict of Interest and the Medieval and Early Modern Legal Profession, 7 U. CHI. L. SCH. ROUNDTABLE 137, 146–47 (2000)).
exchange more “ethical,” yet public perception indicates that additional reforms are necessary.5

When the influence peddler is a lawyer, and the influence seeker is a client, different ethical rules may come into play depending upon the decision maker before whom influence is sought.6 If the decision maker resides in the legislative branch (e.g., a United States congressman), then the Lobbying Disclosure Act and similar laws impose restrictions intended to increase transparency.7 If the decision maker resides in the judicial branch (e.g., a state court judge), then the rules of professional conduct prohibit a lawyer from improperly influencing the judicial decision maker.8 If the decision maker resides in an executive or administrative agency, then “the rules”—to the extent they apply—are not so clear.9

---

5 According to a recent George Washington Battleground Poll, the vast majority of Americans support more transparency in lobbying practices. See The George Washington University Battleground Poll 2006, available at http://www.tarrance.com/files/10420Q.pdf. For example, 86 percent of those polled supported greater disclosure by lobbyists about their work and their level of congressional contacts, and 87 percent supported greater disclosure by members of Congress about their contacts with lobbyists and about campaign contributions from lobbyists. Id.

6 Although the rules may vary depending upon the decision maker, there is a current effort to reconcile at least the very definition of lobbying in these multiple contexts. See Mayer, supra note 2, at 508–18 (chronicling the definition of “lobbying” in various federal and state laws) (citing FRANK R. BAUMGARTNER & BETH L. LEECH, BASIC INTERESTS: THE IMPORTANCE OF GROUPS IN POLITICS AND IN POLITICAL SCIENCE 34 (1998) (positing that the desire to influence the policy process is the “common thread” of all lobbying)).


8 See Mayer, supra note 2, at 529–31 (documenting procedural and institutional limits on interest group influence on judges); MODEL RULES OF PROF’L CONDUCT R. 8.4 (2008) (“It is professional misconduct for a lawyer to: . . . (e) state or imply an ability to influence improperly a government agency or official or to achieve results by means that violate the Rules of Professional Conduct or other law.”). This prohibition may be intended to offset the oft-stated notion that “a good lawyer knows the law; a great lawyer knows the judge.”

9 See LUNEBURG & SUSMAN, supra note 4, at 178–85 (describing the difficulty in determining the applicability of lobbying restrictions to communications with administrative agencies such as the Environmental Protection Agency); Mayer, supra note 2, at 526–29 (describing how influence is exercised over executive branch officials); see also discussion infra Part IV.D (reviewing statutory and executive lobbying restrictions). The general lack of clear ethical guidance on administrative agency lobbying—or “influencing”—means that important public policy decisions may be made in a “no rules” environment. See LUNEBURG & SUSMAN, supra note 4, at 495 (“None of the [Model Rules] directly address[es] the activities of lawyers qua lobbyists.”); see also David B. Wilkins, Who Should Regulate Lawyers?, 105 HARV. L. REV. 799, 815–24 (1992) (critiquing generally applicable Model Rules versus categorical or practice-specific rules and
Any attempt to clarify the ethics of administrative agency influencing should start with a core principle underlying modern legal ethics—the avoidance or resolution of conflicts of interest. Accordingly, this Article addresses a conflict of interest—the “allocating influence conflict”—that, to date, has escaped proper identification or analysis. In its simplest form, an allocating influence conflict emerges when (i) a lawyer properly may, and was retained by the client to, influence an agency decision maker, and (ii) there is a significant risk that allocating influence on behalf of one client is reasonably certain to inhibit substantially the lawyer’s ability to influence the same decision maker on behalf of another client.

Essentially, if a lawyer’s exercise of influence over an agency decision maker on behalf of one client could harm another client or the lawyer himself, then that lawyer likely faces an allocating influence conflict.

suggesting a contextual focus as a future guide); see generally Stanley Sporkin, The Need for Separate Codes of Professional Conduct for the Various Specialties, 7 GEO. J. LEGAL ETHICS 149 (1993) (concluding that “[t]he time is ripe to reassess the ethics rules that guide attorney conduct”).

10 See John S. Dzienkowski, Positional Conflicts of Interest, 71 TEX. L. REV. 457, 458 (1993) (“In recent years, conflicts of interest in the practice of law have presented the legal profession with some of the most serious problems of professional responsibility.”); see also James W. Jones & Bayless Manning, Getting at the Root of Core Values: A “Radical” Proposal to Extend the Model Rules to Changing Forms of Legal Practice, 84 MINN. L. REV. 1159, 1187 (2000) (identifying one core value of the legal profession as “independent judgment in advising the client, uninfluenced by considerations other than the requirements of the law and the client’s best interests”) (citing MODEL RULES OF PROF’L CONDUCT R. 1.7, 1.9–1.13, 2.1)). Accordingly, “this principle thus requires that, insofar as possible, a lawyer must avoid any conflict of interest when rendering legal advice to a client.” Id.

11 See discussion infra Part IV (demonstrating that existing ethical guidance provides a foundation for, but does not adequately address, allocating influence conflicts).

12 “Properly” is an important part of this definition because Rule 8.4(e) already prohibits stating or implying an ability to influence “improperly a government agency or official or to achieve results by means that violate the Rules of Professional Conduct or other law.” MODEL RULES OF PROF’L CONDUCT R. 8.4(e) (2008). Instances in which a lawyer properly may influence a decision maker include when lobbying a member of Congress or her staff or during an ex parte meeting with an administrative agency commissioner or her staff. See discussion infra Part II.A.

13 For the purposes of this Article, one should consider an allocating influence conflict as a subset of the “current client” conflict category. See MODEL RULES OF PROF’L CONDUCT R. 1.7 (2008). However, there may be situations where an allocating influence conflict can arise even absent a current-client conflict.

14 Exactly which officials qualify as “decision makers” varies depending on the circumstances and the applicable agency rules. For example, the Securities and Exchange Commission defines “decisional employee” to mean:

(i) The administrative law judge assigned to the proceeding in question; and
(ii) All members of the staff of the Office of Opinions and Review; and
Although allocating influence conflicts occur frequently in practice—and with particular frequency in administrative law practice—primary legal ethics sources do not explicitly address them. Similarly, although a few scholars have noted individual “problems” that could qualify as allocating influence conflicts, no scholarship has explained how to identify allocating influence conflicts or, perhaps more importantly, how to address them ethically. This lack of guidance means that many conflicts of interest are occurring without proper identification by the affected lawyer and without effective oversight from lawyers charged with enforcing ethical standards. In turn, the practical harm to clients generally is the same harm associated with all other conflicts of interest—the loss of loyalty, independent judgment, and zealous advocacy from one’s lawyer.

Part II of this Article defines an allocating influence conflict. In addition to providing a basic definition, it describes the circumstances in which these conflicts

(iii) The legal and executive assistants to members of the Commission; and
(iv) Any employee of the Commission who has been specifically named by order of the administrative law judge or the Commission in the proceeding to assist thereafter in making or recommending a particular decision; and
(v) Any other employee of the Commission who is, or may reasonably be expected to be, involved in the decisional process of the proceeding.


15 See Jonathan R. Macey & Geoffrey P. Miller, Reflections on Professional Responsibility in a Regulatory State, 63 GEO. WASH. L. REV. 1105, 1109–11 (1995) [hereinafter Macey & Miller, Regulatory State] (explaining how “private-sector lawyers” practicing “before government agencies” frequently face a conflict-like problem given that they are “repeat players” before that agency); see also discussion infra Part V.B (arguing that Macey & Miller’s identification of the “repeat player problem” supports treating the allocating influence problem as a conflict of interest); infra Part II.A (describing particular practice scenarios in which allocating influence conflicts are most likely to occur).

16 See discussion supra Part IV (demonstrating that various sources, including the Model Rules, state rules, bar opinions, and federal and state lobbying statutes do not address allocating influence conflicts).

17 See infra notes 120–129 and accompanying text (summarizing existing legal scholarship, which provides a foundation for addressing allocating influence conflicts).

18 In each state, a disciplinary committee, often composed of practicing lawyers, typically reviews ethics complaints, as delegated and supervised by the state supreme court. See Lisa G. Lerman & Philip G. Schrag, Ethical Problems in the Practice of Law 20–22 (2d ed. 2008).

19 See, e.g., MODEL RULES OF PROF’L CONDUCT R. 1.7 cmt. 1 (2008) (identifying “loyalty and independent judgment” as key interests protected by conflict of interest rules); Id. at pmbl. ¶ 2 (“As advocate, a lawyer zealously asserts the client’s position under the rules of the adversary system.”); see also Jones & Manning, supra note 10, at 1186–90 (describing core values of legal ethics).

20 See discussion infra Part II.B (crafting a specific definition of an allocating influence conflict).
typically emerge and distinguishes the allocating influence problem from other comparatively benign resource allocation conflicts.\(^{21}\)

Part III of this Article demonstrates the pressing need to address allocating influence conflicts given their prevalence and their harmful effects on clients and the profession.\(^{22}\)

Part IV demonstrates that existing ethics sources, though lacking direct recognition of allocating influence conflicts, provide the proper foundation to address them.\(^{23}\)

Part V addresses how to resolve allocating influence conflicts, first by showing how allocating influence conflicts fit within the existing conflict of interest framework and, second, by suggesting specific revisions to the comments of the Model Rules of Professional Conduct.\(^{24}\)

II. DEFINING AN ALLOCATING INFLUENCE CONFLICT

Properly defining an allocating influence conflict requires a three-level inquiry. The definition necessarily begins with the scene of the conflict, i.e., the circumstances in which allocating influence conflicts arise, as shown below in Section A. After establishing the scene, Section B identifies the specific facts that must be present for the scene to generate an ethical dilemma known as an allocating influence conflict. Section C then shows how allocating influence issues, which should be addressed as conflicts of interest, differ from mere resource allocation problems.

A. The Scene of the Conflict: Administrative Law Practice

Private-sector clients often hire administrative lawyers to persuade agency decision makers to alter or retain the agency’s rules so that the client may pursue its preferred business strategy.\(^ {25}\) Administrative lawyers do this “persuading”

\(^{21}\) See discussions infra Part II.A (depicting the typical “scene” of an allocating influence conflict); infra Parts II.B & C (demonstrating how allocating influence conflicts are an ethically significant type of resource allocation conflict); infra Part V.A (illustrating how an allocating influence conflict threatens the core principles underlying conflict of interest prohibitions).

\(^{22}\) See discussions infra Part III.A (analyzing how frequently allocating influence conflicts occur); infra Part III.B (arguing that allocating influence conflicts affect clients and the profession).

\(^{23}\) See discussion infra Part IV.

\(^{24}\) See discussion infra Part V.D (suggesting one additional comment and one revised comment to Model Rule 1.7).

\(^{25}\) Or, in some cases, one may seek to block or complicate a competitor’s preferred legal strategy. For example, if a lawyer’s client provides voice services over a copper wire—subject to multiple agency regulations perceived as burdensome, such as those requiring E911 capability—the lawyer may try to persuade the Federal Communications Commission to impose similar regulations on a new type of competing service provider,
through written submissions to the agency, similar in substance to a litigator’s brief. However, unlike their counterparts in litigation, who generally may not attempt to influence a judge, administrative lawyers may lobby or “influence” agency decision makers through informal ex parte communications, such as an in-person meeting with an agency commissioner, or a phone call to an agency staff person. A lawyer with access to, and influence over, a key agency decision maker often is hired by multiple clients, each expecting that the lawyer’s fee entitles him to the lawyer’s access and influence. As a result, an administrative lawyer often must decide how to allocate influence among these multiple clients.

To better appreciate the scene of the conflict, picture yourself as a lawyer who routinely practices before an administrative agency in Washington, D.C. On one particular day, your office phone rings. Client A needs agency action on a pending waiver request before its board meeting next week. As you dial the number of

such as one that provides voice services using voice over Internet protocol technology. See In re IP-Enabled Services: E911 Requirements for IP-Enabled Service Providers, WC Docket Nos. 04-36 & 05-196, 20 F.C.C.R. 10245 (proposed June 3, 2005) (codified in 47 C.F.R. pt. 9); see also John F. Duffy, The FCC and the Patent System: Progressive Ideals, Jacksonian Realism, and the Technology of Regulation, 71 U. COLO. L. REV. 1071, 1142–43 (2000) (describing how ex parte lobbying efforts were used to influence the FCC’s decisions regarding personal communications service licensing); see generally Kevin Werbach, Only Connect, 22 BERKELEY TECH. L.J. 1233 (2007) (documenting FCC proceedings regarding the regulatory status of voice over Internet protocol service).

26 See Edward Rubin, It’s Time to Make the Administrative Procedure Act Administrative, 89 CORNELL L. REV. 95, 100–01 (2003) (“To make a rule, an agency must publish a proposed version of the rule or a statement of the rule’s subject matter, allow a period of time for private parties to file written comments with the agency, and, after the comments have been received, publish a final version of the rule with a statement of basis and purpose.”). Federal agencies typically perform two basic functions—rulemakings and adjudications. See 5 U.S.C. § 553 (2006) (rulemaking); 5 U.S.C. § 554 (2006) (adjudications). The initial scenario described herein focuses on the rulemaking function. In agency rulemakings, interested parties may file comments and reply comments in response to a notice of proposed rulemaking. See 5 U.S.C. § 553(c) (2006). Under certain agency rules, parties also may engage in written or in-person ex parte presentations. See, e.g., 47 C.F.R. ch. I, subch. A, pt. 1, subpt. H (2008); see also infra notes 43–46 and accompanying text (noting agencies that permit ex parte communications, often subject to limitations).

27 See Rubin, supra note 26, at 119–20 (explaining how the permissibility of ex parte consultations varies based on the nature of the proceeding). Such communications often are permitted under an agency’s rules regarding ex parte communications, which may condition the permissibility of the communication on later public disclosure of the communication. See infra notes 43–46 and accompanying text (reviewing ex parte rules applicable to various administrative agency proceedings).

28 See Richard A. Epstein, The Legal Regulation of Lawyers’ Conflicts of Interest, 60 FORDHAM L. REV. 579, 585 (1992) (positing that “when the stakes are very large, clients want lawyers who have real experience in a given area, just the way a patient with a brain tumor wants to hire a physician who has done many similar surgeries”).

29 For example, Tribune Television Co. sought and obtained two temporary waivers
your most trusted staff contact at the agency to urge quick action on Client A’s request, an e-mail arrives. \(^{30}\) Client B needs the same agency to approve its merger application so it may close on a transaction before the financial quarter ends. \(^{31}\) Your most trusted staff contact at the agency is well-positioned to speed up action on Client A’s request and on Client B’s request. You feel, however, that you cannot ask your staff contact to assist both Client A and Client B (i.e., to grant you two personal favors in a single day) without risking your influential relationship with your staff contact—a relationship established over the past several years. \(^{32}\) Welcome to your first “allocating influence” dilemma for the day—how to allocate your influence over a decision maker between Client A and Client B, and still maintain your relationship with the agency decision maker.


\(^{30}\) Many administrative agencies publish a public phone directory, but only the well-connected lawyer has access to important staff members’ direct-dial numbers. See, e.g., http://www.fcc.gov/fcc-bin/findpeople.pl (enter name in search box) (last visited Sept. 1, 2009) (listing the same general ten-digit number for multiple staff members who possess separate individual, yet unpublished, direct-dial numbers). However, those lawyers with influential relationships are privy to certain direct-dial numbers and often use their “smile dial” capabilities to a client’s advantage. See Macey & Miller, Regulatory State, supra note 15, at 1111 (noting that “law firms that are repeat players before regulatory agencies are truly favored in the sense that they are given preferential treatment by the bureaucrats in the agencies”).

\(^{31}\) The Federal Communications Commission and Federal Trade Commission often must approve certain mergers. See 47 U.S.C. § 214 (2006) (giving the FCC authority to review certain telecommunications mergers via declaration that “[n]o carrier . . . shall acquire or operate any line . . . unless and until there shall first have been obtained from the Commission a certificate that the present or future public convenience and necessity require or will require [such acquisition]”); 47 U.S.C. § 310 (2006) (giving the FCC authority to review various mergers involving licensees via declaration that “[n]o construction permit or station license, or any rights thereunder, shall be transferred, assigned, or disposed of in any manner . . . to any person except upon application to the Commission and upon finding by the Commission that the public interest, convenience, and necessity will be served thereby”).

\(^{32}\) See Macey & Miller, Regulatory State, supra note 15, at 1110 (noting that, when administrative lawyers seek to “appease the agency bureaucrats,” they “may be unwilling to zealously assert the client’s interests if doing so would alienate the bureaucrats” and jeopardize the lawyers’ long-term personal interests).
B. A Basic Definition

The Client A versus Client B conflict described above is a single, practical example of an allocating influence conflict. In more abstract terms, an allocating influence conflict begins to emerge when (i) a lawyer properly may, and was retained by the client to, influence an agency decision maker, and (ii) there is a significant risk that allocating influence on behalf of one client is reasonably certain to inhibit substantially the lawyer’s ability to influence the same decision maker on behalf of another client.33 If these two elements are present, then an allocating influence conflict is present, and a lawyer may not proceed with the representation(s) absent client consent.

C. Distinguishing Allocating Influence Conflicts from Other Resource Allocation Issues

If one casts the above definition of an allocating influence conflict in general terms, the first element requires a lawyer’s possession of or access to a finite resource and the second element requires a risk that use of that resource to one client’s benefit will restrict the lawyer’s ability to use that same resource to benefit another client or the lawyer. In an allocating influence conflict, the finite resource to be allocated is proper influence over a decision maker.34 Other resources that arguably satisfy these first two elements include time, staff support, and even filing cabinet space. A lawyer who must decide how to allocate these everyday resources among clients does not face an ethically significant conflict of interest; indeed, any conflict of interest definition that included these resource allocation issues would be too broad.35 Thus, a narrower definition is necessary.

33 See supra notes 12–13 and accompanying text.

34 See John P. Heinz, The Power of Lawyers, 17 GA. L. REV. 891, 892 n.2 (1983) (“Thus, if one’s work involves allocative decisions about resources that are scarce and valued, the autonomy or control that permits one to make those decisions authoritatively constitutes power or influence.”); see also Burnele V. Powell, What Clients Want and Why They Can’t Have It, 52 EMORY L.J. 1135, 1144 (2003) (“I assert, then, that what clients want, in the absence of a lawyer willing to offer victory, is a lawyer who exudes victory—what I will call a ‘lawyer of stature.’ Clients wish to believe that, despite their lawyer’s inability to say it, she will act as though the client’s question really has been taken seriously.”).

35 Ask a busy lawyer how her day was, and she may offer you a metaphor: “I spent my whole day putting out fires,” with each “fire” being a client problem she addressed. See, e.g., Sue Reisinger, Wachovia Hires an Expert at Putting Out Fires, CORPORATE COUNSEL, Aug. 11, 2008, available at http://www.law.com/jsp/ihc/PubArticleIHC.jsp?id=1202423655554. Deciding how to divide one’s time and other finite resources among different client “fires” is a problem virtually all lawyers face. See Robert W. Gordon, The Independence of Lawyers, 68 B.U. L. REV. 1, 7–8 (1988) (“Ideally independent lawyers freely decide . . . how to divide their time between paying clients and other commitments, what strategies and tactics to follow in pursuit of the clients’ ends, and so forth . . . . The client dictates (as moderated by the lawyer’s advice) the results to be sought and has the
So what is it that separates an allocating influence conflict from any other decision in which a lawyer must allocate a finite resource, such as time in the day? The defining characteristics involve two client-focused elements: (i) the client’s choice to retain the lawyer being based primarily on the lawyer’s access to the finite resource, and (ii) the client’s ignorance regarding both the resource’s limited nature and the consequences of the resource’s allocation to someone else (i.e., that the lawyer could or would use the resource on behalf of another client to her detriment). The presence of these two elements distinguishes an allocating influence conflict, which is harmful to clients and in need of addressing, from more benign resource allocation conflicts, which do not require an ethically sound remedy.

For example, it is highly unlikely that a client would choose to hire a particular lawyer primarily on the basis of his access to filing space or pens. Similarly, it is highly unlikely that a client would not know that a lawyer’s time is limited and that his choice to spend time on one client’s matter limits his ability to spend time on another client’s matter. In contrast, when the finite resource is influence over a decision maker, the client-focused elements flip-flop. Specifically, a lawyer with influence over a decision maker often is retained solely because of his influence. The failure to deliver on this implicit commitment to exercise

---

36 See infra note 40 and accompanying text.
37 Macey & Miller, Regulatory State, supra note 15, at 1111.
38 Consequently, the proper remedies for an allocating influence conflict focus on these two elements. See discussion infra Part IV.
39 Further, the allocation of time is well-documented in a client’s bill, thus providing transparency that is lacking when the resource instead is influence.
40 This phenomenon is part of the broader trend of clients seeking lawyers with narrow specializations. See Epstein, supra note 28, at 585 (demonstrating that legal specialists are in high demand, especially in high-stakes matters). Administrative lawyers, who often specialize in a practice before a single administrative agency or even before a single bureau of a single agency, thus face an increased intensity when it comes to conflict of interest issues. See Macey & Miller, Regulatory State, supra note 15, at 1111 (discussing special treatment afforded to “repeat players before regulatory agencies”).
influence on behalf of a client may even constitute a breach of contract.\footnote{See Epstein, supra note 28, at 580–581 (articulating the contractual principles underlying conflicts of interest).} Further, a client may not be aware that her lawyer must divide his influence among clients and may exercise that influence on behalf of someone else to her detriment.\footnote{See Macey & Miller, Regulatory State, supra note 15, at 1111 (“[L]awyers may trade off the rights of some clients in order to curry favor with the agency and thereby advance the rights of other clients.”).} Thus, there is a lack of transparency with respect to allocating influence that is absent with respect to allocating other finite resources. Ultimately, this lack of transparency, coupled with heightened client expectations, means that lawyers should make influence allocation decisions according to a higher ethical standard than is required for allocation decisions involving other resources.

### III. THE CASE FOR ADDRESSING ALLOCATING INFLUENCE CONFLICTS NOW

Once influence allocation conflicts are defined and understood as a special type of resource allocation issue, the next logical question may be, “Why should we care?” Simply put, the legal ethics community should care about allocating influence conflicts given (A) the frequency with which they occur, and (B) their harmful effects on clients and the profession, as demonstrated in turn below.

#### A. Allocating Influence Conflicts Occur Frequently

The allocating influence conflict scenario described above occurs frequently for lawyers practicing before administrative agencies and other lawyers valued for their influence. This is because administrative agencies, unlike courts, often permit—if not expect—a lawyer to meet directly with agency decision makers to present a client’s position without the “other side” present.\footnote{See, e.g., 47 C.F.R. § 1.1206 (2009) (listing FCC proceedings in which ex parte presentations are permissible). Some federal agencies, like the Securities and Exchange Commission, require that all parties be given notice and opportunity to participate in ex parte communications. See 17 C.F.R. § 201.120(a)(1) (2009) (“Except to the extent required for the disposition of ex parte matters as authorized by law, the person presiding over an evidentiary hearing may not: (1) consult a person or party on a fact in issue, unless on notice and opportunity for all parties to participate.”). Other agencies, such as the Federal Election Commission, permit single-sided ex parte presentations provided that a notice describing the nature of the communication is made part of the public record. See 11 C.F.R. § 201.4(a) (2004) (“A Commissioner or member of a Commissioner’s staff who receives an ex parte communication concerning any rulemaking or advisory opinion . . . shall . . . provide a copy of a written communication or a written summary of an oral communication to the Commission Secretary for placement in the public file of the rulemaking or advisory opinion. The Commissioner or staff member shall advise any person making an oral communication that a written summary of the conversation will be made part of the public record.”). The proceedings in which ex parte communications are
Every ex parte presentation is an opportunity to exercise influence. Thus, any agency that permits these presentations in matters of importance is fertile ground for allocating influence conflicts.\textsuperscript{45} The vast number of administrative agencies at the federal, state, and local levels makes the likely number of potential allocating influence conflicts quite large.\textsuperscript{46} Considering the number of lawyers practicing permissible include informal and formal rulemakings, the heart of any agency’s work, as well as other proceedings, including high-profile merger reviews. See \textit{supra} notes 29–32 and accompanying text.

\textsuperscript{44} See 5 U.S.C. § 557(d) (2006) (describing restrictions on ex parte presentations applicable to federal agencies generally); \textit{see, e.g.}, 47 C.F.R. § 1.1202(b) (2009) (defining the term “ex parte presentation” as used before the Federal Communications Commission). Ex parte presentations may be done in person, by telephone, or even by e-mail. \textit{See, e.g.}, 47 C.F.R. § 1.1202(b). The ability to get such meetings scheduled and the time allowed during each such meeting are finite, highly valued resources. See \textit{ERWIN G. KRASNOW, DAVID R. SIDDALL & MICHAEL D. BERG, FCC LOBBYING xi} (2001) (“More lobbyists representing more interests means less time to make your case before the FCC and more competition for the attention of decision makers. Like the spectrum they allocate, the time of those officials is scarce.”). The fact that some lawyers perpetually are more successful at obtaining ex parte meetings with important decision makers is well-known among lawyers in the administrative law bar and the agencies before whom they practice, but is not as well understood by existing and prospective clients. \textit{Id.} at 33–38 (describing the process for obtaining meetings with FCC decision makers and the associated opportunities for allocating influence). This asymmetrical information problem likely could be reconciled by requiring informed client consent to allocating influence conflicts, as suggested in Part V, \textit{infra}. See generally U.S. GEN. ACCOUNTING OFFICE, FCC SHOULD TAKE STEPS TO ENSURE EQUAL ACCESS TO RULEMAKING INFORMATION 3–18, GAO-07-1046 (2007) (documenting and criticizing one example of the information gap between “stakeholders [that] had access to nonpublic information” and those who lacked access to such information, and describing the harmful effects of that gap on the agency decision-making processes and results).

\textsuperscript{45} Agencies that permit ex parte presentations in at least some proceedings include the Food and Drug Administration, \textit{see} 21 C.F.R. § 10.55 (2009), the Federal Energy Regulatory Commission, \textit{see} 18 C.F.R. § 385.2201 (2009), the Federal Communications Commission, \textit{see} 47 C.F.R. § 1.1206 (2009), and the Federal Election Commission, \textit{see} 11 C.F.R. § 201.4 (2009). Important ex parte communications also occur in non-administrative contexts, such as when a criminal defense attorney speaks informally with a prosecutor regarding a possible plea. In the latter scenario, the judge, and not the prosecutor technically is the “decision maker”; however, in practice, a prosecutor’s recommendation regarding a plea carries significant, and often determinative, weight. See generally Alafair S. Burke, \textit{Prosecutorial Passion, Cognitive Bias, and Plea Bargaining}, 91 MARQ. L. REV. 183 (2007) (discussing factors that influence prosecutorial discretion in plea bargaining).

\textsuperscript{46} \textit{See supra} note 45 (listing federal agencies allowing ex parte presentations). At the state level, some administrative agencies, such as public utilities and corporations commissions, have more flexible ex parte policies than others. \textit{Compare} CAL. PUB. UTIL. COMM’N R. PRACT. & P. 8.2(a) (2007) (permitting ex parte communications with the state Public Utilities Commission “without restriction or reporting requirement” in any “quasi-legislative proceeding”), \textit{with} ARIZ. ADMIN. CODE 14-3-113(C)(1) (2006) (prohibiting any “oral or written communication, not on the public record, concerning the substantive merits
before such agencies and the number of clients each lawyer represents, the potential number of allocating influence conflicts grows exponentially.

**B. Allocating Influence Conflicts Harm Clients and the Profession**

The frequency with which allocating influence conflicts occur, though striking, is an insufficient reason to address them absent some demonstrated harm. If, however, it can be shown that such conflicts cause significant harm, the case for addressing them would be much stronger. As shown below, a significant harm is present. Left unacknowledged and unaddressed, allocating influence conflicts will continue to impose significant costs on clients and the profession.

For a client, the most obvious harm of an allocating influence conflict is the loss of a lawyer’s independent judgment and zealous advocacy. A lawyer with the ability to secure only a single meeting in a given week with an agency decision maker faces a difficult judgment call: which client matter should be discussed in of a contested proceeding or siting hearing to a commissioner [of the Corporation Commission] or commission employee involved in the decision-making process for that proceeding or siting hearing”); see also 4 CODE COLO. REG. 723-1-1105 (2009) (prohibiting “ex parte communications concerning any disputed substantive or procedural issue” with the state Public Utilities Commission, except communications in proceedings such as those “relating to a pending non-adjudicatory proceeding”); OHIO ADMIN. CODE 4901-1-09 (2007) (prohibiting ex parte communications with the state Public Utilities Commission, “unless all parties have been notified and given the opportunity to be present or to participate by telephone, or a full disclosure of the communication insofar as it pertains to the subject matter of the case is made”); ORE. ADMIN. R. 860-012-0015 (2009) (discouraging ex parte communications with the state Public Utility Commission, which, “if made, must be disclosed to ensure an open and impartial decision-making process”); 5 VA. ADMIN. CODE 5-20-50 (2009) (prohibiting ex parte communications with the state Corporation Commission “without giving adequate notice and opportunity for all parties to participate”).

Although it is difficult to pinpoint the number of lawyers practicing administrative law, one helpful metric is the 17,000 lawyer-members of the American Bar Association’s Administrative Law and Regulatory Practice Section. See ABA, Membership Information, http://www.abanet.org/adminlaw/members.html (last visited Sept. 1, 2009).

See infra notes 49–65 and accompanying text.

See Monroe H. Freedman, A Civil Libertarian Looks at Securities Regulation, 35 OHIO ST. L.J. 280, 285 (1974) (noting that pressure to “curry favor” with agency personnel “intimidate[es] attorneys into forego[ing] zealous advocacy on behalf of their clients”). Although protecting independent judgment was a primary motivation for ethical rules going back to the 1908 legal ethics canons, some commentators have identified a decline in the profession’s respect for this bedrock principle. See also MODEL CODE OF PROF’L RESPONSIBILITY (1969) (addressing conflicts of interest under Canon 5, which was titled, “A Lawyer Should Exercise Independent Professional Judgment on Behalf of a Client”); see generally Gordon, supra note 35 (chronicling a perceived decline in the professional independence of lawyers); ABA CANONS OF PROF’L ETHICS, Canon 6 (1908) (“It is unprofessional to represent conflicting interests, except by express consent of all concerned given after a full disclosure of the facts.”).
that meeting? And after making that judgment call, the lawyer faces yet another difficult question: should she emphasize the dire need for immediate action on behalf of Client A—as the client expects—or should she downplay that need given the likelihood that she soon will need to make a similar claim on behalf of Client B?

Such decisions inevitably tempt a lawyer to trade off one client’s interests to benefit another client or even to benefit the lawyer. As one preeminent scholar has noted, some agencies effectively “encourage lawyers to trade off the rights of some clients in order to curry favor with the [agency] and thereby advance the rights of other clients,” which turns a temptation into a near necessity. In these circumstances, the client is harmed because the lawyer no longer is making decisions with the client’s interests paramount. Instead, the allocating influence conflict has compromised the lawyer’s independent judgment. A failure to address allocating influence conflicts will ensure that this harm continues.

The loss of independent judgment is a clear, easy-to-categorize harm resulting from allocating influence conflicts. A more subtle, yet perhaps more significant, harm results from the gap between a client’s expectations regarding influence allocation and the way in which influence is allocated. A client who desires

---

50 Some commentators have considered this problem to be a “conflict” that, as indicated above, results in the loss of “zeal” in the lawyer’s representation. See Freedman, supra note 49, at 285; see also Macey & Miller, Regulatory State, supra note 15, at 1109 (describing the “repeat player problem” as one of two conflicts that arise in administrative law practice that are difficult to resolve under market principles alone).

51 Macey and Miller, drawing upon Freedman, aptly describe this general ethical dilemma as follows: “Thus, at one stage of a proceeding before an agency, a client may benefit because his attorney receives ‘the opportunity, denied to others, to appear before [an agency] at a critical stage’ in the proceedings. At a later stage in the proceedings, however, the lawyers may ‘trade off the rights of some clients in order to curry favor with the [agency] and thereby advance the rights of other clients.’” Macey & Miller, Regulatory State, supra note 15, at 1111 (citation omitted).

52 See id. (arguing that the appearance of agency lawyers before the same decision makers on behalf of multiple clients “creates a conflict that can cause private-sector lawyers who represent clients before government lawyers and bureaucracies to subordinate the interests of their clients to their own long-term interests in maintaining a close and cordial relationship with the government lawyer or bureaucracy”). Like the boy who cried “Wolf!” too many times, a lawyer who insists that every client matter is an emergency worthy of an agency decision maker’s immediate attention soon will find his calls unanswered and pleas ignored. See AESOP, AESOP’S FABLES 91 (George Fyler Townsend trans., Forgotten Books 1965) (“A shepherd-boy, who watched a flock of sheep near a village, brought out the villagers three or four times by crying out, ‘Wolf! Wolf!’ and when his neighbors came to help him, laughed at them for their pains. The Wolf, however, did truly come at last. The Shepherd-boy, now really alarmed, shouted in an agony of terror: ‘Pray, do come and help me; the Wolf is killing the sheep;’ but no one paid any heed to his cries, nor rendered any assistance. The Wolf, having no cause of fear, at his leisure lacerated or destroyed the whole flock.”).


54 Gaps like these are ripe for addressing via ethical rules because the primary
prompt agency action in a matter often seeks a lawyer the client perceives as capable of influencing the agency decision maker.\textsuperscript{55} In many cases, the perceived influence may be the single determinative factor the client uses in hiring a lawyer.\textsuperscript{56}

The quid pro quo is simple and direct—the client pays the higher rate or retainer in exchange for the lawyer’s influence.\textsuperscript{57} Consequently, the client reasonably may expect that, when the lawyer has an opportunity to influence the relevant decision maker, the lawyer will do so on behalf of the client. Gone unacknowledged, however, is that several other clients likely hired the same lawyer based on a similar understanding.\textsuperscript{58} The lawyer’s decision to discuss one client’s matter over other clients’ matters automatically harms the other clients and deprives them of the single thing for which they thought they were paying a

\textsuperscript{55} Macey & Miller, Regulatory State, supra note 15, at 1111 (arguing that “repeat players before regulatory agencies are truly favored in the sense that they are given preferential treatment by the bureaucrats in the agencies” and that “[c]lients, of course, are attracted to law firms whose lawyers qualify for such preferential treatment”).

\textsuperscript{56} See Epstein, supra note 28, at 585 (“[C]lients want lawyers who have real experience in a given area, just the way a patient with a brain tumor wants to hire a physician who has done many similar surgeries.”).

\textsuperscript{57} For example, a client may agree to pay $600 an hour for Lawyer A in lieu of $220 for Lawyer B, although both are equally competent in other areas, solely because Lawyer A, by virtue of a previous stint as a legal advisor to an agency commissioner, has direct access to and influence over that decision maker. See Gordon, supra note 35, at 37 (noting that clients value a lawyer who “has not only pierced the veil of legal mysteries but also has access to persons and milieux exotic to the provincial client”). Gordon further posits that “[l]awyers might acquire such contacts and knowledge through government service, involvement in political campaigns, or just from having been on the scene.” \textit{Id.} Examples of lawyers who may have acquired such “influence” are Wall Street lawyers and “Washington lawyers [who] have served as intermediaries between corporations and government agencies.” \textit{Id.}

\textsuperscript{58} That clients hire certain lawyers based on their perceived “access” is not a new or automatically troubling concern. See Gordon, supra note 35, at 37 (noting that a “client’s dependence on the lawyer’s special knowledge increases[] as the content of representation moves away from the client’s familiar turf onto the lawyer’s”). Complicating matters here, however, is the fact that the lawyer also has his relationship with the decision maker that he may be reluctant to use on behalf of any client at this time. See Freedman, supra note 49, at 285 (documenting the temptation to preserve a relationship with a decision maker at the expense of the current client).
This gap between a client’s expectation and the lawyer’s service delivery creates a loyalty problem that, in turn, serves as an intrinsic justification for addressing allocating influence conflicts through the professional responsibility rules. The disconnect between expectations and delivery also may be viewed as a “backsliding” or “agency cost” problem.

Ultimately, the problem is one of resource allocation, as discussed above, or, in other words, an allocating influence conflict. What makes the harm especially troubling, however, is that the resource to be allocated—influence—is the key resource for which the client is paying. Because the conflict strikes at the

---

59 This may be viewed as an informational asymmetry problem, which is common in lawyer-client relationships. See Macey & Miller, Economic Analysis, supra note 54, at 970–71 (describing the core information asymmetry problem as occurring when “[t]he lawyer possesses a store of specialized knowledge, skill, and judgment that the client lacks”).

60 This gap between a client’s expectations and the lawyer’s delivery served as a significant justification for addressing a “positional” conflict as a conflict of interest. See Dzienkowski, supra note 10, at 494–95; MODEL RULES OF PROF’L CONDUCT R. 1.7 cmt. 24 (2008) (listing “the clients’ reasonable expectations in retaining the lawyer” and “the significance of the issue to the immediate and long-term interests of the client” as factors to consider when evaluating a positional conflict of interest); see also discussion infra Part V (discussing parallels between positional conflicts of interest and allocating influence conflicts of interest). As one conflicts scholar has stated, clients expect that their lawyers will be loyal and, specifically, will not “concurrently . . . attack[] the view previously taken for them.” Dzienkowski, supra note 10, at 494 (citation omitted). Further, the client’s loyalty expectation regarding position-taking is greater with respect to issues at the heart of the client’s case. Id. at 494–95. Thus, because allocating influence conflicts affect the core of a client’s expectations, they are especially problematic conflicts of interest.

61 See Epstein, supra note 28, at 580. In “agency cost” terms, the gap may be described as follows: “In any contract, it is easy to promise the moon, but tempting to deliver only a slice of green cheese. The promise may determine the scope of the obligation, but it is the performance that ultimately matters . . . .” Id. Even viewed in this context, however, a personal interest conflict emerges because “[t]he risks of self-interest are such that the attorney may not undertake actions that work for the benefit of the client because of the high costs” to himself of taking such actions. Id. at 580–81; see also Larry E. Ribstein, Ethical Rules, Agency Costs, and Law Firm Structure, 84 Va. L. Rev. 1707, 1707 (1998) (“Agency costs may arise when clients delegate discretion over their affairs to lawyers. Ethical rules are an important way to minimize these costs and ensure a minimum level of quality in legal services.”).

62 See discussion supra Part II.C.

63 In this sense, the lawyer’s resolution of an allocating influence conflict deprives a client of the primary benefit of the bargain. Although it is possible that market principles theoretically could punish lawyers who resolved allocating influence conflicts unethically, scholars have shown that the “repeat player” problem in allocating influence conflicts is not easily resolved by such market principles. See Macey & Miller, Regulatory State, supra note 15, at 1110–11.
heart of the client’s expectations, the failure to deliver on that expectation is even more harmful to the client than other types of resource allocation conflicts.64

In the best-case scenario, clients may temper their own influence-related expectations based on knowledge that the lawyer has other clients with similar expectations.65 Such clients likely know that they have to share the lawyer’s influence with other clients. What they do not know, however, is how the lawyer decides to divvy up that influence and whether it is allocated ethically. Most likely, even clients with “conflict awareness” expect the lawyer to allocate influence in a proper, ethical way. But absent any recognition of allocating influence conflicts or guidance on how to deal with them, proper influence allocation will not occur. Instead, lawyers will be tempted to resolve allocating influence conflicts based on other, perhaps unsavory, factors—such as which client gives the lawyer the most work per year. Accordingly, even sophisticated clients who know that conflicts likely will emerge are harmed by allocating influence conflicts.

IV. EXISTING ETHICS SOURCES REGARDING ALLOCATING INFLUENCE

Despite the frequency with which allocating influence conflicts occur and the resulting harm to clients if they are left unresolved or resolved improperly, there is little existing guidance regarding such conflicts in any recognized ethics source.66 Existing rules of professional conduct, comments to such rules, bar ethics opinions, and legal scholarship address conflicts of interest generally, but they do not explicitly address allocating influence conflicts.67 Similarly, federal and state statutes that address the ethics of influencing (regarding lobbying and associated disclosures) focus on reducing the potential harm to the public at large versus the harm to the client seeking lobbying assistance.68 Although these sources provide

---

64 See discussion supra Part II.C (distinguishing comparatively benign resource allocation issues).

65 Sophisticated clients likely understand this reality. For example, an in-house lawyer at a major wireless phone company likely understands, when hiring Lawyer X, that Lawyer X will meet with agency decision makers on behalf of other clients when he could be spending that time discussing the wireless phone company’s matter. However, the same understanding may not be held by a less sophisticated client. Further, even if understood, many clients may not “fix” the problem by hiring a different lawyer because only a small number of lawyers have the kind of access and influence critical to resolution of the client’s problem. See supra notes 39–42 and accompanying text; see also Macey & Miller, Regulatory State, supra note 15, at 1111 (predicting that the “repeat player problem” could not “easily be solved by having clients select law firms that were not repeat players before agencies” because that “solution . . . often results in higher costs for the clients . . . [a]nd, of course, there could be no guarantee that the neophyte law firm selected would not use the opportunity presented by the client to attempt to become a repeat player before the agency”).

66 See infra Part IV.

67 See discussions infra Part IV.A (model rules), Part IV.B (state rules), and Part IV.C (bar opinions).

68 See discussion infra Part IV.D (state and federal lobbying statutes).
no explicit identification or resolution of allocating influence conflicts, they do provide a solid foundation for addressing them. Thus, it is important to understand these potential sources of guidance before proposing solutions involving the sources and the ethical principles upon which they are based.

A. Model Rules Regarding Conflicts of Interest

In searching for ethical guidance on allocating influence, one first may consult the American Bar Association’s Model Rules of Professional Conduct (“Model Rules”). The general intent of the Model Rules is to ensure “the highest standards of professional competence and ethical conduct.” One practical purpose of the Model Rules is to provide “suggested” rules to state courts empowered to adopt and enforce ethical rules governing lawyers practicing in their states. Thus, although the Model Rules technically are not binding on anyone, they are binding to the extent that a state has adopted them in whole or in part.

Structurally, the Model Rules contain prohibitions on certain conduct while identifying other conduct as permissible or suggested, subject to a lawyer’s reasonable discretion. For example, the Model Rule regarding confidentiality states that a lawyer “shall not reveal” a client’s confidential information, thus prohibiting the release of client confidences. In contrast, the rule regarding fee

---


71 Id. at Chairperson’s Introduction (“The Model Rules of Professional Conduct are intended to serve as a national framework for implementation of standards of professional conduct.”).

72 As noted below, forty-nine states have adopted the Model Rules in whole or in part. See infra note 90 and accompanying text.

73 MODEL RULES OF PROF’L CONDUCT Scope ¶ 14 (2008) (“The Rules are thus partly obligatory and disciplinary and partly constitutive and descriptive in that they define a lawyer’s professional role.”).

74 Id. at R. 1.6(a). Like many of the Model Rules, the prohibition on revealing client confidences is subject to exceptions based on particular circumstances. Id. (“A lawyer shall not reveal information relating to the representation of a client unless the client gives
agreements merely acts as a suggestion, indicating that such agreements “preferably” be in writing.\textsuperscript{75}

Conflicts of interest fall into the first category—prohibitions—because the Model Rules generally prohibit a lawyer from continuing a representation involving a conflict of interest.\textsuperscript{76} The Model Rules further dictate how such conflicts should be avoided or resolved.\textsuperscript{77} In certain circumstances, the Model Rules indicate, conflicts may be waived by the client.\textsuperscript{78}

Under the Model Rules, conflicts of interest are categorized by the type of client involved—a current client, a former client, or a prospective client.\textsuperscript{79} With respect to current clients, the Model Rules recognize three types of conflicts of interest—direct adversity conflicts, material limitation conflicts, and personal interest conflicts.\textsuperscript{80} Condensing all three types of conflicts into a simple summary leads to the following proposition: when a current client is involved, a conflict of interest occurs if there is a “significant risk” that a lawyer’s representation of the client will be “materially limited” by (i) the lawyer’s responsibilities to another...
client or former client, (ii) the lawyer’s responsibilities to a third person, or (iii) the lawyer’s personal interests.  

Although one side of the conflict equation is broken down into specific interests—those of the client, a third person, or the lawyer himself—the other side of the equation is not so specific. Instead, the rule drafters would have us believe that it merely is a question of whether one of these specific interests, on the one hand, conflicts with a lawyer’s representation of a client, on the other hand. There is little guidance in the rules regarding particular types of lawyers or particular types of clients, despite the fact that within the practice of law there is limitless variation. Thus, a basic assumption underlying the Model Rules regarding conflicts is that they are written, or at least are intended to be written, broadly enough to encompass all practice scenarios.

Without a breakdown in the rules themselves, the primary practice-specific guidance is in the comments accompanying the conflict rules. Comments to the Model Rules often are a good source for practicing lawyers in search of more specific or nuanced guidance. For example, when legal scholars and others grew concerned that litigators were not identifying or addressing “positional” conflicts, the ABA modified an existing comment.

---

81 Id. at R. 1.7(a)(2).
82 See id. at R. 1.7.
83 See generally id. at pmbl. & Scope.
84 See id. at Scope ¶ 14 (“Comments do not add obligations to the Rules but provide guidance for practicing in compliance with the Rules.”). Although the conflict of interest rules do not make practice-area-based distinctions, certain other rules do make such distinctions. See, e.g., id. at R. 3.8 (detailing special duties of criminal prosecutors).
85 See id. at Scope ¶ 21 (“The Comment accompanying each Rule explains and illustrates the meaning and purpose of the Rule. . . . The Comments are intended as guides to interpretation, but the text of each Rule is authoritative.”).
86 Compare id. at R. 1.7 cmt. 24 (“Ordinarily a lawyer may take inconsistent legal positions in different tribunals at different times on behalf of different clients. The mere fact that advocating a legal position on behalf of one client might create precedent adverse to the interests of a client represented by the lawyer in an unrelated matter does not create a conflict of interest. A conflict of interest exists, however, if there is a significant risk that a lawyer’s action on behalf of one client will materially limit the lawyer’s effectiveness in representing another client in a different case; for example, when a decision favoring one client will create a precedent likely to seriously weaken the position taken on behalf of the other client. Factors relevant in determining whether the clients need to be advised of the risk include: where the cases are pending, whether the issue is substantive or procedural, the temporal relationship between the matters, the significance of the issue to the immediate and long-term interests of the clients involved and the clients’ reasonable expectations in retaining the lawyer. If there is significant risk of material limitation, then absent informed consent of the affected clients, the lawyer must refuse one of the representations or withdraw from one or both matters.”) with MODEL RULES OF PROF’L CONDUCT R. 1.7 cmt. 9 (2000) (“A lawyer may not represent parties having antagonistic positions on a legal question that has arisen in different cases, unless representation of either client would be adversely affected. Thus, it is ordinarily not improper to assert such positions in cases pending in different trial courts, but it may be improper to do so in cases
Thirty-five comments accompany the rule governing current-client conflicts of interest. Although some of these comments attempt to clarify how the conflict rules apply in various practice circumstances, none addresses administrative law practice. Nor does any comment address the problem of allocating influence within any practice.

B. State-Level Rules Regarding Conflicts of Interest

Although states are not obligated to adopt the Model Rules, forty-nine of them have adopted the Model Rules in whole or in part. Forty states also have adopted at least some of the comments accompanying the Model Rules. The Model Rules’ failure to anticipate, acknowledge, or address allocating influence conflicts of interest thus has led to similar failures in state-level rules and codes of professional conduct.

Forty-one states and the District of Columbia have adopted, in whole or in part, Model Rule 1.7 regarding conflicts of interest. State variations on the Model

---

88 See LUNEBURG & SUSMAN, supra note 4, at 495 (“Unfortunately, most of the Model Rules and their state counterparts seem to assume that lawyers are either litigators or transactions attorneys.”).
89 For practice-specific comments, see MODEL RULES OF PROF’L CONDUCT R. 1.7 cmts. 24, 26–27, 29–35 (2008) (cmt. 24 (litigation), cmt. 26 (transactional matters), cmt. 27 (trusts and estates), cmts. 29–33 (joint representation), & cmts. 34–35 (organizational clients)).
90 California is the only state with rules not based on the ABA Model Rules. See ABA’s Center for Prof’l Responsibility, Dates of Adoption of the Model Rules of Professional Conduct, http://www.abanet.org/cpr/mrpc/alpha_states.html (last visited Sept. 1, 2009). The District of Columbia and the Virgin Islands also follow the Model Rules’ format. Id. Effective Aug. 1, 2009, Maine was the forty-ninth state to adopt rules that more closely track the Model Rules. See MAINE RULES OF PROF’L CONDUCT (2009), available at http://www.courts.state.me.us/rules_forms_fees/rules/MRProfCond6-4-09.pdf.

---
Rules regarding conflicts of interest generally are minor, such as not requiring that a conflict waiver be in writing\textsuperscript{94} or replacing a phrase such as “a tribunal” with “any tribunal.”\textsuperscript{95} No state variation addresses allocating influence conflicts.\textsuperscript{96} Nor does any state rule or comment provide specific guidance regarding conflicts in administrative law practice.\textsuperscript{97} Thus, state rules and comments—like the Model Rules and comments—provide virtually no guidance regarding allocating influence conflicts.

C. ABA and State Bar Opinions

Given the frequency with which allocating influence conflicts occur in practice,\textsuperscript{98} one might expect ABA formal or state bar opinions to have addressed them. The lack of rule-based guidance, however, makes lawyer-generated complaints unlikely. Similarly, the lack of transparency to clients likely mutes any complaints from clients who suffer as a result of these conflicts.\textsuperscript{99} Because these two sources—fellow lawyers and clients themselves—are unlikely to lodge allocating influence complaints, it makes sense that ABA and state-level


\textsuperscript{94} \textit{Id.}; ILL. RULES OF PROF’L CONDUCT R. 1.7 (1990).
\textsuperscript{95} ABA Model Rule 1.7 Comparison, \textit{supra} note 93; CONN. RULES OF PROF’L CONDUCT R. 1.7(b)(3) (2007).
\textsuperscript{96} ABA Model Rule 1.7 Comparison, \textit{supra} note 93.
\textsuperscript{97} \textit{Id.}
\textsuperscript{98} Although Rule 8.3 and its state equivalents mandate that lawyers report known instances of professional misconduct by other lawyers, in many states the instances of actual reporting are few, and the instances of discipline for non-reporting are fewer. \textit{See} MODEL RULES OF PROF’L CONDUCT R. 8.3(a) (2008) (“A lawyer who knows that another lawyer has committed a violation of the Rules of Professional Conduct that raises a substantial question as to that lawyer’s honesty, trustworthiness or fitness as a lawyer in other respects, shall inform the appropriate professional authority.”); \textit{see also} LERMAN & SCHRAG, \textit{supra} note 18, at 99. (“While nearly every state has adopted a rule requiring lawyers to report misconduct of other lawyers, there are relatively few public reports of discipline of lawyers for not reporting.”).
\textsuperscript{99} \textit{See} Macey & Miller, \textit{Regulatory State, supra} note 15, at 1111; \textit{see also} Macey & Miller, \textit{Economic Analysis, supra} note 54, at 971 (“Because clients often cannot distinguish good legal work from bad, the client will rarely be an effective monitor of the attorney’s behavior (unless the client is a sophisticated party, such as a firm that employs its own in-house counsel.”).
disciplinary boards have not issued any formal opinions providing allocating influence conflict guidance.\textsuperscript{100}

\textit{D. Statutory and Executive Lobbying Restrictions}

Although parts A through C above address all typical sources of ethics guidance applicable to practicing lawyers, federal and state restrictions on lobbyists also are potentially relevant to the extent that they apply to administrative agency lawyers. This is because, fundamentally, much of the “influencing” done by a lawyer practicing before an administrative agency is identical to the influencing done by a traditional Congressional lobbyist.\textsuperscript{101} The location may be different—the offices of an agency in contrast to the halls of Congress—but the exercise of influence on behalf of a paying client is the same. As a result, many restrictions conventionally viewed as applicable to lobbyists also apply to administrative lawyers who influence agency decision makers.\textsuperscript{102}

Federal lobbying restrictions generally require the influence peddler, or lobbyist, to disclose his “lobbying contacts” in a quarterly report.\textsuperscript{103} These disclosure obligations appear, at first glance, to apply to people lobbying or influencing administrative agency officials.\textsuperscript{104} However, via the many exceptions

\textsuperscript{100} Bar opinions—such as those in Alabama, California, and Connecticut—have addressed how other ethical restrictions apply in particular administrative law circumstances. See \textit{Ala. OGC, Formal Op. 1993-12 (1993)} (advising whether a lawyer “may represent clients before a state agency even though [his] partner serves as a hearing officer for said agency”); State Bar of Cal. Standing Comm. on Prof’l Responsibility and Conduct, Formal Op. 2001-156 (2001) (advising whether a conflict of interest may arise “when . . . officials of a city . . . seek legal advice on the same matter and the constituents’ positions on the matter are antagonistic”); Conn. Bar Ass’n, Comm. on Prof’l Ethics, Informal Op. 95-7(1995) (concluding that “it is not a violation . . . for an attorney to communicate ex parte with [an agency] investigator, provided there is no attempt to influence the investigator unlawfully”).

\textsuperscript{101} Historians believe that the term “lobbyist” originated with President Ulysses S. Grant’s practice of speaking with “those who would wait in the lobby of Washington, D.C.’s Willard Hotel to smoke a cigar with [the] President . . . or to meet Congressmen.” Allard, supra note 2, at 37; see also Ron Smith, \textit{Compelled Cost Disclosure of Grassroots Lobbying Expenses}, 6 KAN. J.L. & PUB. POL’Y 115, 122, 170 & n.34 (1996). Of course, at the time of Grant’s presidency, administrative agencies did not exist to the extent they exist today, or else perhaps they, too, would have joined in the fun.

\textsuperscript{102} See \textit{infra} notes 103–114 and accompanying text.

\textsuperscript{103} 2 U.S.C. § 1604 (2006); \textit{see also} 2 U.S.C. § 1602 (2006) (defining “lobbying contact”). The substance of the required disclosure typically includes the name of the lobbyist and client, a brief list of the issues to be addressed in the lobbying, identification of the House of Congress or federal agency lobbied (but not individual names), and estimated amounts of income from the lobbyist’s client(s). 2 U.S.C. § 1604(b) (2006).

\textsuperscript{104} \textit{See LUNEBURG & SUSMAN}, supra note 4, at 177–78 (noting that, after the Lobbying Disclosure Act of 1995, “communications with executive branch officials related to the administration of federal laws are considered lobbying and may trigger registration
listed in the definition of “lobbying contacts” and the limited definition of “covered executive branch official,” much of the influence exerted by administrative lawyers likely is not covered by federal lobbying restrictions.105

The exact applicability of the federal lobbying restrictions to administrative agency influencing is outside the scope of this Article. What is relevant, however, is the lack of guidance in these lobbying restrictions regarding allocating influence conflicts and the like even if they do apply to agency-influencing activities.106 The lack of federal statutory guidance regarding conflicts of interest is not surprising when one considers the primary purpose of these laws. Specifically, federal lobbying restrictions are intended to protect the general public via disclosure and the scrutiny such disclosure facilitates.107 Even President Barack Obama’s recent executive order, “Ethics Commitments by Executive Branch Personnel,” only applied to the pre- or post-government lobbying activities of government personnel to preclude suspicious “side-switching” and increase public confidence in government.108

Although the purposes of these congressional and executive efforts may be laudable, they do not include express protections for another participant in the influence exchange: the influence seeker, or client.109 Rather, the only express

105 See id. at 179 (“The LDA identifies a relatively narrow list of officials with whom communications are considered lobbying contacts.”); id. at 182 (“The LDA includes a long list of exceptions that, to some extent, contract the broad scope of the LDA’s definition of a lobbying contact.”). Another provision of the LDA that contracts its applicability to administrative agency influencers is the “twenty percent rule,” under which 20 percent of the time one spends on behalf of a client must be dedicated to “lobbying activities” in order for one to be deemed a “lobbyist.” See id. at 37 (citing 2 U.S.C. § 1602(10) (2004)).

106 See id. at 499 (noting that lawyer-lobbyists face a difficult challenge “because there is little guidance in the [ethics] literature on applying the conflict rules to the nuances of a lobbying practice”).


109 See William V. Luneburg, Anonymity and Its Dubious Relevance to the Constitutionality of Lobbying Disclosure Legislation, 19 STAN. L. & POL’Y REV. 69, 88 (2008) (“Unlike certain special relationships, like those involving the attorney and client or the association and member, where the law casts a veil of protection over the
protections for the client are set forth in legal ethics rules which, as discussed above, do not adequately address allocating influence conflicts. Thus, federal lobbying restrictions can be eliminated as a potential source of adequate guidance regarding these conflicts.

As with their federal counterparts, state lobbying restrictions focus on disclosure by the lobbyist of the client’s identity, the issue discussed, and the compensation. Even the most recent state lobbying reform efforts—many of which were enacted in response to specific scandals—are directed solely at increased disclosure or at establishing commissions to further investigate allegations of impropriety regarding government employees. Although Pennsylvania recently extended its lobbying disclosure regulations to any attempt to influence “administrative action,” it is too soon to predict whether this effort, or future efforts using it as a model, will provide the kind of direct guidance necessary to prevent and respond to allocating influence conflicts.

V. ADDRESSING THE ALLOCATION OF INFLUENCE THROUGH THE CONFLICT OF INTEREST RULES

The proposal of detailed solutions to allocating influence conflicts may be viewed as premature until a consensus emerges regarding their existence and harms. However, when the allocating influence “problem” is viewed as an actual conflict of interest similar to those already recognized, then the need to address the problem—and the structure in which to do so—becomes quite clear. Ultimately, if one concedes that allocating influence conflicts exist and merit focused guidance, then the most direct way to provide that guidance is to add a comment, and

communications among the parties to the relationship, there is no similar special relationship between lobbyist and client, on the one hand, and legislator or administrator, on the other, that justifies confidentiality of communications.”).  

See discussion supra Part IV.A.  


improve upon an existing comment, to Model Rule 1.7 regarding conflicts of interest.

A. The Allocating Influence “Problem” Is a Conflict of Interest That Should Be Addressed in the Conflict of Interest Rules

Fundamentally, a conflict of interest is a conflict of duties. If a lawyer feels as though his duty (of loyalty, diligence, confidentiality, or another) to one person conflicts with his duty to another person or client (including himself), then he may face a conflict of interest. Under the Model Rules of Professional Conduct, a current-client conflict of interest occurs if there is “direct adversity” or if there is a “significant risk” that a lawyer’s representation of one client will be “materially limited” by the lawyer’s personal interests or by responsibilities to another client, to a former client, or to a third person. Material limitation from any of these sources—another client, a former client, a third person, or the lawyer himself—is sufficient to create a conflict of interest.

An allocating influence dilemma involves significant risk of material limitation from at least three, if not all four, of these sources, making it a clear conflict of interest. The scenario presented above—a lawyer with one ex parte meeting (or a single opportunity to exercise influence) but more than one client (A and B) with a need for such influence—illustrates the presence of these risks and limitations. In that scenario, there is a significant risk that the lawyer’s responsibilities to Client A will be materially limited by his responsibilities (i) to Client B, because use of the twenty-minute meeting to discuss Client A’s matter will make it much more difficult to influence the decision maker on behalf of Client B in the future; (ii) to a third person, the decision maker, with whom the lawyer seeks to maintain an influential relationship; and, similarly, (iii) to the lawyer himself, whose future prospects depend upon maintaining an influential relationship with the decision maker. Thus, in a frequently faced influence allocation scenario, a conflict of interest emerges on several levels.

115 Underlying the lawyer-client relationship and the conflict of interest rules is an expectation that the lawyer’s representation will not be limited by his responsibilities to another. MODEL RULES OF PROF’L CONDUCT R. 1.7 cmt. 1 (2008) (“Loyalty and independent judgment are essential elements in the lawyer’s relationship to a client. Concurrent conflicts of interest can arise from the lawyer’s responsibilities to another client, a former client or a third person or from the lawyer’s own interests.”).

116 Id. at R. 1.7(a).

117 Id. (“Except as provided in paragraph (b), a lawyer shall not represent a client if the representation involves a concurrent conflict of interest. A concurrent conflict of interest exists if: (1) the representation of one client will be directly adverse to another client; or (2) there is a significant risk that the representation of one or more clients will be materially limited by the lawyer’s responsibilities to another client, a former client or a third person or by a personal interest of the lawyer.”).

118 See discussion supra Part II.A.

119 See Macey & Miller, Regulatory State, supra note 15, at 1107 (“[For] private-
B. Treating Allocating Influence Issues as Conflicts of Interest Is a Natural Extension of Previous Conflict of Interest Scholarship

As explained above, viewing allocating influence dilemmas as conflicts of interest is consistent with the conflict of interest rules and the core principles underlying those rules. Treating allocating influence issues as conflicts of interest also is a natural extension of previous conflict of interest scholarship. Although no existing scholarship expressly identifies allocating influence conflicts as done in this Article, a few scholars have addressed one critical aspect of the problems that supports their treatment as conflicts of interest: aspects of administrative law practice lead to frequent, troubling conflicts between a client’s interest and the lawyer’s personal reputational interest.

Three preeminent ethics scholars have suggested that administrative lawyers often are tempted to put their personal reputational interests ahead of a client’s interest. In the 1970s, renowned ethics scholar Monroe Freedman first shared this concern as part of a broader exposé regarding potential ethical abuses at the Securities and Exchange Commission. Essentially, Freedman opined that SEC lawyers, through intimidating and likely unethical practices, were causing the securities bar to sacrifice zealous representation of their clients. Although not expressly stated as such, Freedman’s concern regarding the lack of zealous representation ultimately was a concern that the conflict of interest between the client and the lawyer resulted in a loss of independent judgment.122 Thus, the problem first identified by Freedman is best understood as a conflict of interest that emerges most often in administrative law practice in general, and not just at the SEC.

Freedman’s SEC-related concerns reappeared in the 1990s as part of Jonathan R. Macey and Geoffrey P. Miller’s Reflections on Professional Responsibility in a Regulatory State. In Regulatory State, Macey and Miller submit that market mechanisms that typically control lawyer behavior and mitigate ethical concerns often do not work effectively in the “regulatory sector.” One example of “market failure” that Macey and Miller cite is the problem Freedman first identified, and which they artfully label the “repeat player problem.”

sector lawyers who represent clients before government agencies . . . the client’s interests can be subjugated to the long-term interest of the lawyer or law firm in maintaining a cordial relationship with the particular agency . . . .”

120 Freedman, supra note 49, at 285.
121 See id.
122 See id.
123 See Macey & Miller, Regulatory State, supra note 15, at 1105.
124 See id. at 1106.
125 See id. Another conflict of interest scholar, John S. Dzienkowski, used the “repeat player” terminology but did so when referring to the client, not the lawyer. Dzienkowski, supra note 10, at 486–87 (noting that, in a positional conflict scenario, “the one-time litigant may be most at risk when the other representation involves a repeat player”).
Because administrative lawyers, unlike other lawyers, often must appeal to the same agency officials on behalf of multiple clients over a long period of time, they are viewed as “repeat players.”

These repeat players have a personal interest in maintaining a cordial, working relationship with the agency officials whom they repeatedly must influence. Indeed, lawyers’ solid relationships with these officials often form the primary basis of their practice. In accordance with these long-term personal interests, repeat players are tempted to sacrifice clients’ immediate interest in zealous and persistent advocacy to protect and maintain the lawyers’ fruitful relationships with agency officials. In other words, there is a conflict between the client’s interest in dedicated advocacy and the lawyer’s personal interest in maintaining the relationship for future use.

Like Freedman, Macey and Miller ultimately are concerned that a lawyer might kowtow to agency officials at the expense of clients. Although not explicitly stated, these concerns were justified given existing conflict of interest principles. More importantly for present purposes, the scholars’ identification of the repeat player problem as the cause of personal interest conflicts fundamentally supports treating allocating influence dilemmas as conflicts of interest.

---

126 See Macey & Miller, Regulatory State, supra note 15, at 1106.
127 See id. at 1106, 1111.
128 See id. at 1110.
129 One also could characterize this conflict of interest as an agency cost. See Ribstein, supra note 61, at 1709 (“Agency costs typically involve conflicts between the agent’s and principal’s interests.”). Viewed in these terms, the “[l]awyer’s interests . . . may diverge from those of clients in terms of their willingness to take risk” because “[l]awyers stand to lose in terms of reputation . . . from a bad result but do not share in the client’s gain from a good result.” Id. at 1710. Essentially, this leads to a personal interest conflict. Another personal interest conflict is that “[l]awyers . . . may have selfish reasons to favor the interests of one client over those of another.” Id. at 1709.
130 In identifying this sacrifice of clients’ interests, Freedman primarily assigned blame to the agency instead of to the private sector lawyers practicing before that agency. Specifically, Freedman accused the SEC of “dragooning members of the private Bar . . . as federal police and prosecutors of their own purported clients[,]” Freedman, supra note 51, at 281, and of “depend[ing] upon intimidation of individuals, business firms, and attorneys through aggressive abuse of its power over the economic life and death of those subject to its jurisdiction.” Id. at 282. Macey and Miller, on the other hand, tended to believe that this blaming of SEC lawyers was misplaced. See Macey & Miller, Regulatory State, supra note 15, 1110 (“We believe that Professor Freedman is wrong, however, to blame the lack of forceful representation by the securities bar solely on the misconduct of the SEC, although we have no reason to doubt his analysis in this regard.”).
131 At one point, Macey and Miller describe the repeat player problem as a conflict of interest, but they do not demonstrate how this is so, do not offer a concrete definition of the conflict, and do not offer any guidance regarding how to avoid or resolve such conflicts, presumably because the repeat player problem primarily served as an example to support their more general thesis that the market could not effectively bar certain ethical issues in the regulatory sector. See Macey & Miller, Regulatory State, supra note 15, at 1105–06, 1109.
C. Treating Allocating Influence Dilemmas as Conflicts of Interest Is Consistent with Existing Guidance Regarding Positional Conflicts

Previous scholars’ identification of the repeat player problem—and the conflicts of interest inherent to it—supports treating the allocating influence problem as a conflict of interest. The bar’s current standards on positional conflicts provide further “precedent” for treating allocating influence dilemmas as conflicts of interest. As described above, the Model Rules prohibit certain conflicts of interest in Rule 1.7 and flesh out the applicability of the prohibition to certain practice scenarios via the comments. One scenario addressed in the “Conflicts in Litigation” portion of the comments is a special type of conflict of interest known as a “positional conflict.”

As set forth in the Model Rules, a positional conflict of interest emerges “if there is a significant risk that a lawyer’s action on behalf of one client will materially limit the lawyer’s effectiveness in representing another client in a different case; for example, when a decision favoring one client will create a precedent likely to seriously weaken the position taken on behalf of the other client.” The comment further cites various factors to consider when determining whether to advise a client regarding a positional conflict, including whether the respective cases are at the appellate or trial level and “the client’s reasonable expectations in retaining the lawyer.” If a positional conflict results in “significant risk of material limitation,” then the lawyer must choose between one of the conflicting positions and withdraw from the other or, as with other conflicts of interest, obtain a client waiver.

---

133 Id. at R. 1.7 cmt. 24.
134 Id. Dzienkowski defined a positional conflict as follows: “A positional conflict of interest occurs when a law firm adopts a legal position for one client seeking a particular legal result that is directly contrary to the position taken on behalf of another present or former client, seeking an opposite legal result, in a completely unrelated matter.” Dzienkowski, supra note 10, at 460. Although some have questioned whether positional conflicts should be analyzed differently, see Anderson, supra note 4, at 5, 30–37, the general consensus is that positional conflicts are conflicts of interest in certain defined circumstances and the rules, bar opinions, and other sources generally support this understanding. See, e.g., DEBORAH L. RHODE & DAVID LUBAN, LEGAL ETHICS 610–13 (5th ed. 2009) (describing positional conflicts via a discussion of case examples and existing legal scholarship).
135 MODEL RULES OF PROF’L CONDUCT R. 1.7 cmt. 24 (2008) (“Factors relevant in determining whether the clients need to be advised of the risk include: where the cases are pending, whether the issue is substantive or procedural, the temporal relationship between the matters, the significance of the issue to the immediate and long-term interests of the clients involved and the clients’ reasonable expectations in retaining the lawyer.”). See also supra notes 54–61 and accompanying text (explaining significance of client’s expectations in analyzing conflicts of interest).
136 MODEL RULES OF PROF’L CONDUCT R. 1.7 cmt. 24 (2008) (“If there is significant risk of material limitation, then absent informed consent of the affected clients, the lawyer..."
The concerns that motivated expansion of the positional conflict definition align significantly with the concerns that underlie allocating influence conflicts. Although positional conflicts often are understood as concurrent client conflicts (i.e., taking a position for one client that conflicts with the position taken for another conflict), they also involve a conflict between the client and the lawyer’s personal interest (i.e., a personal interest conflict). Positional conflicts involve personal interest conflicts in that a lawyer considering whether to make an argument inconsistent with one already made for another client must balance the must refuse one of the representations or withdraw from one or both matters.”); id. at R. 1.7(b) (consent). The ABA first expressly acknowledged positional conflicts in 1983. Anderson, supra note 4, at 13; see also id. at 13–15 (positing that “enormous growth and specialization of the legal profession” along with a “philosophical” shift toward realism motivated the ABA to address positional conflicts in the Model Rules). Previously, the Model Rules and comments suggested that positional conflicts existed only if one simultaneously took antagonistic positions in appellate courts. Prior to its amendment in 2002, the positional conflict comment read as follows:

A lawyer may represent parties having antagonistic positions on a legal question that has arisen in different cases, unless representation of either client would be adversely affected. Thus, it is ordinarily not improper to assert such positions in cases pending in different trial courts, but it may be improper to do so in cases pending at the same time in an appellate court.

MODEL RULES OF PROF’L CONDUCT R. 1.7 cmt. 9 (2000). Compare id., with RESTATEMENT (THIRD) OF THE LAW GOVERNING LAWYERS § 128 cmt. f (2000) (“A lawyer ordinarily may take inconsistent legal positions in different courts at different times. While each client is entitled to the lawyer’s effective advocacy of that client’s position, if the rule were otherwise law firms would have to specialize in a single side of legal issues. However, a conflict is presented when there is a substantial risk that a lawyer’s action in Case A will materially and adversely affect another of the lawyer’s clients in Case B. Factors relevant in determining the risk of such an effect include whether the issue is before a trial court or an appellate court; whether the issue is substantive or procedural; the temporal relationship between the matters; the practical significance of the issue to the immediate and long-run interests of the clients involved; and the clients’ reasonable expectations in retaining the lawyer. If a conflict of interest exists, absent informed consent of the affected clients under § 122, the lawyer must withdraw from one or both of the matters.”). Scholarly criticism regarding the limited reach of the ABA’s positional conflict of interest definition led to an ABA formal opinion expanding the definition and, ultimately, to a more expansive Model Rule comment. See ABA Comm. on Ethics and Prof’l Responsibility, Formal Op. 93-377 (1993); Anderson, supra note 4, at 15–18. The two primary lines of criticism were the comment’s ambiguity and its failure to say anything about “notice and disclosure of positional conflicts to clients.” Anderson, supra note 4, at 15. A 1993 ABA ethics opinion rejected the distinction between trial and appellate courts in favor of a more nuanced, factor-based inquiry. Id. at 15–17 (summarizing ABA Comm. on Ethics and Prof’l Responsibility, Formal Op. 93-377 (1993)). Thus, in 2002, the ABA modified a comment to Rule 1.7 regarding “positional” conflicts of interest to read as described above. Id. at 17. 137 See MODEL RULES OF PROF’L CONDUCT R. 1.7 cmt. 10 (2008).
second client’s interest with his own personal interest in keeping the first client happy and inclined to send more work the lawyer’s way. Essentially, “where a client provides repeat business, a lawyer would not want to offend that client by taking on a case or making an argument of which that client disapproves.”

In this respect, a positional conflict—or at least the avoidance of a positional conflict—may be more properly viewed as a personal interest conflict and a concurrent client conflict. Because it involves both types of conflicts, a positional conflict shares much in common with an allocating influence conflict. Given these fundamental similarities, it makes sense to use a comment-based solution, similar to the positional conflict comment, to address allocating influence conflicts of interest.

D. The ABA Should Amend the Comments to Model Rule 1.7 to Address Allocating Influence Conflicts

A comment regarding allocating influence conflicts could track the structure already used with respect to positional conflicts—a description of when an allocating influence conflict exists, supplemented by an example, followed by a general but consentable prohibition. The descriptive or definitional element largely could track the definition presented above, followed by a specific example from administrative law practice. Together, these elements would help practicing lawyers determine whether they face an allocating influence conflict. Further, the comment should identify the circumstances in which an allocating influence conflict may be waived and whether it may be waived prospectively.

Based on the framework already in place in the Model Rules, I propose the following amendment to the comments following Model Rule 1.7, heading “Nonlitigation Conflicts,” to be inserted after current comment 27:

---

138 Anderson, supra note 4, at 31–32 (demonstrating that the avoidance of a positional conflict of interest involves a “business conflict” for the lawyer).

139 Id. at 3. Anderson further posits that “[i]t is when a lawyer decides not to make a contrary argument for one client in order to avoid offending or harming another client that an ethical problem is likely to be present.” Id. at 3–4 (emphasis in original). In this respect, the lawyer may avoid a pure positional conflict but does so based on prioritizing his own personal or economic interest, which may be a more serious problem than the positional conflict the lawyer sought to avoid. Id. at 31–32.

140 One also could list factors that make such a conflict more or less likely to be present in any individual case, as done with positional conflicts. See MODEL RULES OF PROF’L CONDUCT R. 1.7 cmt. 24 (2008).

141 See, e.g., id. at 1.7(b) (defining circumstances in which client may waive or consent to a conflict of interest).

142 Because allocating influence conflicts of interest occur most frequently in administrative law practice, any comment addressing them likely should be placed under the “Nonlitigation Conflicts” heading instead of the “Conflicts in Litigation” heading. Current comments under the Nonlitigation Conflicts heading address practice areas such as transactional matters, estate planning, and negotiating a settlement. See id. at R. 1.7 cmt. 26 (referencing comment 7, noting that “[d]irectly adverse conflicts can also arise in
A lawyer who properly may, and is retained by a client to, influence a decision maker via personal contact (such as a lawyer lobbying Congress or a lawyer meeting with an administrative agency official) often may face a conflict of interest when he must choose how to allocate a finite amount of influence among multiple clients. This type of conflict, an allocating influence conflict, occurs if there is a significant risk that allocating influence on behalf of one client is reasonably certain to inhibit substantially the lawyer’s ability to influence the same decision maker on behalf of another client. If such a significant risk is present, then, absent informed consent of the affected client(s) under 1.7(b), the lawyer shall refuse one of the representations or withdraw from one or both matters.143

Because an allocating influence conflict of interest also involves a conflict between one client’s interest and the lawyer’s own personal interest in maintaining a relationship with a decision maker, guidance specific to this type of “personal interest” conflict is appropriate. Accordingly, I propose that comment 10 following Model Rule 1.7, heading “Personal Interest Conflicts,” be revised as follows, with new language in italics:

The lawyer’s own interests should not be permitted to have an adverse effect on representation of a client. For example, if the probity of a lawyer’s own conduct in a transaction is in serious question, it may be difficult or impossible for the lawyer to give a client detached advice. Similarly, when a lawyer has discussions concerning possible employment with an opponent of the lawyer’s client, or with a law firm representing the opponent, such discussions could materially limit the

transactional matters, with an example), cmt. 27 (describing how “conflict questions may arise in estate planning and estate administration”), & cmt. 28 (addressing circumstances in which nonlitigation conflicts may be waived). The fact that comment 26 begins by stating that “[c]onflicts of interest under paragraphs (a)(1) and (a)(2) arise in contexts other than litigation” further supports the placement of allocating influence conflict guidance.143

As an alternative approach, one could advise lawyers how to resolve allocating influence conflicts based on certain preferred factors rather than prohibiting them altogether. Under such an approach, the comment could continue as follows: “In circumstances in which obtaining prior consent is not practicable, a lawyer may continue the representation, despite the allocating influence conflict, only if the lawyer resolves the conflict based on the lawyer’s professional judgment and not based on unreasonable, subjective factors such as by favoring the client that provides the lawyer with the most annual billable hours.” This resolution approach, in contrast to a prohibition approach, may be preferable to the extent that allocating influence conflicts, like positional interest conflicts, “cannot be resolved by a simple rule that permits or prohibits their existence; instead, a series of considerations must determine how a particular positional conflict of interest is handled.” See Dzienkowski, supra note 10, at 520–21 & n.257.
lawyer’s representation of the client. In addition, a lawyer may not allow related business interests to affect representation, for example, by referring clients to an enterprise in which the lawyer has an undisclosed financial interest. Finally, a lawyer who properly may, and is retained by a client to, influence a decision maker via personal contact (such as a lawyer lobbying Congress or a lawyer meeting with an administrative agency official) shall not permit his personal interest in maintaining an ongoing beneficial relationship with a decision maker to materially limit the lawyer’s representation of the client. See Rule 1.8 for specific Rules pertaining to a number of personal interest conflicts, including business transactions with clients. See also Rule 1.10 (personal interest conflicts under Rule 1.7 ordinarily are not imputed to other lawyers in a law firm).

These two modest revisions to the Model Rule comments would serve several purposes, including mitigation of the harms documented above. At the very least, the comments would alert lawyers who lobby or influence decision makers that they face a special obligation—in addition to those imposed upon non-lawyer lobbyists—to exercise their influence consistent with a lawyer’s general duty of independent judgment for each and every client. Further, this increased awareness, combined with the obligation to avoid allocating influence conflicts or obtain client consent to such conflicts, would better ensure that clients are aware of how they will be represented. Clients, empowered with this information, can then make an informed decision either to continue with a representation despite the presence of an allocating influence conflict, or to hire another lawyer.

144 See supra Part III.A.
145 Increasing awareness in this manner is one of the primary purposes of the comments. See MODEL RULES OF PROF’L CONDUCT Scope ¶15 (2008) (stating comments “are sometimes used to alert lawyers to their responsibilities . . . .”).
146 Although permitting advance waivers of conflicts of interest admittedly may “raise[] several questions,” such as whether a client may be “consent[ing] to a problem without a full appreciation of the facts and the potential damage,” two aspects of an allocating influence conflict make such questions less critical. See Dzienkowski, supra note 10, at 528. First, because an allocating influence conflict is, at its core, a resource allocation problem, which many clients deal with often in their day-to-day lives, it is very likely that clients will be able to appreciate the nature of the problem. Second, because ex parte meetings with agency decision makers generally must be documented through a post-meeting, publicly available notice, it at least is possible that clients may track how “thinly spread” their lawyers’ influence is at any given time. The easy-to-understand nature of the conflict, coupled with relatively easy monitoring capabilities, should balance any concerns regarding advance waivers. See id. at 528–29 (questioning the efficacy of an advance waiver because “it requires a client to consent to a problem without a full appreciation of the facts and the potential damage” and because “once the conflict begins to affect the representation, the lawyer would need to obtain a new consent based on the new information”); c.f. Macey & Miller, Economic Analysis, supra note 54, at 1004 (suggesting that per se bans on attorney conflicts of interest are unwarranted, at least in purely
VI. CONCLUSION

A lawyer’s influence is a finite resource for which a client pays significant value. Because the exercise of this influence in the client’s favor is the heart of a client’s expectations, the lawyer must allocate this resource ethically and with due regard to conflict of interest principles. Regarding conflicts of interest generally, lawyers are subject to ethical rules intended to protect clients. Lobbyists, on the other hand, are subject to state and federal statutes intended to protect the general public. Although it is understood that these sources of ethical guidance apply to lawyers practicing before administrative agencies, exactly how they apply is, at best, unclear. One particular area in which lawyers need additional guidance is how the legal ethics “conflict of interest” concept applies to a lawyer attempting to influence an agency decision maker. As demonstrated in this Article, providing concrete guidance in this area is important given the frequency of, and harm caused by, allocating influence conflicts. The best way to provide this necessary guidance is in the comments to the Model Rules themselves. Only after doing so will we, as a profession, begin to ensure that lawyers are allocating influence ethically.
Collateral Damage and Securities Litigation

Bradford Cornell and James C. Rutten*

Damages arise in securities cases when misstatements or omissions inflate the price of a security above its true value. Investors who buy the securities at these inflated prices are damaged when subsequent revelation of the misstatements or omissions causes the price of the securities to return to their fair value. But what if the subsequent revelation causes the price to fall much more than the original inflation? For example, with respect to stock option backdating, economic research indicates that failure to account properly for employee options does little to inflate security prices, but in numerous instances the subsequent revelation of option backdating led to significant stock price declines. The term coined for the decline in excess of the original inflation is “collateral damage.” Here we examine collateral damage from both an economic and legal perspective. We conclude that while collateral damage can have a material impact on securities prices, declines associated with collateral damage are not, and should not be, recoverable under section 10(b) of the Securities Exchange Act of 1934.

Table of Contents

I. Introduction .................................................. 718
II. Over-Disclosure and Collateral Damage ................. 719
III. Value Creation and Collateral Damage ................... 723
IV. Does Section 10(b) Permit Recovery for Collateral Damage? ...... 727
   A. The Supreme Court’s Analytical Framework for Construing the Scope of the Implied Right of Action Under Section 10(b) ........................................... 730
   B. The Elements of the Section 10(b) Cause of Action .................... 738
      1. The Existence of a Duty to Disclose ........................................... 738
      2. Materiality .............................................................................. 741
      3. Transaction Causation .............................................................. 743
      4. Loss Causation ........................................................................ 744
      5. The “In Connection With” Requirement ................................. 745
V. Conclusion ....................................................... 747

* © 2009 Bradford Cornell & James C. Rutten. The authors are grateful to Mathew Dark and Miriam Steele of Allens Arthur Robinson, and John Haut of Charles River Associates for helpful comments.
I. INTRODUCTION

Despite recent rulings, including the groundbreaking decision in *Dura Pharmaceuticals, Inc. v. Broudo*,¹ the concept of inflation remains the linchpin of damages analysis in securities litigation. If a misrepresentation or omission fails to inflate the price of a company’s securities, then damages are zero regardless of any other considerations, and there is no reason for plaintiffs to proceed. Nonetheless, the problem of measuring inflation has received less attention in recent years as focus has shifted to other issues, such as causation. This has led to conceptual gaps in the legal and financial analysis of inflation. This Article takes a step in redressing that deficiency.

The best way to highlight the contribution of this Article is first to explain what it does not address. This is not an article on the economic and legal issues associated with event studies. It is assumed that the standard event study approach accurately measures the impact of a company’s disclosures on its stock price. More specifically, it is assumed that a model can be developed that produces residual returns accurately reflecting the valuation impact of information disclosures.² This is, of course, a highly controversial assumption. In most securities cases, there is active debate regarding which model should be used to net out market and industry effects. Once a model is selected, there remains the problem of confounding information that arises when more than one disclosure occurs on a given day. In addition, the authors demonstrate³ that by using hindsight in combination with the market inefficiency proven to exist by Grossman and Stiglitz,⁴ plaintiffs can select those unique days on which the residual return markedly overstates the fundamental value of the information disclosed. These and other issues related to event studies are all important, but they all have been analyzed extensively. Because this article focuses on conceptual issues unrelated to the application of

³ See Bradford Cornell & James C. Rutten, *Market Efficiency, Crashes and Securities Litigation*, 81 TUL. L. REV. 443, 443 (2006) (“In light of theoretical and empirical research in finance, we show that the failure to understand the differing implications of the efficient market hypothesis for proving reliance and assessing damages introduces a significant plaintiff’s bias in securities class action litigation.”).
⁴ See Sanford J. Grossman & Joseph E. Stiglitz, *On the Impossibility of Informationally Efficient Markets*, 70 AM. ECON. REV. 393, 405 (1980) (“We have argued that because information is costly, prices cannot perfectly reflect the information which is available, since if it did, those who spent resources to obtain it would receive no compensation.”).
event study techniques, it is assumed that all of these problems can be surmounted. It is important to stress that in making this assumption it also is implicitly assumed that the market is efficient. If such were not the case, then the residual return would not necessarily reflect the fundamental value of an information disclosure.

Given these assumptions, estimating inflation is straightforward in the case of a company making an unexpected and false statement. An example would be a company’s announcement that new oil reserves had been discovered on the company’s property when, in fact, no such discovery had occurred. Under these circumstances, the residual return on the day of announcement would measure the inflating impact of the misrepresentation. The problem is that such cases are rare. More commonly, misstatements involve expected announcements, half-truths, and omissions. A common example is a situation in which a company artificially inflates stated earnings to meet the estimates of Wall Street analysts. In that situation, the stock price of the company should be unchanged by the earnings announcement because expectations are met. Inflation occurs because the stock price should have fallen to its “true value” had the correct financial information been disclosed. In this situation, inflation cannot be measured simply by looking at the residual return on the day of misrepresentation because there is no residual return.

The standard solution to this problem has been to go forward in time to the date on which the misinformation is corrected and the stock price falls. The event study approach is used to estimate the residual return on that date. It is assumed that this residual return—i.e., the valuation impact of the corrective disclosure—is equal to the amount of the original inflation. We argue in this Article that this final critical assumption is typically faulty. The reaction to a corrective disclosure will commonly exceed—often by a large amount—the inflation caused by the original misstatement. This occurs not because the event study approach fails to measure properly the impact of the corrective disclosure—we assume that it does—but because the residual return associated with the disclosure is a poor estimate of the original inflation. The remainder of the Article explains why and discusses the legal implications of this assertion.

II. OVER-DISCLOSURE AND COLLATERAL DAMAGE

There are two major reasons why the valuation impact of a disclosure that corrects a misstatement can exceed the amount of the inflation caused by the original misstatement: over-disclosure and collateral damage. Of the two, over-disclosure has been analyzed more extensively, beginning with Cornell and Morgan. For that reason, our analysis focuses on collateral damage. Nonetheless,

---

5 It is possible that the corrective disclosure occurs over several days. While this complicates the computations, it does not change the nature of the conceptual issues analyzed here. Therefore, to avoid unnecessary complication, it is assumed that the corrective disclosure occurs in one day.

6 See Cornell & Morgan, supra note 2, at 889–97.
before turning to a discussion of collateral damage, it is helpful to review briefly what is meant by over-disclosure. Basic, Inc. v. Levinson⁷ provides a good example despite the fact that the misstatement at issue was associated with deflation rather than inflation. Except for the direction of the price movement caused by the misstatement, the issues related to the analysis of inflation are identical.

In Basic, representatives of Basic met with and called representatives of Combustion Engineering in September 1976 regarding a possible merger.⁸ During 1977 and 1978, however, Basic made public statements denying any potential merger.⁹ Then, on December 18, 1978, Basic abruptly asked the New York Stock Exchange to suspend trading in its shares pending a news announcement.¹⁰ The following day, Basic’s board of directors endorsed Combustion Engineering’s $46 per share offer for its common stock. On December 20, the agreed-upon merger was publicly announced.¹¹

If the merger announcement is taken as the corrective disclosure, it clearly “over-discloses” previous misstatements.¹² While Basic was denying any potential merger in 1977 and early 1978, it knew that its statements were false, but only to the extent that merger discussions were in fact occurring. Basic did not know the merger negotiations would be successful. Consequently, the corrective disclosure that announced the merger agreement did more than correct the original misstatements; it represented an over-disclosure that arguably had a greater impact on the stock price than a disclosure of ongoing merger discussions would have had.

The solution to the over-disclosure problem is to estimate what Cornell and Morgan call “the equivalent disclosure.”¹³ The equivalent disclosure is the price at which a security would have traded if the omitted and misrepresented information—and only that information—had been accurately disclosed in the first place.¹⁴ Though easy to state in theory, estimating the equivalent disclosure price is difficult in practice. Courts and scholars have struggled with this issue since Elkind v. Liggett & Myers, Inc.¹⁵ It remains a vexing problem, but not one we address here. This Article’s focus is collateral damage.

To our knowledge, the term “collateral damage” was coined by Ferrell and Saha¹⁶ but their application of the term is confusing because they use it in two different ways. On the one hand, they discuss collateral damage in the context of an accounting fraud and note that the disclosure of an accounting restatement can

---

⁸ Id. at 227.
⁹ Id.
¹⁰ Id. at 227–28.
¹¹ Id. at 228.
¹² Cornell & Morgan, supra note 2, at 889–90.
¹³ Id. at 894–97.
¹⁴ Id. at 894.
¹⁵ 472 F. Supp. 123, 129 (S.D.N.Y. 1978) (explaining an expert’s analysis of how disclosure would have affected a defendant corporation’s stock price), aff’d in part and rev’d in part, 635 F.2d 156 (2d Cir. 1980).
¹⁶ See Ferrell & Saha, supra note 2, at 179–85.
have collateral damage effects, including reassessment of the quality of a firm’s management and/or internal controls, as well as possible disruptive legal action.\(^{17}\) On the other hand, they argue at another juncture that misstatements can be divided into two categories.\(^{18}\) In the first category are misstatements with direct implications for future cash flows. In the second category are misstatements that “do not have any bearing on the future cash flows of the firm or the discount rate that should apply to those cash flows . . . .”\(^{19}\) Collateral damage is placed in the second category.

This second description of collateral damage does not square with the statement that reassessments of management credibility or a company’s internal controls are examples of collateral damage. As discussed in detail below, management credibility or a company’s internal controls will have an impact—in some cases a large impact—on future cash flows. If they did not, reassessments of management credibility or internal controls would not affect the stock price.

To avoid this confusion, there is a more direct and useful way to define collateral damage: the valuation impact of a corrective disclosure that does not correspond to the original inflation. This definition is best illustrated by an example, and the one presented by Ferrell and Saha is instructive. They posit: “[A] misstatement might be an accounting statement by a firm that falsely states that the firm has $100 more in cash than it really does while falsely understating, in the same statement, the firm’s corporate holdings of U.S. treasury bonds by an equivalent amount, $100.”\(^{20}\) From a financial point of view, such a misstatement—at the time it occurred—would have no impact on value. Treasury bonds are sufficiently liquid that a company can move from bonds to cash or vice versa almost instantaneously at a negligible transaction cost. However, Ferrell and Saha go on to assume that the firm fails promptly to report the error. It is not until a later date that the firm issues a corrective disclosure regarding its holdings of cash and Treasury bonds. At that time, the stock price falls. Because the original misstatement could not have inflated the stock price in an efficient market, the decline following the corrective disclosure must be due to collateral damage.\(^{21}\)

The example raises at least two questions related to collateral damage. First, from an economic perspective, what is the source of the collateral damage? Second, from a legal perspective, should collateral damage be recoverable? That is, should it be taken into account when assessing materiality, reliance, causation, and the other elements of liability? Or should it be considered as independent of the alleged misstatements? Part IV addresses these legal questions, but first the Article focuses on the economic issues.

\(^{17}\) Id.
\(^{18}\) Id.
\(^{19}\) Id. at 180.
\(^{20}\) Id.
\(^{21}\) Id. Of course, this example is a bit unrealistic from an economic standpoint. Presumably, companies engage in securities fraud only when management perceives that perpetrating a fraud will create some benefit. With respect to misstating holdings of cash and Treasury bonds, there is no apparent benefit to be gained.
Before turning to the analysis, it is helpful to graphically summarize the concept of collateral damage. The impact of collateral damage in a hypothetical case is depicted in Figure 1.

**Figure 1: Inflation and Collateral Damage**

As shown in the figure, price and value are equal until the company receives negative information (for instance, that its earnings will be less than expected). At that juncture, rather than accurately disclosing the information, the company misstates its financials so that earnings appear to meet market expectations. As a result, the market price of the stock does not fall as indicated by the dotted line, but its true value does as indicated by the solid line. The figure shows that if the correct information had been disclosed, the price would have fallen by $5 per share. Consequently, the stock price is now inflated because true value has fallen but the price remains constant. The inflation equals the difference between the solid line and the dotted line. At a later date, the company corrects the omission by providing accurate historical financial information, and the stock price falls by $15 per share. As shown, the drop exceeds the original inflation by $10 per share. As a result, the ultimate stock price is now $10 per share below what it would have been in the but-for world in which there was no misstatement. The added drop represents the collateral damage. The economic question is, what is the source of this additional decline?
Value creation does not occur in a vacuum. It is affected by the relationships between the various classes of stakeholders, including managers, employees, customers, distributors, and investors, who are party to the nexus of both legal contracts and informal agreements that define the corporation. Both the contractual and informal relationships among stakeholders depend upon the flow of information among counterparties. Disclosure of negative information can damage these working relationships and thereby affect the ability of the firm to generate cash flow. We argue that the primary source of collateral damage comes from the impact that corrective disclosures have on stakeholder relationships. In this context, collateral damage is not a secondary consideration from a valuation standpoint. In many situations, collateral damage can have a larger impact on a company’s stock price than the original misstatement or omission with which it is associated. Nonetheless, as we argue below, this does not mean it results in recoverable damage.

The stakeholder approach builds off the large literature in economics regarding reputational capital. The idea is that relationships between various corporate stakeholders, including managers, investors, customers, and employees, are multifaceted and complex. While some aspects of those relationships can be reduced to legal contracts, many cannot. In this context, managers and companies create what can be called reputational value by properly managing the intangible aspects of those relationships. One widely studied example in this regard is the development of brand-name capital. Consumers pay higher prices for brand-name products because the brand name provides information about product quality at the time of purchase. Companies with brand names, in turn, take extra care to assure consumers that they continue to produce high-quality products so as to maintain the reputation associated with the brand and, thereby, continue to charge premium prices.

As shown by Klein and Leffler, Cornell and Shapiro, and several subsequent authors, such reputational capital often constitutes a significant fraction

---

22 See Bradford Cornell & Alan C. Shapiro, Corporate Stakeholders and Corporate Finance, 16 FIN. MGMT. 5, 10 (1987).
23 See, e.g., W. Bentley MacLeod, Reputations, Relationships, and Contract Enforcement, 45 J. ECON. LITERATURE 595, 595–96 (2007) (exploring “the relationship between the concern for one’s reputation and the effect of reputation on the efficiency and form of contracts that are enforced informally”).
24 Id. at 597 (“If one can assign a value to one’s reputation, one can reformulate the problem in which each party trades off the benefit of breaching against the cost of losing one’s reputation.”).
25 Id. at 611–16.
26 Id.
of a company’s market value. In addition to brands, reputational value can be associated with perceptions regarding the honesty, integrity, and skill of management, the internal controls of the company, and the company’s quality control, among other things.29

Because reputational value depends on the perceptions of investors and other stakeholders, it is highly sensitive to negative information regarding the company, its management, or its products. A variety of empirical studies demonstrates the impact of disclosures on the value of reputational capital. One source of evidence is the market response to corporate litigation. The Texaco-Pennzoil case provides a particularly dramatic example.30 On November 19, 1985, a Houston jury found that Getty Oil breached its contract with Pennzoil for Pennzoil to buy Getty, and because Texaco had indemnified Getty for any resulting litigation, the jury directed Texaco to pay Pennzoil $10.53 billion in damages plus interest.31 In the six trading days following the verdict, the market value of Texaco dropped by approximately $1.8 billion, while the market value of Pennzoil rose by only about $800 million.32 After the judge upheld the award, Pennzoil stock rose further, and Texaco fell. Overall, Texaco lost about $2.58 billion in market value, whereas Pennzoil’s market value rose by about $935 million.33 One explanation for the $1.64 billion differential between Texaco’s loss and Pennzoil’s gain is the impact of the trial on reputational capital. Given the financial distress caused by the award, customers, suppliers, and business partners were no longer willing to do business with Texaco on the same terms.34 For instance, Atlantic Richfield sent a letter to its staff in early December urging them to use “prudence” in doing business with Texaco.35 This loss of confidence increased the cost of doing business for Texaco, thereby reducing the value of the company.36 On the other hand, receiving a windfall would have been unlikely to significantly affect Pennzoil’s reputational capital. If stakeholders were already satisfied with their relationship with Pennzoil, changes associated with the windfall would have been small. Though it is a dramatic illustration, the Texaco-Pennzoil example is not unique. Large corporate awards are typically associated with asymmetric valuation changes. Cornell and

have long considered ‘reputations’ and brand names to be private devices which provide incentives that assure contract performance in the absence of any third-party enforcer.”). 28 See Cornell & Shapiro, supra note 22, at 8–9 (describing the effect that failure to deliver on an implicit claim could have had on IBM’s company name).

29 Id. (detailing IBM’s reaction to setbacks with a product).


31 Id. at 158.

32 Id. at 161.

33 Id.

34 Id. at 167–68.


36 Id.
Engelman\textsuperscript{37} study five cases in detail and report results similar to those for Texaco and Pennzoil.

The reputational effect also explains the response of stock prices to recalls by drug and automobile manufacturers. In a classic study, Jarrell and Peltzman report that the drop in shareholder wealth accompanying recalls is twelve times the size of the direct costs of the recalls on average.\textsuperscript{38} These findings are consistent with the view that the reputational capital associated with product quality control is an important component of the value of a drug company.

The impact of defective Firestone tires on Ford Explorer sport-utility vehicles in 2000 provides another example. Bridgestone’s stock price dropped by more than $9 billion over the four months after Bridgestone (Firestone’s Japan-based parent company) announced the recall of the defective tires.\textsuperscript{39} Overall, the drop reached nearly 55 percent of Bridgestone’s stock market value.\textsuperscript{40} Ford’s stock price did not drop initially, but it eventually fell about 18 percent relative to the S&P 500 index over the same period, as information was revealed that Ford knew tire failure was possible.\textsuperscript{41} These stock market declines amounted to losses of about $7 billion in Bridgestone’s market value and nearly $10 billion in Ford’s market value.\textsuperscript{42} These costs were substantially greater than the direct costs associated with the recall and the related litigation, as estimated by the companies.\textsuperscript{43}

Finally, a comprehensive study by Karpoff, Lee, and Martin\textsuperscript{44} examines the impact of collateral damage in securities-related matters directly. The authors analyze 384 firms investigated by the Securities and Exchange Commission (“SEC”) for accounting-related issues from 1978 through 2002, for which Compustat financial data is available. The authors divide the total losses in value at the sample firms into four categories: (1) fines and sanctions imposed by regulators, (2) amounts paid to settle related securities class actions, (3) the valuation impact of financial restatements, and (4) remaining reputational costs.\textsuperscript{45} The authors find that the fourth component, reputational costs, accounted for more than 66 percent of the decline in value.\textsuperscript{46} Furthermore, if reputational costs are compared with component (3) alone—the component associated with the original inflation—they are more than two-and-a-half times as large on average.\textsuperscript{47}

\textsuperscript{37} See id. at 398.
\textsuperscript{40} Id.
\textsuperscript{41} Id.
\textsuperscript{42} Id.
\textsuperscript{43} Id.
\textsuperscript{44} Jonathan M. Karpoff, D. Scott Lee & Gerald S. Martin, The Cost to Firms of Cooking the Books, 43 J. FIN. & QUANTITATIVE ANALYSIS 581 (2008).
\textsuperscript{45} Id. at 582–83.
\textsuperscript{46} Id. at 582.
\textsuperscript{47} Id. at 583.
The underlying point is that collateral damage is not a secondary consideration from a valuation standpoint. Loss in reputational value associated with the company and its management often greatly exceeds the amount of the original inflation. As a result, great care must be taken to differentiate between the extent to which a disclosure corrects an original misstatement and thereby removes inflation, and the collateral damage caused by the disclosure.

With this background, we return to the example of the company that hides an earnings miss by accounting manipulation and thereby inflates its stock price by $5 per share. When the accounting misstatement is disclosed, the stock price drops by $15 per share—$5 per share to correct for the accounting misstatement and an additional $10 per share associated with reputational collateral damage. In this example, the reputational effects are most likely associated with the honesty, integrity, and competence of management, as well as the reliability of the company’s internal controls. When the corrective disclosure occurs, the market acquires two—and only two—pieces of information it did not possess beforehand: (1) the substantive information it would have had if accurate information had been conveyed originally, and (2) the knowledge that the company made a false statement. Disclosure of the first category corrects the inflation.48 Disclosure of the second leads to collateral damage by causing the market to reevaluate the reputation of the company’s management and its internal controls.49

In summary, a misrepresentation can cause investors to suffer real financial harm—at least in a “but for” sense—in the form of collateral damage as a result of

---

48 Removing the inflation may itself involve the devaluation of reputational capital as the market considers whether an earnings shortfall bears on the quality of the company’s management, but only reputational capital that would have been devalued anyway if accurate information had been disclosed. The devaluation of this reputational capital does not represent collateral damage.

49 The market price also can decline in an amount greater than the original inflation if the market expects that the company will face securities fraud lawsuits or regulatory sanctions as a result of its misstatement. Richard Booth demonstrates how this anticipation of a lawsuit theoretically can cause the stock price to drop by a multiple of the percentage decline attributable to the news itself. See Richard A. Booth, Who Should Recover What in a Securities Fraud Class Action? 4, 7 (U. of Maryland Sch. Law Faculty Publications, Working Paper No. 11, 2005), available at http://digitalcommons.law.umd.edu/fac_pubs/11/. While Booth is theoretically correct, the exacerbation of the decline is likely limited by the fact that not all misstatements associated with stock price declines lead to litigation or regulatory sanctions. Also, many lawsuits are dismissed, and among those that proceed, settlements typically are only small fractions of claimed damages. For example, recovery rates in securities class actions were only 2.7 percent of investor losses in 2002 and 2.8 percent of investor losses in 2003. Elaine Buckberg et al., Recent Trends in Securities Class Action Litigation: 2003 Update, 5 CLASS ACTION LITIG. REP. 304, 307 (2004). Because investors are presumptively aware of these facts, the amount by which the initial stock price decline is exacerbated should not be a multiple as Booth suggests, although it is not a factor that can be ignored. Along with the economic issues discussed above, it will widen the potential wedge between the inflation caused by the misstatement and the amount by which the value of a company declines upon a corrective disclosure.
altered assessments of the reputational capital of the company and its management. Furthermore, the changes in value caused by these reputational effects often significantly exceed the inflation caused by the initial misstatement or omission. However, the size of the reputational effect does not answer the critical legal question: to what extent is recovery for collateral damage permitted by section 10(b) of the Securities Exchange Act of 1934 (the “Exchange Act”)\(^5\) and SEC Rule 10b-5?\(^5\)

IV. DOES SECTION 10(B) PERMIT RECOVERY FOR COLLATERAL DAMAGE?

There is good reason to tread cautiously in permitting recovery for collateral damage. Collateral damage is inherently nebulous and therefore difficult to identify and to measure. What may appear at first blush to be collateral damage may be attributable, in whole or in part, to market inefficiencies and random crashes in security prices bearing no relation to changes in the value of a company’s reputational capital.

An example illustrates the point. On September 21, 2000, Intel Corp. issued a brief and seemingly unexceptional press release.\(^5\) It projected that revenue for the third quarter would be 3 to 5 percent higher than the second quarter’s revenue of $8.3 billion—results that fell short of the company’s previous forecast of 7 to 9 percent growth.\(^5\) Intel attributed the lower growth to a temporary decline in European sales associated with currency exchange rate fluctuations. An Intel officer later expressed his view that the press release was relatively insignificant in that it reflected only short-run developments in Europe (the decline of the Euro) and did not reflect “any change in Intel’s long-run strategy, product mix, competitive position, or even . . . the long-run demand for Intel’s products.”\(^5\) Intel’s stock price could have been expected to decline slightly on this news, but instead it plummeted. Its stock fell nearly 30 percent over the next two trading days, and $122 billion in shareholder wealth—more than twice the market capitalization of Enron Corp. at its peak—evaporated.\(^5\)

Bradford Cornell, a co-author of this Article, has compared the $122 billion drop in market value with estimates of the change in the company’s fundamental value associated with the announcement.\(^5\) Cornell used analysts’ forecasts of financial performance to compute the discounted present value of future cash flows for Intel before and after the September 21, 2000 announcement.\(^5\) He found that, depending on the assumptions used, the drop in the discounted cash flow value of...

\(^5\)\) See Bradford Cornell, Is the Response of Analysts to Information Consistent with Fundamental Valuation? The Case of Intel, 30 Fin. Mgmt. 113, 113 (2001).
\(^\)\) See id.
\(^\)\) See id.
\(^\)\) See id. at 119–28.
\(^\)\) See id. at 123, 127.
the company accounted for between 1 percent and 4.5 percent of the market drop. The remaining $116.5 billion to $120.8 billion in lost market capitalization must have been attributable to something else.  

Plaintiffs often see a bonanza in facts like these. They invariably invoke efficient market theory in presenting their damages models, following a devastatingly simple chain of reasoning to the ultimate conclusion that the entire stock price drop (net of the market and the industry) is attributable to the fraud. Plaintiffs argue that if “[t]he market is acting as the unpaid agent of the investor, informing him that given all the information available to it, the value of the stock is worth the market price,” then the entire stock price drop after the corrective disclosure must be attributable to the fraud. It has been observed that the stock price drop in many cases is disproportionate to the fundamental value of the information conveyed by the corrective disclosure, and, therefore, the entire drop cannot represent the amount the shareholder overpaid for the stock—i.e., it cannot represent out-of-pocket damages. Plaintiffs have developed a convenient response to this observation. They portray the delta as collateral damage arising from changed perceptions regarding the value of the reputational capital associated with the company and its management. They argue that this is a foreseeable consequence of securities fraud that is therefore recoverable not as out-of-pocket damages, but as consequential damages. After all, they argue, if the market is efficient, what other explanation can there be?  

As we demonstrated in a previous article, however, there can be many explanations for a stock price drop that is out of all proportion to the fundamental value of the information conveyed by a corrective disclosure. As we show in that article, even the most efficient of markets contains inefficiencies that can cause unwarranted crashes in stock prices irrespective of changed perceptions regarding the value of reputational capital. Certainly in the case of Intel, it is difficult to imagine how a disclosure relating to short-term fluctuations in currency exchange rates could change perceptions of management to the tune of $120 billion—yet determining whether any portion of the stock price drop is attributable to such changed perceptions is a hazardous enterprise at best. Estimating how much of the drop might represent collateral damage is wholly speculative.

58 See id. at 126.
60 See, e.g., Programs v. Leighton, 90 F.3d 1442, 1449 (9th Cir. 1996) (“Rule 10b-5 plaintiffs may recover consequential damages which can be proven with reasonable certainty to have resulted from the fraud.”) (internal quotation marks, emphasis and citation omitted); Garnatz v. Stifel, Nicolaus & Co., 559 F.2d 1357, 1360 (8th Cir. 1977) (“As applied to a fraudulently induced purchase of securities, [the out-of-pocket] rule provides for the recovery of the difference between the actual value of the securities and their purchase price. . . . Recovery is also allowed for any consequential damages proximately resulting from the fraud.”).
61 Cornell & Rutten, supra note 3, at 464.
The problem is compounded by ex post selection bias on the part of plaintiffs. Even the largest and most sophisticated plaintiffs’ firms have finite resources and must pick and choose their cases based not only on the strengths and weaknesses of a case, but also on its potential for recovery. A plaintiff’s lawyer naturally will prefer to take on cases with large drops in market capitalization, all else being equal. It is difficult to differentiate the portion of a stock price drop caused by collateral damage from a portion attributable to other factors, such as market inefficiencies. Thus, permitting plaintiffs to recover for collateral damage threatens to permit recovery completely unrepresentative of collateral damage—recovery that normally would not be permitted under any legal theory.

The federal securities laws are carefully structured to avoid such results. The securities laws are designed to strike a balance between deterring and remedying wrongdoing, and imposing liabilities so draconian that they effectively deter capital formation. The securities laws therefore carefully circumscribe recoverable damages, often more sharply than the common law would. Section 28 of the Exchange Act, for example, precludes punitive or exemplary damages by providing that “no person permitted to maintain a suit for damages under the provisions of this Act shall recover . . . a total amount in excess of his actual damages on account of the act complained of.” Additionally, section 21D of the Exchange Act, enacted as part of the Private Securities Litigation Reform Act of 1995, imposes a “look back” requirement because Congress was concerned that “the current method of calculating damages in 1934 Act securities fraud cases” was too “uncertain” and resulted in “windfall damages.”

This is not to suggest that collateral damage, where it exists, does not stem from securities fraud—at least in a “but for” sense, as discussed below—or that it does not constitute actual financial harm to shareholders. Nor is it meant to suggest that difficulties in identifying and measuring collateral damage are sufficient reasons to preclude recovery, although prohibitions on recovering damages that

---

62 See, e.g., H.R. Rep. No. 104-369, at 31 (1995) (Conf. Rep.) (“The overriding purpose of our Nation’s securities laws is to protect investors and to maintain confidence in the securities markets, so that our national savings, capital formation and investment may grow for the benefit of all Americans.”); Stoneridge Inv. Partners, LLC v. Scientific-Atlanta, Inc., 552 U.S. 148, 128 S. Ct. 761, 772 (2008) (declining to adopt an expansive interpretation of section 10(b) in part because companies “might find it necessary to protect against these threats, raising the costs of doing business,” and “[o]verseas firms with no other exposure to our securities laws could be deterred from doing business here,” which, “in turn, may raise the cost of being a publicly traded company under our law and shift securities offerings away from domestic capital markets”); United Hous. Found., Inc. v. Forman, 421 U.S. 837, 849 (1975) (“The primary purpose of the Acts of 1933 and 1934 was to eliminate serious abuses in a largely unregulated securities market. The focus of the Acts is on the capital market of the enterprise system: the sale of securities to raise capital for profit-making purposes, the exchanges on which securities are traded, and the need for regulation to prevent fraud and to protect the interest of investors.”).


64 Id. § 78u-4(e).

cannot reasonably be estimated might well be dispositive in most, if not all, cases. The point is that the _in terrorem_ effects of permitting recovery for collateral damage cannot be ignored in evaluating whether section 10(b) should be construed to permit such recovery.

Below, we apply the United States Supreme Court’s governing analytical framework for construing the scope of implied rights of action in the securities laws to show that recovery for collateral damage is not permitted by section 10(b), precisely because of its _in terrorem_ effects. We also show that permitting recovery for collateral damage would be contrary to nearly every element of the section 10(b) cause of action.

**A. The Supreme Court’s Analytical Framework for Construing the Scope of the Implied Right of Action Under Section 10(b)**

Section 10(b)’s language is the starting point in construing the scope of its right of action. Section 10(b) does not expressly address the damages a private plaintiff may recover, which is unsurprising because the statute does not expressly create a private right of action at all—that right is judicially implied. The statute nevertheless contains language relevant to whether collateral damage is recoverable.

Section 10(b) prohibits fraud only “in connection with the purchase or sale of any security.” The statute’s “purchase or sale” requirement generally is regarded as the

---

66 See, e.g., Ambassador Hotel Co. v. We-Chuan Inv., 189 F.3d 1017, 1030 (9th Cir. 1999) (“Consequential damages may also be awarded [under section 10(b)] if proved with sufficient certainty.”) (citation omitted); Pelletier v. Stuart-James Co., 863 F.2d 1550, 1557–58 (11th Cir. 1989) (“The measure of damages in a Rule 10b-5 case is limited to actual pecuniary loss suffered by the defrauded party, and does not include any speculative loss of profits.”); Hershock v. Fiascki, No. 90-0497, 1992 WL 164739, at *7 (E.D. Pa. July 2, 1992) (“The important element in allowing a recovery for other than out-of-pocket damages is whether the asserted damage is actual and nonspeculative.”).

67 See, e.g., Cent. Bank of Denver, N.A. v. First Interstate Bank of Denver, N.A., 511 U.S. 164, 173 (1994) (“With respect [to] . . . the scope of conduct prohibited by § 10(b), the text of the statute controls our decision. . . . [A] private plaintiff may not bring a 10b-5 suit against a defendant for acts not prohibited by the text of § 10(b). To the contrary, our cases considering the scope of conduct prohibited by § 10(b) in private suits have emphasized adherence to the statutory language . . . .”).

68 See id. (“[Determining] the elements of the 10b-5 private liability scheme[] has posed difficulty because Congress did not create a private § 10(b) cause of action and had no occasion to provide guidance about the elements of a private liability scheme.”); Superintendent of Ins. v. Bankers Life & Cas. Co., 404 U.S. 6, 13 n.9 (1971) (endorsing the consensus of the lower courts and commentators that there is an implied right of action under section 10(b) and Rule 10b-5).

as a standing requirement,\textsuperscript{70} but is also relevant to the damages question because it indicates that Congress was focused on the purchase transaction.\textsuperscript{71} The statutory language indicates that Congress was focused on the out-of-pocket harm befalling investors who purchase at an inflated price, only to later suffer losses when the value of their holdings declines in a corresponding amount upon a corrective disclosure\textsuperscript{72}—i.e., their out-of-pocket damages. There is no indication that Congress was concerned with losses investors suffer from additional declines in value due to changes in reputational capital not corresponding with an amount by which the investors overpaid.\textsuperscript{73}

The legislative history of section 10(b) is consistent with its language. It suggests that section 10(b) was designed to ensure that the prices investors pay reflect only the legitimate forces of supply and demand.\textsuperscript{74}

Although section 10(b)’s text and legislative history suggest that the statute does not permit recovery for collateral damage, they are anything but definitive. Determining whether section 10(b) permits such recovery thus requires additional analysis.

The Supreme Court repeatedly has held that when the scope of the section 10(b) right of action cannot be gleaned from the statutory language, courts should refrain from determining its scope based on policy considerations or the

\textsuperscript{70} See Blue Chip Stamps v. Manor Drug Stores, 421 U.S. 723, 731–33 (1975) (holding that persons who neither purchased nor sold securities lack standing to sue under section 10(b)).

\textsuperscript{71} For simplicity, we discuss the issues in terms of defrauded purchasers, not defrauded sellers.

\textsuperscript{72} As \textit{Dura} explains, even if an investor pays an inflated price for stock, the investor suffers no harm unless and until the market price subsequently declines as a result of a corrective disclosure—because at “the moment the transaction takes place, . . . the inflated purchase payment is offset by ownership of a share that at that instant possesses equivalent value.” \textit{Dura Pharm., Inc. v. Broudo}, 544 U.S. 336, 342, (2005). \textit{Dura} teaches that there must be a corrective disclosure before there can be a loss and that a plaintiff’s out-of-pocket damages cannot exceed the amount by which the stock price subsequently declines.

\textsuperscript{73} Cf. Strougo \textit{ex rel. Brazil Fund, Inc. v. Scudder, Stevens & Clark, Inc.}, 964 F. Supp. 783, 803 (S.D.N.Y. 1997) (“\textit{Blue Chip Stamps} and the other cases cited by the defendants in support of an out-of-pocket rule involved securities fraud claims . . . . The defendants cite no authority imposing a similar ‘purchaser-seller’ requirement in cases arising under [another statute]. Moreover, \textit{Blue Chip Stamps} was decided under Section 10(b) and Rule 10b-5, which expressly require that the plaintiff be injured ‘in connection with the purchase or sale of any security.’”).

\textsuperscript{74} For example, one of the Senate reports states: “Certain devices employed for the purpose of artificially raising or depressing security prices are specifically prohibited by the act. Others have not been forbidden outright but have been placed under the control of the Securities and Exchange Commission.” S. REP. NO.1455, at 54 (1934) (emphasis added). The reference to “other[]” provisions appears to be a reference to section 10(b), which prohibits fraud in connection with the purchase or sale of securities “in contravention of such rules and regulations as the Commission may prescribe.” 15 U.S.C. § 78j(b) (2006).
like, as courts frequently do with common law causes of action. Rather, the Court has held that because the cause of action is statutory, courts should attempt to infer how the 1934 Congress would have defined the parameters of the section 10(b) cause of action had it enacted it expressly—looking for guidance in the causes of action Congress actually did enact expressly.

In *Musick, Peeler & Garrett v. Employers Insurance of Wausau*,75 for example, the Court considered “whether a right to contribution is within the contours of the 10b-5 action.”76 The Court criticized the parties for “devot[ing] considerable portions of their briefs to debating whether a rule of contribution or of no contribution is more efficient or more equitable,” and “declined to rule on such matters.”77 The Court explained:

> Our task is not to assess the relative merits of the competing rules, but rather to attempt to infer how the 1934 Congress would have addressed the issue had the 10b-5 action been included as an express provision in the 1934 Act. We do this not as an exercise in historical reconstruction for its own sake, but to ensure that the rules established to govern the 10b-5 action are symmetrical and consistent with the overall structure of the 1934 Act and, in particular, with those portions of the 1934 Act most analogous to the private 10b-5 right of action that is of judicial creation. . . . [O]ur goals in establishing limits for the 10b-5 action [are] . . . to ensure the action does not conflict with Congress’ own express rights of action, to promote clarity, consistency, and coherence for those who rely upon, or are subject to, 10b-5 liability, and to effect Congress’ objectives in enacting the securities laws.78

Similarly, in *Central Bank of Denver, N.A. v. First Interstate Bank of Denver, N.A.*,79 the Court considered whether defendants can be liable for aiding and abetting violations of section 10(b). The Court applied the same approach:

> When the text of § 10(b) does not resolve a particular issue, we attempt to infer how the 1934 Congress would have addressed the issue had the 10b-5 action been included as an express provision in the 1934 Act. For that inquiry, we use the express causes of action in the securities Acts as the primary model for the § 10(b) action.80

---

76 Id. at 294.
77 Id.
78 Id. at 294–95 (citations omitted).
80 Id. at 178 (internal quotation marks and citation omitted).
The Court explained: “The reason is evident: Had the 73d Congress enacted a private § 10(b) right of action, it likely would have designed it in a manner similar to the other private rights of action in the securities Acts.”

The Court repeatedly has applied this approach in numerous cases presenting questions about the scope of section 10(b), such as whether only purchasers or sellers of securities have standing to sue, whether there is a scienter requirement, whether there is a reliance requirement, the appropriate statute of limitations, and where to draw the line between primary and secondary liability.

The most relevant express causes of action to consult in construing the scope of section 10(b) are contained in sections 9 and 18 of the Exchange Act, because they are “close in structure, purpose, and intent to the 10b-5 action.” They “both target the precise dangers that are the focus of section 10(b), and the intent motivating all three sections is the same—to deter fraud and manipulative

---

81 Id.
82 See Blue Chip Stamps v. Manor Drug Stores, 421 U.S. 723, 735–36 (1975) (“The principal express nonderivative private civil remedies, created by Congress contemporaneously with the passage of § 10(b), for violations of various provisions of the 1933 and 1934 Acts are by their terms expressly limited to purchasers or sellers of securities. . . . It would . . . be anomalous to impute to Congress an intention to expand the plaintiff class for a judicially implied cause of action beyond the bounds it delineated for comparable express causes of action.”).
83 See Ernst & Ernst v. Hochfelder, 425 U.S. 185, 200 (1976) (“It is . . . evident that Congress fashioned standards of fault in the express civil remedies in the 1933 and 1934 Acts on a particularized basis. Ascertainment of congressional intent with respect to the standard of liability created by a particular section of the Acts must therefore rest primarily on the language of that section.”).
85 See Lampf v. Gilbertson, 501 U.S. 350, 359 (1991) (“[W]e are faced with the awkward task of discerning the limitations period that Congress intended courts to apply to a cause of action it really never knew existed. . . . We can imagine no clearer indication of how Congress would have balanced the policy considerations implicit in any limitations provision than the balance struck by the same Congress in limiting similar and related protections. When the statute of origin contains comparable express remedial provisions, the inquiry usually should be at an end.”) (citations omitted). The Supreme Court decided Lampf before Congress amended 28 U.S.C. § 1658 to provide an express statute of limitations for securities fraud claims.
86 See Stoneridge Inv. Partners LLC v. Scientific-Atlanta, Inc., 552 U.S. 148, 128 S. Ct. 761, 768 (2008) (“The Court doubted the implied § 10(b) action should extend to aiders and abettors when none of the express causes of action in the securities Acts included that liability.”).
88 Id. § 78r.
practices in the securities markets, and to ensure full disclosure of information material to investment decisions.”90 Section 9, like Rule 10b-5(a) and (c), generally prohibits manipulative securities transactions, such as wash sales and the like. Section 18, similar to Rule 10b-5(b), prohibits “false or misleading” statements, albeit limited to the context of documents filed with the SEC.91

Both statutes expressly discuss the damages a private plaintiff may recover. Section 9 provides that “[a]ny person who willfully participates in any act or transaction in violation of [the statute] shall be liable to any person who shall purchase or sell any security at a price which was affected by such act or transaction, and the person so injured may sue . . . to recover the damages sustained as a result of any such act or transaction.”92 Similarly, section 18 provides that “[a]ny person [who makes a false statement in violation of the statute] shall be liable to any person . . . who, in reliance upon such statement, shall have purchased or sold a security at a price which was affected by such statement, for damages caused by such reliance.”93 Sections 9 and 18 are similar to section 10(b) in that they both contain a “purchase or sale” requirement. Moreover, sections 9 and 18 explicitly reference “a price which was affected by” the fraud,94 thereby strongly suggesting that Congress was focused on the harm that can befall investors when they purchase shares at an inflated price and later suffer losses from a corresponding decline in value—i.e., their out-of-pocket damages. Section 9, by providing that only an investor “so injured” may sue, further underscores that Congress was focused on investors’ out-of-pocket harm. The legislative history of sections 9 and 18 is in accord.95 Not surprisingly, the statutes have been construed, like section 10(b), to permit recovery of out-of-pocket damages.96

90 Id. (internal quotation marks and citations omitted).
92 Id. § 78i(e) (emphasis added).
93 Id. § 78r(a) (emphasis added).
94 Id. §§ 78i(e), 78r(a).
95 S. REP. NO. 792, at 7 (1934) (“Several devices are employed for the purpose of artificially raising or depressing security prices.”) (emphasis added); id. at 8 (stating that the anti-manipulation provisions of the bill were designed to prohibit “transactions specifically designed to manipulate the price of a security”) (emphasis added); id. at 12–13 (“[T]he bill provides that any person who unlawfully manipulates the price of a security . . . shall be liable in damages to those who have bought or sold the security at prices affected by such violation or statement.”) (emphasis added); id. at 17 (“To manipulate the price of a security . . . is prohibited by [section 9].”) (emphasis added); id. at 21 (section 18 creates liability in any person “who in reliance on the statement . . . has purchased the security to which it relates at a price affected by it”) (emphasis added); S. REP. NO. 1455, at 30 (1934) (“The true function of an exchange is to maintain an open market for securities, where supply and demand may freely meet at prices uninfluenced by manipulation.”); id. at 54 (“Certain devices employed for the purpose of artificially raising or depressing security prices are specifically prohibited by the act.”); H.R. REP. No. 1383, at 10 (1934) (“The bill seeks to give to investors markets where prices may be established by the free and honest balancing of investment demand with investment supply.”) (emphasis added); id. (“False
The other express causes of action in the Exchange Act, though less relevant because they are less analogous to section 10(b), underscore that Congress was careful in defining recoverable damages so as to avoid creating liability that is unduly open-ended and unpredictable. Section 16 relates to short swing trading and permits recovery of only the “profit realized by [the trader].” Section 20(a) imposes secondary liability on “controlling persons” only “to the same extent [that] a controlled person” is liable under one of the substantive sections of the Exchange Act. Section 20A relates to insider trading, and although it is arguably irrelevant here because it was enacted in 1988 instead of by the 1934 Congress, it too is careful to cap recoverable damages at a definite and certain amount; it provides that “[t]he total amount of damages [recoverable by persons who traded in the market at the same time as the insider] shall not exceed the profit gained or loss avoided in the transaction or transactions.”

The Securities Act of 1933 (the “Securities Act”) has a somewhat different focus but does not lead to a different conclusion. The Securities Act contains two express causes of action (other than a cause of action against “controlling persons” that is identical insofar as relevant here to section 20(a) of the Exchange Act). Section 11 of the Securities Act relates to false and misleading registration statements and prescribes a damages measure of “the difference between the amount paid for the security (not exceeding the price at which the security was offered to the public)” and the market price at certain subsequent points in time, subject to a “negative causation” defense that “any portion or all of such damages represents other than the depreciation in value of such security resulting from [the] misleading statements designed to induce investors to buy when they should sell and sell when they should buy are . . . outlawed and penalized.” (emphasis added).


99 Id. § 78t(a).
100 Id.
101 Id. § 78t-1.
102 See Musick v. Employers Ins. of Wausau, 508 U.S. 286, 296 (1993) (declining to consider section 20A for this reason).
104 Id. § 77o.
105 Id. § 77k.
registration statement . . . not being true.” 106 Section 12107 of the Securities Act relates to sales of unregistered securities or pursuant to false or misleading prospectuses or oral communications, and it provides for rescission of the transaction or a rescissory measure of damages, 108 again subject to a negative causation defense. 109 Sections 11 and 12, rather than referencing the amount the plaintiff overpaid, explicitly look to the amount by which the plaintiff’s investment declined—an amount that could include collateral damage—while taking care to ensure that the plaintiff cannot recover for damages not proximately caused by the misstatement.

Sections 11 and 12, however, are vastly different from section 10(b). Section 11 relates to sales pursuant to false or misleading registration statements, a context in which the sale price is set by the issuer, not the open market—and in the case of initial public offerings, before there is any secondary market for the securities at all. Accordingly, section 11—unlike section 10(b)—by definition cannot look to market-price inflation in calculating the plaintiff’s damages. Although section 11 does permit recovery in connection with open-market purchases, it requires an open-market purchaser plaintiff to “trace” its shares to the challenged registration statement,110 which, in light of the way modern-day securities trades are cleared,111

106 Id. § 77k(e).
107 Id. § 77l.
108 See id. § 77l(a) (“Any person who [violates the statute] shall be liable . . . to the person purchasing such security from him, who may sue . . . to recover the consideration paid for such security with interest thereon, less the amount of any income received thereon, upon the tender of such security, or for damages if he no longer owns the security.”); Randall v. Loftsgaarden, 478 U.S. 647, 649 (1986) (stating that investors are “entitled under § 12(2) . . . to rescind the fraudulent transaction or obtain rescissory damages”); Farley v. Henson, 11 F.3d 827, 837 (8th Cir. 1993) (addressing rescissory damages under section 12).
110 See, e.g., Lee v. Ernst & Young LLP, 294 F.3d 969, 977 (8th Cir. 2002) (ruling that “§ 11 is broad enough to encompass some aftermarket purchases, subject, of course, to the long-recognized requirement” that the security was indeed issued under that registration statement and not another); Krin v. pOrder.com, Inc., 210 F.R.D. 581, 586 (W.D. Tex. 2002) (noting that even where 91 percent of the stock in the market came from the challenged registration statement, the plaintiffs could not prevail under section 11 because “[p]laintiffs must demonstrate all stock for which they claim damages was actually issued pursuant to a defective statement, not just that it might have been, probably was, or most likely was, issued pursuant to a defective statement”) (emphasis removed); In re Quarterdeck Office Sys., Inc. Sec. Litig., No. CV 92-3970-DWW(GHKx), 1993 WL 623310, at *2–3 (C.D. Cal. Sep. 30, 1993) (holding that plaintiff lacked section 11 standing even though fully 97 percent of the shares outstanding were issued pursuant to the allegedly defective registration statement); Lorber v. Beebe, 407 F. Supp. 279, 287 (S.D.N.Y. 1975) (“[I]t is insufficient that stock ‘might’ have been issued pursuant to a defective registration statement. A plaintiff must show that it actually was so issued.”).
111 Most broker-dealers clear trades through the Depository Trust Company (“DTC”)
is usually impossible if the same class of securities has been offered pursuant to more than one registration statement. 112 Further, section 11 provides that even when a plaintiff can trace (e.g., where there is only one registration statement), the plaintiff’s purchase price shall be deemed to be not more than “the price at which the security was offered to the public.” 113 Section 11 is therefore narrowly focused on registration statements, provides far more limited remedies, and, unlike section 10(b), provides no remedy for purchase-price inflation as such. Section 11 provides little support for construing section 10(b) to permit recovery for subsequent declines in the price of a security that do not correspond with inflation in the market price at the time of purchase.

Section 12 is similar. Section 12 creates liability only on the part of persons who “offer[] or sell[]” securities in certain prohibited circumstances, and provides a remedy only to “the person purchasing such security from him.” 114 Section 12 thus imposes what is effectively a privity requirement, authorizing suit against only the seller (and as construed by the case law, certain persons acting on behalf of or in conjunction with the seller). 115 Section 12 therefore tends to be limited to direct seller-purchaser transactions at prices the parties specifically negotiate or that one side dictates (e.g., face-to-face transactions, direct purchases from the issuer in a public offering), rather than open-market transactions at prices set by the market. Section 12 is therefore much more limited than section 10(b) and provides little support for construing section 10(b) to permit recovery for subsequent declines in the price of a security that do not correspond with inflation in the market price at the time of purchase.

In sum, section 10(b) and its analogous provisions, sections 9 and 18, suggest that Congress focused on the out-of-pocket harm that can befall investors when they overpay for stock and later suffer a loss due to a decline in the value of their


112 See Abbey v. Computer Memories, Inc., 634 F. Supp. 870, 872–73 (N.D. Cal. 1986) (holding that because the plaintiff’s shares “were purchased in the open market,” and at one time were “part of the common pool of . . . shares held in DTC’s vault,” the plaintiff could not trace his shares to the relevant registration statement); Kirkwood v. Taylor, 590 F. Supp. 1375, 1379 (D. Minn. 1984) (concluding that the way the DTC operates precludes tracing where there is more than one extant registration statement).


114 Id. § 77l.

holdings upon a corrective disclosure. This harm contrasts with losses investors suffer from further declines due to changed perceptions of the reputational value associated with a company and its management that do not correspond to purchase-price inflation. The other express causes of action further demonstrate that Congress was careful to craft remedies that were not utterly open-ended and unpredictable, as section 10(b) would threaten to be if recovery for collateral damage were permitted. The language of section 10(b) and the express causes of action all are consistent with the overall balance struck by the securities laws: protecting investors by deterring and remedying wrongdoing, while not imposing liabilities so burdensome that they deter capital formation. This balance counsels against permitting recovery for collateral damage under section 10(b).

B. The Elements of the Section 10(b) Cause of Action

A misleading statement or omission that causes collateral damage can be viewed as having two components: a substantive statement or omission (e.g., “We had third-quarter revenues of $100 million,” when actually they were $90 million); and an omission to disclose that management was dishonest, reckless, or incompetent in making the statement or omission, or that the information management relied on was the product of deficient internal controls. Permitting recovery for losses caused by the former is what section 10(b) entails if the other elements of liability are satisfied. As shown below, permitting recovery for losses caused by the latter would be contrary to nearly every element of the section 10(b) cause of action, including the existence of a duty to disclose, materiality, transaction causation, loss causation, and that the defendant’s statement or omission be made in connection with the purchase or sale of a security.

1. The Existence of a Duty to Disclose

Courts have made clear that section 10(b) prohibits only dishonest statements and conduct in connection with the purchase or sale of a security, not dishonest character, managerial incompetence, or deficient corporate controls. Courts consistently reject section 10(b) claims that allege nothing more than factors such as these, and they reject attempts by plaintiffs to bootstrap such issues into section 10(b) claims by alleging that defendants failed to disclose dishonesty, incompetence, or mismanagement.

Santa Fe Industries, Inc. v. Green,116 is the seminal case in this area. In Santa Fe, a 95 percent shareholder of a company tried to effect a short-form merger by cashing out the minority shareholders at a price of $150 per share—a $25 per share premium above the value set forth in an appraisal that the majority shareholder had commissioned and provided to the minority shareholders when it made the offer. The minority shareholders concluded that the value of their shares was significantly higher, but rather than availing themselves of their appraisal rights in

In 2009, Collateral Damage and Securities Litigation 739 state court, they sued under section 10(b). The minority shareholders alleged that the majority shareholder had commissioned a “fraudulent appraisal” and that providing the appraisal to them was an attempt to “lull [them] into erroneously believing that [the majority shareholder] was generous” when in fact it was trying to “freeze [them] out . . . at a wholly inadequate price.”

The Supreme Court held that section 10(b) did not extend to the challenged conduct, because the majority shareholder had fully disclosed the bases of the $150 offer, including the appraisal itself, and the plaintiffs’ remedy was to pursue their appraisal rights in state court—even though the plaintiffs clearly alleged that the majority shareholder harbored a dishonest motive in making the offer and transmitting the appraisal. The Court concluded that the complaint essentially alleged a breach of fiduciary duty, and the Court rejected the notion that “the term ‘fraud’ in Rule 10b-5 . . . bring[s] within the ambit of the Rule all breaches of fiduciary duty in connection with a securities transaction.” The Court explained:

[T]he claim of fraud and fiduciary breach in this complaint states a cause of action under any part of Rule 10b-5 only if the conduct alleged can be fairly viewed as “manipulative or deceptive” within the meaning of the statute. . . . [T]here was no “omission” or “misstatement” in the information statement accompanying the [$150 per share] offer. . . . [T]he cases do not support the proposition . . . that a breach of fiduciary duty by majority stockholders, without any deception, misrepresentation, or nondisclosure, violates the statute and the Rule. . . . [W]e do not think [Congress] . . . meant to bring within the scope of § 10(b) instances of corporate mismanagement such as this, in which the essence of the complaint is that shareholders were treated unfairly by a fiduciary. . . . We thus adhere to the position that Congress by § 10(b) did not seek to regulate transactions which constitute no more than internal corporate mismanagement.

The courts have consistently interpreted Santa Fe as standing for the proposition “[t]he securities laws do not require corporate management to direct conclusory accusations at itself or to characterize its behavior in a pejorative manner,” much less “to confess one’s corporate wrongdoing.” For example, in In re Citigroup, Inc. Securities Litigation, the plaintiff alleged that Citigroup “structur[ed] commodities and other financial transactions with Enron and other entities that permitted those entities to disguise debt on their books, with Citigroup carrying the transactions on its own books as equity investments rather than

117 Id. at 467.
118 Id. at 472.
119 Id. at 473–79.
extensions of credit.\textsuperscript{123} The plaintiff contended that such “activity supports a section 10(b) claim because the risks inherent in participating in allegedly illegal and deceptive conduct with Enron and other companies, as well as the effect disclosure of the transactions as loans would have had on Citigroup’s balance sheet, rendered misleading Citigroup’s public disclosures regarding its risk management policies.”\textsuperscript{124} The court, relying heavily on \textit{Santa Fe}, held that a failure to disclose corporate mismanagement or malfunctioning internal controls could not give rise to liability under section 10(b):

Even taking as true Plaintiff’s claims that Citigroup and its defendant officers knowingly participated in the structuring of transactions designed to mislead investors in Enron and other companies as to the true financial status of those companies, and that Citigroup improperly carried the transactions on its own books as equity investments rather than loans, Plaintiff’s section 10(b) claim here—that participation in such transactions was inconsistent with Citigroup’s stated risk management policies and historical business practices—amounts to nothing more than a charge that Citigroup’s business was mismanaged. Such allegations of mismanagement, even where a plaintiff claims that it would not have invested in an entity had it known of the management issues, are insufficient to support a securities fraud claim under section 10(b).\textsuperscript{125}

The court found “likewise unavailing” the plaintiff’s allegation that “Citigroup’s failure to disclose that its revenues were derived from ‘unsustainable and illegitimate sources’ violated section 10(b).”\textsuperscript{126} The court held: “[T]he federal securities laws do not require a company to accuse itself of wrongdoing.”\textsuperscript{127} Similarly, in \textit{Iron Workers Local Pension 16 Fund v. Hilb, Rogal & Hobbs Co.},\textsuperscript{128} the plaintiff alleged that the defendant insurance broker “hid [the broker’s] significant reliance on non-standard commissions in reporting its revenue and earnings” and that the broker “should have disclosed that its practices were illicit and improper.”\textsuperscript{129} The court, citing \textit{Santa Fe}, rejected this argument and held that the “[f]ederal securities laws do not require a company to accuse itself of wrongdoing.”\textsuperscript{130} The court explained:

Rule 10b-5 was not intended to provide shareholders with an avenue for relief against executives for alleged illegal practices or corporate

\begin{itemize}
\item \textsuperscript{123} \textit{Id.} at 375.
\item \textsuperscript{124} \textit{Id.}
\item \textsuperscript{125} \textit{Id.} (emphasis removed).
\item \textsuperscript{126} \textit{Id.} at 377.
\item \textsuperscript{127} \textit{Id.}
\item \textsuperscript{128} \textit{432 F. Supp. 2d} 571 (E.D. Va. 2006).
\item \textsuperscript{129} \textit{Id.} at 575, 586.
\item \textsuperscript{130} \textit{Id.}
mismanagement. ... Thus, even if [the defendant’s] commission practices were improper, Plaintiff has no claim for securities fraud on that basis alone. Moreover, securities laws do not require that [the company] had a duty to disclose illegal and illicit activities.\textsuperscript{131}

These cases and many others like them\textsuperscript{132} teach that section 10(b) requires statements to be \textit{substantively} accurate with respect to their subject matter, and does not require disclosures that management has acted dishonestly or incompetently, that management is inherently dishonest or incompetent, or that internal controls are weak. These cases explain why recovery is permitted for the losses caused by substantively misleading statements, and refute the notion that recovery is permitted for collateral damage: if recovery were permitted for collateral damage, recovery would be available for a failure to confess wrongdoing, dishonesty, or mismanagement—precisely what the case law holds there is no duty to disclose.\textsuperscript{133}

2. Materiality

Section 10(b) liability requires the defendant’s misstatement or omission to have been material, i.e., to have been of such a character that “disclosure of the omitted fact would have been viewed by the reasonable investor as having significantly altered the ‘total mix’ of information made available.”\textsuperscript{134} Section

\textsuperscript{131} Id. at 586–87 (internal quotation marks and citations omitted).

\textsuperscript{132} See, e.g., Biesenbach v. Guenther, 588 F.2d 400, 409 (3d Cir. 1978) (“In effect, appellants are stating that the failure to disclose the breach of fiduciary duty is a misrepresentation sufficient to constitute a violation of the Act. We refuse to adopt this approach which would clearly circumvent the Supreme Court’s holding in \textit{Santa Fe}.”); \textit{In re Am. Express Co. S’holder Litig.}, 840 F. Supp. 260, 269–70 (S.D.N.Y. 1993) (stating a company is not required to “accuse itself of antisocial or illegal policies”) (quoting GAF Corp. v. Heyman, 724 F.2d 727, 740 (2d Cir. 1983)); Ciresi v. Citicorp, 782 F. Supp. 819, 823 (S.D.N.Y. 1991) (“[T]he law does not impose a duty to disclose uncharged, unadjudicated wrongdoing or mismanagement.”); Merritt v. Colonial Foods, Inc., 499 F. Supp. 910, 914 (D. Del. 1980) (“[T]he gravamen of plaintiff’s federal claim ... reduces inescapably to a claimed failure to confess one’s corporate wrongdoing. ... [T]he Supreme Court’s holding in \textit{Santa Fe} is not to be easily circumvented.”).

\textsuperscript{133} This is not to say that failures to disclose facts about management’s honesty or competence can never give rise to liability. Where, for example, a company misrepresents management’s education or experience, section 10(b) liability might well arise. See, e.g., Suez Equity Investors LP v. Toronto-Dominion Bank, 250 F.3d 87, 92 (2d Cir. 2001). The issue we address is different; the issue is whether a false or misleading statement on a subject that does not particularly relate to management’s honesty or competence (quarterly revenues, for example) can give rise to liability for collateral damage when the fact of the misrepresentation is revealed and the market reassesses management’s honesty or competence.

10(b)’s materiality requirement would appear at first blush to be satisfied whenever a company fails to disclose that its management is dishonest or incompetent in making a statement to investors, or that its statements are unreliable because they are the product of weak internal controls. After all, what reasonable investor would not view such a disclosure as affecting the “total mix” of information to consider in making an investment decision?

That observation proves too much, however, for it would write the materiality requirement out of the law. Suppose, for example, that a plaintiff shows that management knowingly lied to investors by telling them that the corporate offices had just been painted blue, when in fact they had been painted green. The plaintiff would have satisfied two elements of section 10(b) liability—falsity and scienter. But, absent facts that are hard to imagine, the plaintiff would not be able to establish materiality because reasonable investors do not care what color the corporate offices are painted. The plaintiff’s section 10(b) claim would fail even though management made a false statement, even though it did so knowingly, and even though reasonable investors care very much if management is going around lying to investors. If the plaintiff could show materiality on the theory that a knowingly false statement is ipso facto material—or put another way, that management failed to disclose that the statement was knowingly false—then there would be no separate materiality requirement; it would be subsumed within the falsity and scienter elements. A plaintiff, therefore, must show that a statement is substantively material, i.e., material as to its subject matter.

The courts have held as much for decades. In Britt v. Cyril Bath Co.,\textsuperscript{135} for example, the court explained:

\begin{quote}
[A] nondisclosure is neither material nor immaterial; it is the fact itself which is known to the corporate insider but not disclosed when there is a duty to do so, which is either material or immaterial to the market price of the stock. The difference between the two approaches is not a mere question of semantics because the proper focus of inquiry should be on the materiality of the undisclosed facts themselves. The materiality of the undisclosed facts is the prime issue in determining whether a nondisclosure is a violation of Rule 10b-5.\textsuperscript{136}
\end{quote}

Similarly, in United States v. Reyes,\textsuperscript{137} the defendant was convicted of engaging in securities fraud in connection with stock option backdating, and he argued in his motion for a new trial that the prosecutor improperly had invited the jury to infer materiality from the mere fact that a misstatement had been made. The Reyes court, in discussing this issue, stated: “[I]nvestors must care about the

\footnotesize{\begin{small}
\textsuperscript{135} 417 F.2d 433 (6th Cir. 1969).
\textsuperscript{136} Id. at 437.
\textsuperscript{137} No. C 06-00556-1 CRB, 2007 WL 2462147 (N.D. Cal. Aug. 29, 2007), rev’d, Nos. 08-10047, 08-10140, 2009 WL 2501920 (9th Cir. Aug. 18, 2009) (overturning defendant’s conviction because of prosecutorial misconduct).
\end{small}}
information misrepresented, not merely *that* it was misrepresented. If a misrepresentation is deemed material simply because it is a misrepresentation, then the law’s materiality requirement is altogether meaningless.”

The foregoing case law makes clear that a failure to confess managerial dishonesty or incompetence is immaterial as a matter of law as far as the securities laws are concerned. Permitting recovery for collateral damage would permit recovery not just for the portion of the defendant’s misstatement that is substantively false or misleading, but for the portion that consists of not disclosing dishonesty or mismanagement in making the statement—precisely what the foregoing case law teaches cannot give rise to liability.

3. Transaction Causation

The foregoing case law regarding materiality also bears on the element of transaction causation, or “but for” causation.139 Case law makes clear that transaction causation generally requires a showing that the plaintiff actually or constructively relied on the defendant’s statement or omission.140 The cases hold, however, that reasonable investors, as a matter of law, do not rely on statements or omissions that are immaterial.141 The cases holding failures to disclose managerial

138 Id. at *9.
139 See, e.g., Berkeley Inv. Group, Ltd. v. Colkitt, 455 F.3d 195, 222 (3d Cir. 2006) (“In the securities realm, but for causation is referred to as . . . transaction causation . . . . In order to establish . . . transaction causation, a Section 10(b) plaintiff must prove that but for the fraudulent misrepresentation, the investor would not have purchased or sold the security.”); Binder v. Gillespie, 184 F.3d 1059, 1065 (9th Cir. 1999) (“The requirement of transaction causation is equivalent . . . in tort liability terms[ ] [to] but-for causation.”); Harris v. Union Elec. Co., 787 F.2d 355, 366 (8th Cir. 1986) (“Transaction causation is established when the plaintiff shows that the defendant’s fraudulent conduct caused the plaintiff to engage in the transaction in question. This is nothing more than ‘but for’ causation . . . .”) (citation omitted).
140 See Stoneridge Inv. Partners LLC v. Scientific-Atlanta, Inc., 128 S. Ct. 761, 768 (2008) (“In a typical § 10(b) private action a plaintiff must prove . . . reliance upon the misrepresentation or omission . . . .”); Basic, Inc. v. Levinson, 485 U.S. 224, 241–49 (1988) (transaction causation can be satisfied by the presumption of reliance afforded by the fraud-on-the-market doctrine). Some courts have held that transaction causation can be satisfied in other ways in unusual cases, such as where the plaintiff alleges that but for the reliance of a third party (such as a government regulator), the plaintiff would not have purchased the security. See Arthur Young & Co. v. United States Dist. Court, 549 F.2d 686, 695 (9th Cir. 1977); In re Am. Cont’l Corp. / Lincoln Sav. & Loan Sec. Litig., 140 F.R.D. 425, 433–35 (D. Ariz. 1992); Mishkin v. Peat, Marwick, Mitchell & Co., 658 F. Supp. 271, 273–74 (S.D.N.Y. 1987).
141 See, e.g., In re Amdocs Ltd. Sec. Litig., 390 F.3d 542, 548 (8th Cir. 2004) (“Alleged misrepresentations can be immaterial as a matter of law if . . . no reasonable investor would rely upon them . . . .”); Amalgamated Bank v. Coca-Cola Co., No. 1:05-CV-1226, 2006 WL 2818793, at *3 (N.D. Ga. Sep. 29, 2006) (“[E]ven where a statement is
dishonesty, incompetence, or weak internal controls immaterial as a matter of law thus also lead to the conclusion that, as a matter of law, there can be no transaction causation with respect to such failures to disclose.

4. Loss Causation

Loss causation is the securities law equivalent of proximate causation, and requires a showing “that the act or omission of the defendant . . . caused the loss for which the plaintiff seeks to recover damages.” The loss causation element ensures that the securities laws “protect [investors] against [only] those economic losses that misrepresentations actually cause.”

A misrepresentation can “actually cause” collateral damage, at least in some sense. When a misrepresentation is revealed, the market may reassess the reputation of management and the company and accordingly adjust the price of the stock downward. Absent the misrepresentation, such an adjustment would not have taken place (holding all other factors constant), and the misrepresentation thus would have “led to [the] loss.” Further, that a misrepresentation can have this effect is entirely foreseeable because it is reasonable that revelation of a false statement may cause the market to reassess management’s honesty or competence or the quality of the company’s controls—and foreseeability long has been the touchstone of proximate causation.

Loss causation nevertheless would appear to be lacking as a matter of law when it comes to collateral damage. “[T]o establish loss causation, a plaintiff must

intentionally misleading as to the speaker’s true opinion or prediction, it may be held immaterial as a matter of law [if] no reasonable investor would rely upon it.”).

See Dura Pharm., Inc. v. Brodsky, 544 U.S. 336, 344 (2005) (noting “the common-law roots of the securities fraud action (and the common-law requirement that a plaintiff show actual damages)” and holding that the defendant’s statement must “proximately cause the relevant economic loss”); Berkeley, 455 F.3d at 222 (“Causation in the securities context is strikingly similar to the familiar standard in the torts context, but with different labels. In the securities realm, . . . ‘proximate cause’ is known as ‘loss causation.’”).

143 15 U.S.C. § 78u-4(b)(4) (2006); see also Dura, 544 U.S. at 342 (requiring “a causal connection between the material misrepresentation and the loss”).

144 Dura, 544 U.S. at 345.

145 Id. at 342.

146 See generally Palsgraf v. Long Island Rail Co., 162 N.E. 99 (N.Y. 1928) (seminal case on proximate causation and the closely related tort law concept of duty); see also McCabe v. Ernst & Young LLP, 494 F.3d 418, 430 (3d Cir. 2007) (“The loss causation inquiry typically examines how directly the subject of the fraudulent statement caused the loss, and whether the resulting loss was a foreseeable outcome of the fraudulent statement.”); Lattanzio v. Deloitte & Touche LLP, 476 F.3d 147, 156 (2d Cir. 2007) (“Loss causation . . . is intended to fix a legal limit on a person’s responsibility even for wrongful acts, and . . . it requires that the plaintiff’s loss be foreseeable.”) (internal quotation marks and citations omitted).
What is the “subject of the fraudulent statement or omission” in this context? Again, a false statement or omission can be viewed as having two components: a substantive statement or omission; and an omission to disclose that management was dishonest, reckless, or incompetent in making the statement or omission, or that the statement or omission was the product of weak internal controls. The first component might well cause harm when the truth is revealed and might well result in the recovery of out-of-pocket damages, but it does not cause collateral damage. Only the second component causes collateral damage, but, as shown above, that is the component with which the securities laws do not concern themselves. Accordingly, when it comes to the relevant “subject of the fraudulent statement or omission,” loss causation is lacking.

Moreover, loss causation, like common law proximate causation, is not just analytically driven—it is policy driven as well. “[L]oss causation focuses on whether the defendant should be held responsible as a matter of public policy for the losses suffered by the plaintiff.” There may be strong equitable arguments for holding defendants liable for collateral damage, as discussed below. There are also strong arguments for disallowing recovery for collateral damage, however, even if permitting the same would be analytically sound. These include the inherent difficulty, if not impossibility, of estimating the amount of the stock price decline attributable to collateral damage, and the in terrorem effects of allowing open-ended and unpredictable liability, among others. Such policy considerations counsel in favor of treading cautiously in this area.

5. The “In Connection With” Requirement

The “in connection with” requirement of section 10(b) is liberally construed and easily satisfied in many cases. Although the “in connection with”
requirement in the abstract neither supports nor undercuts the notion of permitting 
recovery for collateral damage, it is relevant here because it tends to rebut, at least 
to some degree, what would otherwise be a strong policy argument in favor of 
permitting such recovery.

The argument is one of culpability and foreseeability. When a defendant 
violates section 10(b) by making a false statement to investors with scienter, the 
defendant in many cases should be able to foresee that when the falsity is revealed, 
collateral damage may result. As between the culpable defendant—who could 
foresee that investors would suffer the collateral damage—and the innocent 
investors, it would seem entirely appropriate to require the defendant to be the one 
to bear that loss.

This policy argument, however, though certainly equitable, proves too much 
because it runs headlong into the “in connection with” requirement. This argument 
counsels in favor of permitting recovery for collateral damages not only for 
defrauded purchasers, but also for existing stockholders who did not purchase 
during the relevant period.

For example, suppose that a defendant fraudulently states on July 22 that 
second-quarter revenues were $100 million when in fact they were $90 million, 
and that this false information inflates the stock price by $5 per share. Suppose 
further that the defendant’s fraud is revealed a month later on August 22 and that, 
as in our example above, the stock price declines $15 per share—$5 per share to 
correct the inflation once the market impounds the truth about the revenues, and 
$10 per share in collateral damage. The defendant would be liable under 
section 10(b) to everyone who purchased between July 22 and August 22 (the class 
period). If the defendant is to be liable to the class not only for the $5 loss per 
share it suffered when the market corrected the inflation, but, on a theory of equity, 
also for the additional $10 loss per share in the form of collateral damage, why 
stop there? Why not also hold the defendant liable to everyone who purchased 
before the class period and held the shares through the subsequent decline? After 
all, the defendant equally could foresee that those investors also would suffer 
collateral damage when the price of their shares declined. Yet the defendant cannot 
be liable to such “holders” under section 10(b) because they did not purchase their 
shares “in connection with” the fraud. Arguing that the defendant should be

153 See Dabit, 547 U.S. at 80 (noting that only persons who purchased or sold in 
connection with the fraud have standing to sue under section 10(b)); Blue Chip Stamps v. 
Manor Drug Stores, 421 U.S. 723 (1975). Holders can avail themselves of statutory or 
common law claims under the laws of numerous states. See, e.g., Small v. Fritz Cos., 65 
P.3d 1255 (Cal. 2003). However, the Securities Litigation Uniform Standards Act of 1998, 
112 Stat. 3227 (“SLUSA”), precludes such causes of action from being brought on a class 
basis. See 15 U.S.C. § 78p(b) (2006) (“No covered class action based upon the statutory or 
common law of any State . . . may be maintained in any State or Federal court by any 
private party alleging . . . an untrue statement or omission of material fact in connection
liable for collateral damage based on equitable considerations therefore tends to prove too much.

V. CONCLUSION

Dura teaches that an investor does not suffer a loss—even if the investor overpaid for the stock—unless and until the stock price declines upon a corrective disclosure.\textsuperscript{154} Dura explains that focusing on whether the stock price declines is the proper way to assess causation of damages. Dura’s approach, however, all too often is used to estimate the amount of those damages—with the plaintiffs, and sometimes the defendants, assuming that the entire amount of the stock price decline upon the corrective disclosure represents the amount of the inflation for which recovery should be allowed.

Causation and damages, however, are two distinct inquiries, and the issues should not be conflated. Causation focuses on the stock price decline; damages focus on inflation before the decline. Once causation is established, the analysis shifts, and the parties and their experts must set about estimating the amount of the inflation so as to avoid permitting recovery for collateral damage—which as shown above, is not recoverable under section 10(b).

\footnote{Dabit, 547 U.S. at 71 (holding that SLUSA precludes holders’ claims brought on a class basis).}

\footnote{See supra note 72 and accompanying text.}
FROM IDENTITY POLITICS TO IDEOLOGY POLITICS

Jessica Knouse*

TABLE OF CONTENTS

I. INTRODUCTION .................................................................750
II. THE PROBLEM: IDENTITY POLITICS .................................751
   A. Introducing Identity Politics .............................................751
   B. Understanding Identity Group Members ............................752
      1. Three Types of Attributes ...........................................752
      2. Two Conflicting Influences ..........................................755
   C. Understanding Identity Groups ........................................758
      1. Existing Vocabularies .................................................758
      2. A New Vocabulary ....................................................759
   D. Examining the Impact of Identity Politics .........................761
      1. On Identity Group Members and Identity Groups ..........762
      2. On Society as a Whole ...............................................763
III. THE SOURCE OF THE PROBLEM: EQUAL PROTECTION DOCTRINE 764
   A. Our History of Subordination .........................................765
   B. Our Constitutional Commitment to Equality ....................767
      1. Abolishing the Black Codes ........................................768
      2. Instituting an Equal-Birth Principle ............................769
   C. The Supreme Court’s Equal Protection Doctrine ...............770
      1. Phase One—Protecting One Identity Group ...................770
      2. Phase Two—Protecting Multiple Identity Groups ..........772
      3. Phase Three—Preventing Identity-Based Groupings .........782
IV. A SOLUTION: A NEW DOCTRINE AND A NEW POLITICS .........785
   A. Arguments Against Existing Equal Protection Doctrine ........785
      1. From Pragmatism ......................................................785
      2. From Originalism .....................................................787
      3. From Structuralism ..................................................788
   B. Suggestions for a New Doctrine and a New Politics ..........792
V. CONCLUSION .................................................................795

* © 2009 Jessica Knouse, Assistant Professor at the University of Toledo College of Law. She would like to thank her colleagues at the College of Law—especially Rebecca Zietlow, Melissa Hamilton, Nicole Porter, Lee Strang, Geoff Rapp, Bruce Campbell, and Bill Richman—the attendees of the January 2009 Ohio Legal Scholarship Workshop, and Lynn Izod for all their comments and guidance. She would additionally like to thank Holly Dombkowski for her expert research assistance, the College of Law for its generous financial support, the Utah Law Review for its impeccable editing, and, most of all, Mike Loegering for his encouragement, wisdom, and love.
I. INTRODUCTION

We live in a lacuna between identity politics and post-identity politics. Although we understand that identity groups like African Americans and women are not ideologically monolithic, we continue to use race and gender as proxies for ideology. Although some have proclaimed the 2008 presidential election the first “post-identity politics election,”¹ issues of race and gender continue to dominate the public forum.² In short, although we want to have transcended identity politics, we have not. This Article asks why not, and concludes that the existing equal protection doctrine is partly to blame. It proposes that we abandon the existing doctrine, which not only has mired us in identity politics, but also has dishonored our constitutional commitments to equality, autonomy, and democracy. It suggests that we replace the existing doctrine with a new doctrine, which would inaugurate a new regime of “ideology politics” and honor our broad range of constitutional commitments.

Part II illustrates that identity politics is a harmful institution. It begins by defining identity politics and then examines how identity politics impacts: (1) identity group members as individuals, (2) identity groups as groups, and (3) society as a whole. Part II concludes that, although identity politics may benefit identity groups as groups by increasing their political power, it harms identity group members as individuals by diminishing their autonomy, and society as a whole by impeding democracy.

Part III considers why such a harmful institution so robustly persists. It begins by observing that our history of racism, sexism, homophobia, and other subordinating social institutions has produced numerous identity groups. It argues that, although social institutions have created these groups, legal institutions—such as equal protection doctrine—have converted them into entrenched political actors. Because our constitutional commitment to equality, expressed in the Fourteenth Amendment, has been deployed through a doctrine structured around identity groups, we have come to view ourselves and our political landscape in terms of such groups. The existing doctrine thus supports identity politics and all of its resulting harms.

Part IV argues that existing equal protection doctrine is fatally flawed—not only because it supports identity politics, but also because it prevents us from honoring our constitutional commitments to equality, autonomy, and democracy. It calls for the replacement of existing doctrine with a new doctrine that would inaugurate a regime of “ideology politics” free from the harms of identity politics, while also honoring the full range of our constitutional commitments.

II. THE PROBLEM: IDENTITY POLITICS

Part II is divided into four subparts. The first three subparts develop the vocabulary that is used throughout the fourth subpart to illustrate the harms of identity politics. Part II(A) defines “identity politics.” Part II(B) defines the attributes that collectively create individual identity (for purposes of discussing individual identity group members), and Part II(C) defines “identity groups.” Part II(D) analyzes the impact of identity politics and concludes that, although it may benefit identity groups as groups, it detriments identity group members as individuals along with society as a whole and is, on balance, a harmful institution.

A. Introducing Identity Politics

We all know vaguely what “identity politics” means. When women vote for female candidates because of their sex or African Americans vote for African American candidates because of their race, that is identity politics. In mainstream culture (via Wikipedia), identity politics is defined as “political action to advance the interests of members of a group whose members perceive themselves to be oppressed by virtue of a shared and marginalized identity.” In more rigorous sociological terms, it is “activism engaged in by status-based social movements,” and it can be contrasted with class activism and cultural activism. Identity politics rejects class activism’s neo-Marxist view that economic inequality is the primary source of oppression and cultural activism’s postmodern (or queer) view that identity categories are themselves the primary source of oppression.

While identity politics may be incompatible to varying degrees with both class and cultural activism, it is not, as is often assumed, incompatible with a respect for the intersection of multiple identities within individuals. Although the origin of the term “identity politics” is uncertain, some have traced it to the Combahee River Collective of the 1970s—a group that described itself as
“committed to struggling against racial, sexual, heterosexual, and class oppression.”

History, then, reveals that identity politics is not inherently incapable of respecting intersectionality and that, when contemporary identity groups ignore the multiple identities that define their members, they are making a choice rather than accepting an inevitability.

From the above, we can preliminarily define “identity politics” as “political activism by identity groups.” This definition is, however, meaningless without elaboration. It tells us nothing about individual identity group members, the characteristics of identity groups, or the nature of their political activism. The following three subparts are, thus, respectively dedicated to developing the vocabulary necessary to understand individual identity group members,9 the characteristics of identity groups,10 and the nature of their political activism.11

B. Understanding Identity Group Members

Part II(B) identifies and defines three types of attributes that, in my view, collectively create individual identity. Although these attributes are present in all individuals, this Article focuses on individuals in identity groups.12

1. Three Types of Attributes

It is useful to view individuals as composed of three basic types of attributes: (1) anatomical attributes, (2) quasi-anatomical attributes, and (3) non-anatomical attributes. These three types of attributes obviously do not capture every complexity of individual identity, nor are they meant to. They simply divide a relatively continuous spectrum of individual identity into a few meaningful segments for purposes of discussion.

Anatomical attributes are instantiated in the body and are readily detectable through physical examination. They include eye color, hair color, skin color, and sex organs. Some anatomical attributes, such as eye color and hair color, are of little cultural significance; others, such as skin color and sex organs, are of deep cultural significance.13 Although anatomical attributes are subject to alteration, those that are of deep cultural significance are often so difficult to alter—requiring

8 Harris, supra note 7, at 300. Smith, an instrumental member of the Collective, explained that “identity politics” originally referred to “a politics that grew out of our objective material experiences as Black women.” Id.

9 See infra Part II.B.

10 See infra Part II.C.

11 See infra Part II.D.

12 See infra Part II.D.

13 I categorize traits as “culturally significant” if they have been and continue to be markers of egregious social inequality. Although I do not dispute the importance of eye and hair color—or deny that they are ever markers of inequality—they have not been so instrumental in creating and maintaining social hierarchies as skin color and sex organs have.
surgery, hormone treatments, and other invasive procedures—that we may consider them for our present purposes functionally immutable.\textsuperscript{14}

Quasi-anatomical attributes are quite similar to anatomical attributes. They are instantiated in the body, but are not readily detectable through physical examination. They include handedness, sexual orientation, and certain disabilities.\textsuperscript{15} Some quasi-anatomical attributes, such as handedness, are of little cultural significance; others, such as sexual orientation, are of deep cultural significance.\textsuperscript{16} Although quasi-anatomical attributes are sometimes subject to alteration (or, at least, to the appearance of alteration),\textsuperscript{17} they are often so difficult to alter—requiring years of therapeutic and other treatments, which may ultimately fail—that we may consider them for present purposes as functionally immutable.\textsuperscript{18}

\textsuperscript{14} The fact that some anatomical attributes, such as eye color and hair color, may be easily altered—with, respectively, contact lenses or chemical-based dyes—is irrelevant for our present purposes because these anatomical attributes do not generally take on deep cultural significance.

\textsuperscript{15} Examples of disabilities that are instantiated in the body but not readily detectable through physical examination might include learning disabilities, certain heart conditions, fibromyalgia, and chronic fatigue syndrome.

\textsuperscript{16} Although handedness may have in the past been a marker of egregious social inequality, it is at present much less so. See Elaine F. Costas, The Left-Handed: “Their Sinister” History 2 (Education Resources Information Center, Paper No. 399519, 1996), available at http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/14/b3/93.pdf (describing the historical discrimination suffered by left-handed children, but concluding that “[p]arents and teachers are more realistic now and accept children’s natural tendencies for left- or right-handedness”). I do not assert that left-handedness is no longer a source of oppression, but merely that the oppression currently experienced by left-handed individuals is not on the same level as that experienced by gay, lesbian, or bisexual individuals. This is arguably evidenced by the fact that a Westlaw search for variations of “left handed” within ten words of variations of “discriminate” retrieves no cases, whereas a search for variations of “homosexual” within ten words of variations of “discriminate” retrieves 645 cases. Some, however, might argue that, rather than providing evidence of a lack of discrimination, these statistics provide evidence of a lack of awareness of discrimination. See John J. Sciortino, Essay, Sinistral Legal Studies, 44 SYRACUSE L. REV. 1103, 1118 (1993) (stating that despite “the overwhelming right-handedness of the world [and the natural expectation] that legal challenges to these biases would abound . . . American courts have only heard one case in which someone alleged that he was discriminated against because of being left-handed”).

\textsuperscript{17} The theme of attributes being subject to alteration or the appearance of alteration resurfaces numerous times. The distinction between true alteration and apparent alteration is important but virtually impossible to quantify in any given individual. It is possible that an individual may begin with one set of attributes, be culturally pressured into creating an appearance of alteration, and over time incorporate the appearance of alteration so fully into the individual’s identity that he or she is most accurately described as having actually altered the attributes.

\textsuperscript{18} Attempts to alter sexual orientation are, for example, notoriously unsuccessful. See, e.g., Gregory M. Herek, Myths About Sexual Orientation: A Lawyer’s Guide to Social Science Research, 1 LAW & SEXUALITY 133, 151 (1991) (“Even for the relatively small
Non-anatomical attributes differ from their counterparts. They are neither directly instantiated in the body—though they are, as will be discussed, influenced by biology\(^\text{19}\)—nor readily detectable through physical examination. They include personality, which encompasses degree of openness, conscientiousness, extraversion, agreeableness, and neuroticism;\(^\text{20}\) behavior, which encompasses dress, hairstyle, and mannerisms; and ideology, which encompasses politics, philosophy, and religion. Some non-anatomical attributes are of little cultural significance (e.g., certain behavioral attributes); others are of great cultural significance (e.g., certain political positions like socialism and anarchism and certain religious beliefs). Non-anatomical attributes, as will be illustrated, may in some circumstances be more susceptible than their counterparts to alteration—or, again, at least the appearance of alteration—under the influence of cultural norms.\(^\text{21}\)

\(^{19}\) Consider, for example, the so-called “God gene” that creates a propensity toward faith in some people. See, e.g., Jeffrey Kluger, Is God in Our Genes?, TIME, Oct. 25, 2004, at 62.

Our genome may contain information about our personalities, behaviors, and ideologies. Prominent psychologist and cognitive scientist Steven Pinker recently made the following statement:

The most prominent finding of behavioral genetics has been summarized by the psychologist Eric Turkheimer: “The nature-nurture debate is over. . . . All human behavioral traits are heritable.” By this he meant that a substantial fraction of the variation among individuals within a culture can be linked to variation in their genes. Whether you measure intelligence or personality, religiosity or political orientation, television watching or cigarette smoking, the outcome is the same. Identical twins (who share all their genes) are more similar than fraternal twins (who share half their genes that vary among people). Biological siblings (who share half those genes too) are more similar than adopted siblings (who share no more genes than do strangers). And identical twins separated at birth and raised in different adoptive homes (who share their genes but not their environments) are uncannily similar.

Steven Pinker, My Genome, My Self, N.Y. TIMES MAGAZINE, Jan. 11, 2009, at 28.

\(^{20}\) These traits derive from a common personality model. See Lewis R. Goldberg, The Structure of Phenotypic Personality Traits, 48 AM. PSYCHOLOGIST 26, 26–27 (1993) (discussing the “Big-Five” personality factors).

\(^{21}\) Altering one’s personality, behavior, or ideology, while significant, seems to require less effort than altering one’s quasi-anatomical or anatomical attributes. Compare examples of the three types of alteration: first, a person with the “God-gene” in an atheist culture might stop attending church in an attempt to reform his ideology; second, a gay male in a heterosexist culture might undergo therapy in an attempt to reform his sexual orientation; third, a large-nosed person in a small-nose-obsessed culture might undergo surgery in an attempt to reform his anatomy. Regardless of whether these methods of
Different types of attributes provide the bases for different types of groups. Anatomical and quasi-anatomical attributes, due to their functional immutability, often provide the bases for relatively stable political groups. African Americans, for example, due to the immutability of skin color, cannot easily dissociate themselves from the Civil Rights Movement. Openly gay men similarly—due to the immutability of sexual orientation—cannot easily dissociate themselves from the gay rights movement. In contrast, non-anatomical attributes, due to their relative mutability, often fail to produce stable political groups. Socialists, for example, due to the mutability—or apparent mutability—of ideology in response to cultural pressures, can easily dissociate themselves from and thereby diffuse socialist political movements. Because anatomical and quasi-anatomical attributes are more stable than non-anatomical attributes, they have often been the preferred bases for political groups.23

2. Two Conflicting Influences

Having established that individual identity is the product of three types of attributes, this Article now considers how each type of attribute acquires its content within a given individual. I posit that each of the three types of attributes is influenced by biology and culture, as depicted in the following diagram:24

reforms are successful or not, the first method appears far less burdensome than the second and third.

22 African Americans are presumptively part of any African American political movement. They may attempt to defect by making a public statement. Such statements, however, are often not heard or are not understood due to the strength of the initial presumption that they are part of the movement.

23 Daria Roithmayr contemplates the relative advantages of anatomical and quasi-anatomical groups over non-anatomical groups:

I have long had my doubts about identity politics as a useful source of political struggle. . . . On the political level, I have wondered whether we are better off mobilizing a strategic coalition of fellow travelers who share our particular political commitments rather than our . . . identities. . . . [But if] I buy into identity politics, I buy into it as a strategic matter, as a way of advancing the political commitments I care about. I do so because I acknowledge that the rest of the world thinks strategically about identity.


24 This diagram is similar, though not identical, to those included in Jessica Knouse, Intersexuality and the Social Construction of Anatomical Sex, 12 CARDOZO J.L. & GENDER 135, 137 (2005) [hereinafter Knouse, Intersexuality], and Jessica Knouse, Using Postmodern Feminist Legal Theory to Interrupt the Reinscription of Sex Stereotypes Through the Institution of Marriage, 16 HASTINGS WOMEN’S L.J. 159, 162 (2005), both of which were inspired by the model set forth by Kenji Yoshino. See Yoshino, supra note 24, at 865-75.
Biology emanates outward, exerting a strong influence on anatomical attributes, a weaker influence on quasi-anatomical attributes, and a still weaker influence on non-anatomical attributes. Culture, in contrast, focuses inward, exerting a strong influence on non-anatomical attributes, a weaker influence on quasi-anatomical attributes, and a still weaker influence on anatomical attributes.

Biology and culture are in constant competition. Because biology is immutable and culture is highly variable, the result of the competition varies from culture to culture. Consider the contrasting results in cultures with strong versus weak norms. In a culture with strong norms, biology might be repressed over all three types of attributes—rendering culture the clear winner. Imagine, for example, a culture that required schoolchildren to recite a pledge of allegiance, criminalized same-sex sodomy, and performed genital normalizing surgery on its intersex infants. The requirement of pledge recitation would normalize ideology; the

---

25 Norms are, in short, social rules. See Merriam-Webster’s Online Dictionary, Norm, http://www.merriam-webster.com/dictionary/norm (last visited July 10, 2009) (defining a norm as, inter alia, “a principle of right action binding upon the members of a group and serving to guide, control, or regulate proper and acceptable behavior”). As will be illustrated, norms often arise from cultural associations between anatomical and non-anatomical attributes. There is, for example, a norm that all individuals with female sex organs wish to be wives and mothers.
criminalization of same-sex sodomy would normalize sexuality; and the performance of genital normalizing surgery would normalize sex organs. Imagine, in contrast, a culture with weak norms, where biology was expressed over all three types of attributes. In the former culture, individuals would be deprived of autonomy; in the latter, they would be free to self-determine according to their own biological predilections.

We all, however, reside in multiple cultures—some defined broadly and some narrowly. A broadly defined culture might extend across geographic and temporal boundaries to include all of Western civilization since the rise of ancient Greece. A narrowly defined culture might be limited by geographic, temporal, and identity-group boundaries. It might, for example, be described as “white women living in Toledo, Ohio, in 2009.” Inasmuch as we all reside in multiple cultures, we are all subject to multiple—and often conflicting—sets of norms. For purposes of clarity, I use the term “cultural norms” to refer to the norms imposed by Western society generally, and the term “group norms” to refer to the norms imposed by identity groups specifically. Both cultural and group norms compete with biological influences, and such influences can reliably prevail only when both cultural and group norms are relatively weak.

The result of the competition between culture and biology within any given individual, thus, depends upon the particular cultures the individual inhabits and the relative strength of the norms those cultures impose. Although the result varies among individuals, most individuals in modern American society exhibit certain similarities. On average, biology seems most likely to govern anatomical attributes and least likely to govern non-anatomical attributes. Indeed, while biology is rarely repressed in anatomical attributes, it is sometimes repressed in quasi-anatomical attributes, and often repressed in non-anatomical attributes. Individuals, for example, rarely undergo surgery to comply with anatomical norms, but sometimes “pass” as straight or nondisabled to comply with quasi-anatomical

---

26 This is, indeed, a very long period, since the rise of Greece is often dated to the late eighth century B.C. See Jasper Griffin, Introduction to THE OXFORD HISTORY OF GREECE AND THE HELLENISTIC WORLD 1, 1–2 (John Boardman, Jasper Griffin & Oswyn Murray eds., 2001).

27 Anatomical attributes are rarely repressed to comply with cultural or group norms. When repression does occur, it may be accompanied by consent—as in the case of an individual electing to undergo cosmetic surgery—or imposed by force—as in the case of a doctor performing genital normalizing surgery on an intersex infant. See Knouse, Intersexuality, supra note 24, at 139 (“Infants born with ambiguous external morphology are often surgically altered such that they conform with either male or female anatomical stereotypes.”).

28 The term “pass” refers to the repression (without actual alteration) of a quasi-anatomical attribute. Professor Yoshino has identified “passing” as a middle-ground between “converting,” where an individual actually alters the attribute in question, and “covering,” where an individual does not alter or repress, but merely “downplay[s],” the attribute in question. See Kenji Yoshino, Covering, 111 Yale L.J. 769, 772 (2002) (stating that “[p]assing means the underlying identity is not altered, but hidden”).
norms, and often alter their personalities, behaviors, and ideologies to comply with non-anatomical norms. This final phenomenon—the cultural normalization of ideology—is caused, in part, by identity politics.

C. Understanding Identity Groups

Having developed a vocabulary for discussing individual identity group members, we can now develop a vocabulary for distinguishing identity groups from other types of groups. The first subpart of Part II(C) examines past scholarship to identify existing vocabularies for discussing various types of groups. The second subpart draws on that scholarship to develop a new vocabulary for discussing the three types of groups most relevant to this project: (1) phenomenal groups, (2) identity groups, and (3) ideology groups.

1. Existing Vocabularies

Inasmuch as a group is simply an “aggregate of individuals having some characteristic in common,” one could easily identify numerous types of groups. Iris Young, in *Justice and the Politics of Difference*, distinguishes among three primary types of groups: (1) aggregates, (2) associations, and (3) social groups. Young defines aggregates as groups based on a shared attribute, which may be either voluntary (e.g., make of car or street of residence) or involuntary (e.g., eye color, skin color, or age). She defines associations as groups based on a shared practice, which is generally voluntary. Associations include political parties, churches, and corporations. She defines social groups as groups based on a shared identity, explaining: “Though sometimes objective attributes are a necessary condition for classifying oneself or others as belonging to a certain social group, it is identification with a certain social status, the common history

———

29 Quasi-anatomical attributes, such as gayness, are sometimes repressed (or made to appear altered) to comply with cultural or group norms. Although during our nation’s early history gayness was often repressed, in more recent years it has become more socially acceptable and thus less often repressed. See Kenji Yoshino, *Covering: The Hidden Assault on Our Civil Rights* 31–107 (2006) (tracing the diminishing demands to repress gayness from conversion to passing to covering).

30 Non-anatomical attributes, such as behaviors related to dress and grooming, are often repressed to comply with cultural or group norms. The repression of non-anatomical attributes is generally not viewed as such a grave harm as the repression of quasi-anatomical or anatomical attributes.


33 *Id.* at 43–44.

34 *Id.* at 44.

35 *Id.*

36 *Id.*
that social status produces, and self-identification that define the group as a group."\(^{37}\) Social groups, which include blacks, women, and gays, have special significance. Although the attributes that define aggregates are “accidental” to identity and the practices that define associations are “external” to identity, the commonalities that unite social groups are directly constitutive of identity.\(^{38}\) Young’s social groups might well be the kind of groups that engage in identity politics, but they are not the only candidates.

Avishai Margalit and Joseph Raz, in *National Self-Determination*,\(^ {39}\) introduce the concept of “encompassing groups.” Margalit, in *The Decent Society*,\(^ {40}\) further elaborates on the concept. Encompassing groups are somewhat similar to Young’s social groups and, taken together, the two concepts provide a useful platform for defining what this Article ultimately refers to as “identity groups”—that is, the groups that engage in identity politics.

Margalit and Raz define encompassing groups as exhibiting six characteristics: (1) their members share a common culture, which may include language, music, cuisine, and dress; (2) their members’ children are “marked” by the common culture, such that it influences their life choices; (3) membership is based on mutual recognition; (4) membership is a central aspect of members’ identity; (5) membership is based on “belonging rather than achievement[,]” and is therefore nonvoluntary; and (6) their members do not personally know all other members, so group-related symbols are employed to enable mutual recognition.\(^ {41}\)

Margalit and Raz elaborate upon the fourth characteristic, the centrality of encompassing group membership to individual identity, by explaining that “people’s sense of their own identity is bound up with their sense of belonging to encompassing groups and . . . their self-respect is affected by the esteem in which these groups are held.”\(^ {42}\) This assertion about encompassing groups is quite similar to Young’s assertion that social groups are, at least in part, constitutive of their members’ identities.\(^ {43}\) There is, then, an important overlap between Margalit and Raz’s encompassing groups and Young’s social groups. Some combination of the two definitions might very likely describe the type of group that engages in identity politics.

2. *A New Vocabulary*

This subpart draws on Young, Margalit, and Raz to develop a new vocabulary of groups. It defines the three types of groups most relevant to this project: (1)

\(^{37}\) *Id.*  
\(^{38}\) *Id.*  
\(^{41}\) See Margalit & Raz, *supra* note 39, at 443–47.  
\(^{42}\) *Id.* at 449.  
\(^{43}\) YOUNG, *supra* note 32, at 44.
phenomenal groups;\(^\text{44}\) (2) identity groups, which engage in identity politics and currently dominate public forum debate; and (3) ideology groups, which I will later argue ought to be more prevalent in public forum debate. These groups can be defined by reference to their formation around various combinations of the previously discussed anatomical, quasi-anatomical, and non-anatomical attributes. As previously mentioned, different types of attributes will provide the bases for different types of groups.

Phenomenal groups, similar to Young’s aggregates,\(^\text{45}\) are defined by a shared anatomical or quasi-anatomical attribute.\(^\text{46}\) They are virtually infinite in number and include people with blue eyes, people with blond hair, people with female sex organs, left-handed people, etc. We all belong to numerous phenomenal groups, most of which have little or no cultural significance. Phenomenal groups can, however, under certain social conditions, evolve into identity groups.

Identity groups, similar to both Young’s social groups and Margalit and Raz’s encompassing groups, are phenomenal groups whose shared anatomical or quasi-anatomical attributes have become associated with subordinating cultural norms that dictate their members’ non-anatomical attributes.\(^\text{47}\) The phenomenal group of “people with female sex organs,” for example, evolved into the identity group of “women” when female sex organs became associated with subordinating cultural norms dictating that women ought to exhibit personalities entailing neuroticism and agreeableness,\(^\text{48}\) behaviors such as wearing dresses, makeup, and jewelry.\(^\text{49}\)


\(^\text{45}\) They are somewhat similar to Young’s aggregates except that, whereas aggregates can be based on either voluntary or involuntary attributes, phenomenal groups are based on (relatively) involuntary attributes. I characterize anatomical and quasi-anatomical attributes as “involuntary” because, although it is possible to voluntarily alter them, alteration is rare.

\(^\text{46}\) Note that, although the attribute must be instantiated in the body, it need not be immediately visible.

\(^\text{47}\) This association of anatomical or quasi-anatomical attributes with non-anatomical attributes is commonly referred to as “stereotyping.” See, e.g., Eliot R. Smith, Mental Representation and Memory, in 1 THE HANDBOOK OF SOCIAL PSYCHOLOGY 391, 400 (1998) (“Stereotypes have often been conceptualized as associative links between a node representing a social group and various traits and/or evaluations.”).

\(^\text{48}\) Studies suggest that women in many cultures view themselves as more neurotic and agreeable than men in those same cultures view themselves—but that sex-based personality differences are (somewhat surprisingly) most pronounced in Western cultures, specifically in European and American cultures. See Paul Costa, Jr., et al., Gender Differences in Personality Traits Across Cultures: Robust and Surprising Findings, 81 J. PERSONALITY & SOC. PSYCHOL. 322, 322 (2001). I argue that the apparent exaggeration of sex-based personality differences in Western cultures provides some evidence that women in Western cultures are encouraged and expected to exhibit personalities entailing neuroticism and agreeableness.

\(^\text{49}\) The facts of Price Waterhouse v. Hopkins, 490 U.S. 228 (1989), provide one
and ideologies reflecting domesticity. Such associations have generated numerous identity groups beyond women, including African Americans, individuals with disabilities, gays, etc.

Although identity groups are numerous, they are far fewer in number than phenomenal groups. Identity groups may, therefore, be of greater significance than phenomenal groups. Indeed, the social consequences of belonging to an identity group are far more serious than those of belonging to a phenomenal group. Identity groups, as will be illustrated, dictate their members’ level of political power (or powerlessness) and diminish their members’ personal autonomy.

Ideology groups, somewhat similar to Young’s associations, are defined by the shared non-anatomical attribute of ideology, which, in this context, refers to a philosophy or system of beliefs. Ideology groups include feminists, pro-choice advocates, queers, pacifists, and environmentalists. Although they resemble Young’s associations, there is one critical difference. While Young describes the practices that unite associations as external to identity, I view ideology as internal to identity—that is, as directly influenced by biology. Although biology influences ideology, biological influences on ideology are often repressed by cultural and group norms. As will be illustrated, identity politics supports the repression of biological influences on ideology, with highly problematic results.

D. Examining the Impact of Identity Politics

Based on the preceding discussion, we can elaborate on our preliminary definition of identity politics. We can say that “identity politics” is political activism by identity groups—i.e., phenomenal groups whose shared attributes have

example of how women are encouraged to wear dresses, makeup, and jewelry. Ann Hopkins, a Price Waterhouse employee, was advised that her chances of making partner might improve if she would “walk more femininely, talk more femininely, dress more femininely, wear make-up, have her hair styled, and wear jewelry.” Whether such advice is explicit or unspoken, many women understand that they are expected to wear dresses, makeup, and jewelry.

Justice Bradley’s concurrence in Bradwell v. Illinois, 83 U.S. 130 (1873) (allowing Illinois to deny women admission to the bar), provides a historical example of women being encouraged and expected to exhibit ideologies of domesticity. Justice Bradley, quite directly, stated that “[t]he paramount destiny and mission of woman are to fulfill [sic] the noble and benign offices of wife and mother.” Id. at 141 (Bradley, J., concurring).

50 Justice Bradley’s concurrence in Bradwell v. Illinois, 83 U.S. 130 (1873) (allowing Illinois to deny women admission to the bar), provides a historical example of women being encouraged and expected to exhibit ideologies of domesticity. Justice Bradley, quite directly, stated that “[t]he paramount destiny and mission of woman are to fulfill [sic] the noble and benign offices of wife and mother.” Id. at 141 (Bradley, J., concurring).

51 See infra Part II.D.

52 Ideology is often limited in meaning to refer to belief systems that “legitimate an unjust form of power.” TERRY EAGLETON, IDEOLOGY: AN INTRODUCTION 43 (1994). As John B. Thompson famously wrote: “To study ideology . . . is to study the ways in which meaning (or signification) serves to sustain relations of domination.” Id. I employ a broader meaning, which extends to all belief systems, including those that challenge the power of a dominant social group.

53 YOUNG, supra note 32, at 44–45 (“A person’s identity and sense of self are usually regarded as prior to and relatively independent of association membership.”).

54 See infra Part II.D.
become associated with subordinating cultural norms that dictate their members’ ideologies. Now that we understand what identity groups are, we can begin to understand what motivates them to engage in identity politics.

Although we are all subject to cultural norms, identity group members are uniquely subject to *subordinating* cultural norms. Thus, although cultural norms deprive us all of a certain measure of autonomy, they deprive identity group members of not only autonomy but also political power. For example, although cultural norms pressure men and women to adopt sex-specific roles, they uniquely pressure women to adopt roles that have historically been associated with political powerlessness (e.g., the roles of wife and mother). Identity groups engage in “identity politics” to vitiate the latter deprivation—the deprivation of political power—but not to vitiate the former, more universal deprivation of autonomy.

Part II(D) is divided into two subparts that examine the impact of identity politics at different levels of abstraction. The first subpart examines the impact of identity politics on identity group members and identity groups. It illustrates that, although identity politics may generate political power for identity groups, it does not provide ideological autonomy for identity group members. The second subpart examines the impact of identity politics on society as a whole. It illustrates that, in portraying identity groups as ideologically monolithic, identity politics distorts public forum debate and prevents us from making sound policy decisions. Identity politics, in sum, benefits identity groups while harming identity group members and society as a whole. It is, on balance, a harmful institution.

1. On Identity Group Members and Identity Groups

Identity groups engage in political activism to combat subordinating cultural norms. As previously mentioned, such norms inflicts at least two separate harms: they deprive identity group members of (1) ideological autonomy and (2) political power. Identity groups engage in political activism to vitiate the former, but not the latter, of the two harms. They, in fact, vitiate the former harm by *perpetuating* the latter harm. That is to say, identity groups acquire their political power by depriving their members of ideological autonomy. This is accomplished by portraying their members as an ideological monolith in pursuit of a single set of political objectives. Providing their members with ideological autonomy would, indeed, be directly adverse to the political interests of most identity groups.

Identity groups combat subordinating cultural norms primarily by replacing them with empowering group norms. Feminists, for example, have combated the subordinating cultural norm that all women want to be wives and mothers not by arguing that women are ideologically diverse, but rather by imposing the

---

55 See, e.g., Roithmayr, *supra* note 23, at 501 (“Perhaps playing identity politics is a war of position, in which we are forced to resist on the oppressor’s terms by mobilizing identity politics on behalf of our political commitments, to counteract the ways in which identity is used against us.”).
empowering group norm that all women espouse feminist ideologies. Although this tactic may have purchased a certain measure of political power for women as a group, it has not provided ideological autonomy for women as individuals. Identity politics is, therefore, best described as a battle over which norms ought to control rather than whether norms ought to control. Regardless of who wins this battle, identity group members will remain under the dominion of norms and without ideological autonomy. By waging the battle that it does, identity politics not only disregards a significant aspect of the harm inflicted by cultural norms, it also ignores the harm inflicted by group norms.

2. On Society as a Whole

This subsection illustrates that identity politics is not only bad for identity group members, but also bad for society—or, more specifically, for democratic societies that rely on public forum debate to make policy decisions. Identity politics, by portraying identity group members as ideologically monolithic, distorts public forum debate in at least two ways: (1) by converting a diverse array of individual viewpoints into a few identity group meta-ideologies, and (2) by misrepresenting public sentiment. As will be illustrated, the first distortion harms us by limiting our policy choices; the second, by leading us to choose poorly from among those limited choices.

Identity politics first distorts public forum debate by converting a diverse array of viewpoints into a few meta-ideologies. Under an identity politics regime, individuals enter the public forum as members of identity groups with recognized meta-ideologies rather than as ideological independents. When individuals are silent, their silence is interpreted as acquiescence in their identity group’s meta-ideology; when individuals speak, their speech is interpreted as a variant of their identity group’s meta-ideology. Identity politics, thus, prevents society as a whole from hearing unique or uncommon viewpoints, and causes public forum debate to operate in terms of oversimplified meta-ideologies. It reduces a biologically diverse array of individual ideologies into a few meta-ideologies. In silencing unique and uncommon viewpoints, identity politics limits our policy choices and thereby harms society as a whole.

---

56 Professor Frances Olsen has noted that “Western feminists are accused of falsely assuming that all women share important interests and desires.” Frances E. Olsen, Feminism in Central and Eastern Europe: Risks and Possibilities of American Engagement, 106 Yale L.J. 2215, 2228 (1997).

57 “A metanarrative is a story that wants to be more than just a story, that is to say, one which claims to have achieved an omniscient standpoint above and beyond all the stories that people have told before.” UNDERSTANDING CONTEMPORARY SOCIETY: THEORIES OF THE PRESENT 28 (Gary K. Browning, Abigail Hatic, & Frank Webster eds., 2000).

Identity politics also distorts public forum debate by misrepresenting public sentiment. Under an identity politics regime, individual viewpoints count only to the extent that they can be categorized as variants of the recognized meta-ideologies. Individuals who express alternative or third-party viewpoints are not only misunderstood, they are miscounted as supporters of their identity group’s meta-ideology. Because we collapse individual ideologies into identity group meta-ideologies, we misunderstand public sentiment and make poor policy choices. This too harms society as a whole.

Take, as an example of these phenomena, the debate over same-sex marriage. In the late 1990s, when the gay-rights movement galvanized to fight for the right to civil marriage,59 gays and lesbians entered the public forum as presumptive supporters of same-sex marriage. While they could perhaps overcome the presumption by declaring themselves opponents of same-sex marriage—that is, by directly contradicting their identity group’s recognized meta-ideology—they could not overcome the presumption by articulating an alternative or third-party ideology—for example, the abolition of civil marriage. All alternative or third-party ideologies were reduced to variants of the pro- or anti-same-sex marriage meta-ideologies. Individuals that espoused such ideologies were either not heard or not understood, and always miscounted.

Identity politics is, on balance, a harmful institution. Its benefits to identity groups are far outweighed by its detriments to identity group members and society as a whole. A regime that deprives individuals of autonomy and society of the ability to function as a true democracy ought not to be tolerated by an enlightened polity.

III. THE SOURCE OF THE PROBLEM: EQUAL PROTECTION DOCTRINE

Having established that identity politics is harmful, this Article investigates in Part III why such a harmful institution initially developed and so robustly persists. Part III(A) illustrates that our past embrace of subordinating social institutions like racism, sexism, and homophobia produced the present proliferation of identity groups. Part III(B) illustrates that our constitutional commitment to equality, expressed in the Equal Protection and Citizenship Clauses, was originally intended to address not only the subordination experienced by identity groups, but also the subordination experienced by other groups—perhaps including wealth-, age-, or religion-based groups. Part III(C) reveals that, contrary to original intent, our commitment to equality has been deployed through a doctrine that revolves around identity groups and is ill-equipped to address subordination arising from any other source. Part III(D) argues that, as long as equal protection doctrine centers on identity groups, it will perpetuate identity politics.

59 Arthur S. Leonard, Going for the Brass Ring: The Case for Same-Sex Marriage, 82 CORNELL L. REV. 572, 592 (1997) (explaining why “the campaign to win the right to marriage for same-sex couples . . . emerged as the key issue for the lesbian and gay rights movement in the 1990s”).
A. Our History of Subordination

Part III(A) identifies three social conditions necessary for identity group formation: (1) an awareness of difference, (2) an awareness of inequality, and (3) an awareness of subordination. It also posits that the prevalence of these conditions throughout American history has produced the present proliferation of identity groups.

Because identity groups exist only relationally, an “awareness of difference” is the first condition of their formation.  

Young uses Native American tribes to illustrate this point. She explains that certain tribes began to conceive of themselves as something akin to identity groups only when they began to associate with other tribes: “As long as they associated solely among themselves . . . [tribe members] thought of themselves only as ‘the people.’ The encounter with other [tribes] created an awareness of difference; the others were named as a group, and the first group came to see themselves as a group.” In modern society, some of the anatomical and quasi-anatomical differences that provide the bases for phenomenal groups have come to provide the bases for identity groups. An awareness of these differences is the first condition of identity group formation.

Once an awareness of difference has been established, an awareness of inequality may arise. Any phenomenal group may experience difference, but only one that also experiences inequality may evolve into an identity group. Contrast, for example, the phenomenal groups of blue- and green-eyed people, which experience difference but not inequality, with the phenomenal groups of light- and dark-skinned people, which experience both difference and inequality. While eye-color is of minimal cultural significance, the concentration of melanocytes in skin is of deep cultural significance. Thus, while none of the phenomenal groups defined by eye color have evolved into identity groups in our society, some of the phenomenal groups defined by skin color have. An awareness of inequality is, thus, the second condition of identity group formation.

Notice that the first condition, the awareness of difference, is a necessary prerequisite to the second condition, the awareness of inequality. Notice also that certain differences are more conducive to creating the awareness of inequality. Visible differences are most conducive to creating the awareness of inequality, and such differences may arise from either biological or cultural sources. One visible difference arising from a biological source to provide a platform for inequality is skin tone; one visible difference arising from a cultural source to provide a platform for inequality is sex-based dress and grooming standards. Whether they arise from biological or cultural sources, however, visible differences are always capable of providing extremely robust platforms for inequality.

60 Young, supra note 32, at 43.
61 Id.
62 Id.
63 There is, for example, a substantial body of scholarship discussing the importance of visible differences to maintaining gender hierarchies. See Paulette Caldwell, A Hair
Collective awareness of difference and inequality is necessary but not sufficient by itself to trigger identity group formation. Not all phenomenal groups that experience difference and inequality evolve into identity groups. Inequality creates both privileged and subordinated groups, and only the latter may evolve into identity groups. Although both light-skinned and dark-skinned groups experience difference and inequality, for example, it is only dark-skinned groups—which uniquely experience subordination—that have evolved into identity groups. Similarly, although both males and females experience difference and inequality, it is only females—who uniquely experience subordination—who evolved into an identity group. An awareness of subordination, then, is the third and final condition of identity group formation.

When a phenomenal group becomes aware of its difference, inequality, and subordination, it will very likely evolve into an identity group. A phenomenal group can be characterized as an identity group when, as previously stated, its members’ shared anatomical or quasi-anatomical attribute becomes associated with

---

**Piece: Perspectives on the Intersection of Race and Gender**, 1991 DUKE L.J. 365, 393 (explaining that “[j]udgments about aesthetics do not exist apart from judgments about the social, political, and economic order of a society”); Mary Whisner, Gender-Specific Clothing Regulation: A Study in Patriarchy, 5 HARV. WOMEN'S L.J. 73, 73 (1982) (explaining that “appearance conveys a multitude of messages about class, occupation, race, physical freedom, . . . and gender”).

64 For purposes of clarity within this Article, I have defined identity groups as limited to subordinated groups. This is not, however, a universally accepted definition.

65 This is, of course, a gross oversimplification of racial identity groups and is merely intended to illustrate a broad point about group formation.

66 Although my focus in this project is on the conditions of identity group formation, it should be noted that the same conditions can also cause other types of groups to form. Young identifies five “faces” of oppression (which can be viewed as five forms of subordination): (1) exploitation, (2) marginalization, (3) powerlessness, (4) cultural imperialism, and (5) violence. YOUNG, supra note 32, at 40. Different faces of oppression lead to the formation of different types of groups.

Exploitation, for example, which involves inequality in “the transfer of the results of the labor of one social group to benefit another [social group.]” might lead to the formation of class groups. *Id.* at 49. The labor movement of the early twentieth century provides an example of exploitative oppression leading to the formation of a class group. Class groups are distinct from phenomenal, ideological, and identity groups. Although they may share some characteristics of Young’s social groups and Margalit and Raz’s encompassing groups, they do not arise from a phenomenal group and therefore cannot be classified as an identity group.

Contrasted with exploitation, cultural imperialism, which involves inequality in “the universalization of a dominant group’s experience and culture, and its establishment as the norm[,]” might lead to the formation of identity groups. *Id.* at 59. The gay-rights movement of the late twentieth and early twenty-first centuries provides an example of cultural-imperialist oppression leading to the formation of an identity group. Inasmuch as a single group may experience multiple forms of oppression, the relationship between the face of oppression and the type of group formed is complex.

67 *See supra* Part II.C.2.
subordinating cultural norms that dictate their non-anatomical attributes. Even the most cursory review of American history reveals that numerous phenomenal groups have been subjected to experiences of difference, inequality, and subordination. As a result, modern American society is populated by a multitude of identity groups, including African Americans, women, people with disabilities, and gays.\textsuperscript{68}

Although the above-listed identity groups are all presently political actors, they were not so initially. Rather, these identity groups came into being due to their lack of political power. Although identity groups are created by social institutions, it is legal institutions—such as, in the American case, equal protection doctrine—that convert them into political actors. The following two sections, therefore, respectively describe our original commitment to equality and its deployment through an equal protection doctrine that has transformed numerous latent identity groups into political actors.

\textbf{B. Our Constitutional Commitment to Equality}

We have attempted to dismantle a number of the subordinating social institutions that produce identity groups by making a constitutional commitment to equality. This commitment, expressed in the Equal Protection and Citizenship Clauses, may have originally been intended to address not only subordination experienced by identity groups but also subordination experienced by an array of other types of groups—perhaps including wealth- or religion-based groups.\textsuperscript{69}

The Equal Protection Clause mandates that “[n]o State shall . . . deny to any person within its jurisdiction the equal protection of the laws”;\textsuperscript{70} the Citizenship Clause declares that “[a]ll persons born . . . in the United States . . . are citizens of the United States. . . .”\textsuperscript{71} The Thirty-ninth Congress, sitting from 1865 through 1867,\textsuperscript{72} proposed these clauses with a variety of intents: some, such as abolishing the Black Codes, were highly specific; others, such as instituting an equal-birth principle, were quite general. Different intents, of course, suggest different doctrines. The specific intent to abolish the Black Codes, as will be illustrated by Part III(B)(1), suggests a doctrine centered on identity groups; the general intent to institute an equal-birth principle, as will be illustrated by Part III(B)(2), suggests a doctrine operating outside the narrow context of identity groups. I will later argue that to fulfill our original commitment to equality, we need an equal protection doctrine capable of furthering the Thirty-ninth Congress’s specific as well as general intents.

\textsuperscript{68} See supra Part II.C.2.
\textsuperscript{69} See, e.g., AKHIL REED AMAR, AMERICA’S CONSTITUTION: A BIOGRAPHY 381–82 (2005).
\textsuperscript{70} U.S. CONST. amend. XIV, § 1.
\textsuperscript{71} Id.
\textsuperscript{72} The Thirty-ninth Congress met during the second Lincoln administration. WILLIAM HORATIO BARNES, HISTORY OF THE THIRTY-NINTH CONGRESS OF THE UNITED STATES 13 (1868).
1. Abolishing the Black Codes

The Thirty-ninth Congress undeniably intended the Equal Protection Clause to abolish the so-called “Black Laws” or “Black Codes.” In 1865, the Thirteenth Amendment’s prohibition of slavery prompted southern states to enact laws that reproduced the conditions of slavery by “singling out blacks for unequal treatment in matters of labor, land ownership, criminal penalties, and so on.” In response to these Black Codes, the Thirty-ninth Congress’s Joint Committee on Reconstruction drafted the Fourteenth Amendment. The amendment was proposed to the states in 1866 and ratified in 1868. As many scholars have observed, the Black Codes “defined the paradigm case of impermissible legislation” under the Fourteenth Amendment.

Abolishing the Black Codes required an equal protection doctrine extending, at the very least, to protect the African Americans of the 1860s South. These African Americans might be characterized as our nation’s paradigm identity group—they were a phenomenal group defined by an anatomical attribute (skin color) that had become associated with subordinating cultural norms dictating many aspects of their personalities, behaviors, and ideologies. While abolishing the Black Codes required a doctrine capable of protecting at least that paradigm identity group of 1860s African Americans, it suggested a doctrine capable of protecting multiple identity groups.

73 See Jed Rubenfeld, Revolution by Judiciary: The Structure of American Constitutional Law 41 (2005) (“The Fourteenth Amendment was enacted, in the most concrete, historically determinate sense, to abolish the black codes.”); see also Michael C. Dorf, Equal Protection Incorporation, 88 Va. L. Rev. 951, 958 (2002) (“There is general agreement that the central, original purpose of the Equal Protection Clause, indeed of the entire Fourteenth Amendment, was to protect African Americans against the Black Codes.”).

74 Rubenfeld, supra note 73, at 41; see also, e.g., Slaughter-House Cases, 83 U.S. 36, 70 (1872) (“[African-Americans] were in some States forbidden to appear in the towns in any other character than menial servants. They were required to reside on and cultivate the soil without the right to purchase or own it. They were excluded from many occupations of gain, and were not permitted to give testimony in the courts in any case where a white man was a party.”).

75 See Richard A. Primus, The Riddle of Hiram Revels, 119 Harv. L. Rev. 1680, 1711 n.121.


77 Amar, supra note 69, at 384 (“The Black Codes . . . defined the paradigm case of impermissible legislation.”); see also Rubenfeld, supra note 73, at 42 (describing “the abolition of the black codes” as “the foundational paradigm case” guiding the interpretation of the Fourteenth Amendment).
2. Instituting an Equal-Birth Principle

While the Thirty-ninth Congress undeniably intended the Equal Protection Clause to abolish the Black Codes, it arguably also intended the clause to institute what Akhil Amar has described as “an equal-birth principle.” Such a principle might “repudiate a multitude of inequalities beyond Black Codes and race laws,” including not only inequalities experienced by other identity groups, but also inequalities operating outside the context of identity groups.

The intent to institute an equal-birth principle becomes evident through a holistic reading of the Fourteenth Amendment, considering the Equal Protection and Citizenship Clauses together. Both clauses mandate equality—the former, explicitly; the latter, implicitly. As Professor Amar explains, the Citizenship Clause “aimed to . . . mak[e] clear that everyone born under the American flag—black or white, rich or poor, male or female, Jew or Gentile—was a free and equal citizen.”

Taken together, the two clauses suggest an equal-birth principle, described by Professor Amar in the following terms:

[T]he grand idea that humans were born free and equal opened itself to broad[] interpretations—some plainly invited by Reconstruction Republicans, others less clearly foreseen yet nonetheless textually permissible. Laws . . . that heaped disabilities on anyone born a Jew or born female, or that gave special privileges to scions of the wealthy—all such legislation could plausibly be seen as violative of the equal-birth principle.

Our original commitment to equality, then, was intended to address inequities that went beyond the narrow context of identity groups.

To summarize, the Equal Protection Clause was proposed with a variety of intents that each required different doctrinal capabilities. Accomplishing the specific intent of abolishing the Black Codes required a doctrine capable of

---

78 AMAR, supra note 69, at 384.
79 Id. at 383 (The equal-birth principle suggested by the Equal Protection and Citizenship Clauses, thus, might “repudiate a multitude of inequalities beyond Black Codes and race laws.”). One example of an inequality beyond Black Codes and race laws that the framers of the Fourteenth Amendment might have intended to repudiate is sex-based inequality. As Professor Amar explains: “Whereas the Founding text used the word ‘men’ in describing the principle of birthright equality, its Reconstruction descendant did not—and for good reason. Far more than is generally recognized today, the framers of the Reconstruction Amendments focused not merely on the race issue but also on intersecting issues of gender.” Id. at 385.
80 Id. at 381–82 (“Though the word ‘equal’ did not explicitly appear in the Fourteenth Amendment’s first sentence, the concept was strongly implicit. . . . [Indeed,] a later Supreme Court case glossed the citizenship clause as follows: ‘All citizens are equal before the law.’”).
81 Id. at 384.
protecting identity groups, whereas accomplishing the general intent to institute an equal-birth principle required a doctrine capable of responding to inequities operating outside the narrow context of identity groups. Fulfilling all of the framers’ intents would clearly require a flexible doctrine, of the sort that will be proposed in Part IV(B)—but, as will be illustrated presently, the existing doctrine is strikingly inflexible.

C. The Supreme Court’s Equal Protection Doctrine

Part III(C) illustrates that, although the Equal Protection Clause was proposed with a variety of intents, it has been interpreted almost exclusively by reference to only one of them. The Court has focused on the specific intent to abolish the Black Codes and other similarly subordinating institutions, while consistently declining to recognize the general intent to institute an equal-birth principle. Its present doctrine is, therefore, relatively well-equipped to protect identity groups but markedly ill-equipped to address any inequities operating outside the context of identity groups. It does not, for example, adequately protect against inequities that affect wealth- or religion-based groups. So long as equal protection doctrine continues to address only this narrow set of inequities, our original commitment to equality will remain largely unfulfilled.

The history of equal protection doctrine might be described as a series of generalizations from the Thirty-ninth Congress’s specific intent to abolish the Black Codes. This subpart traces the doctrine through three overlapping phases. In the first phase, from Reconstruction until World War II, the doctrine protected only the paradigm identity group of African Americans.82 In the second phase, from World War II until the late twentieth century, the doctrine expanded to protect multiple identity groups.83 And in the third phase, from the late twentieth century through the present, the doctrine further expanded to protect individuals from privileged groups (e.g., white men) who were disadvantaged by their exclusion from identity groups (e.g., through affirmative action programs).84

As will become evident, each of the three phases represents a practical, but not conceptual, move. From a practical perspective, the doctrine has expanded, protecting more and more identity groups in each successive phase. From a conceptual perspective, however, the doctrine has remained static, always structured entirely around identity groups and never willing to go beyond them.

1. Phase One—Protecting One Identity Group

The first phase, from Reconstruction until World War II, reveals a doctrine so rigidly constrained by the abolition of the Black Codes that it protected only the paradigm identity group of African Americans. The Slaughter-House Cases,

82 See infra Part III.C.1.
83 See infra Part III.C.2.
84 See infra Part III.C.3.
decided in 1872—only four years after the Fourteenth Amendment’s ratification—provided the Court with its first opportunity to interpret the Equal Protection Clause. At that early date, the justices could not even imagine a doctrine that protected any group other than African Americans.

Justice Miller, writing for the majority, spoke at length of the intent behind the Reconstruction Amendments. He described their “pervading purpose” as securing “the freedom of the slave race, the security and firm establishment of that freedom, and the protection of the newly-made freeman and citizen from the oppressions of those who had formerly exercised unlimited dominion over him.” He justified the Court’s focus on African Americans, who were explicitly mentioned in only one of the three Reconstruction Amendments, as follows: “It is true that only the fifteenth amendment, in terms, mentions the negro by speaking of his color and his slavery. But it is just as true that each of the other articles was addressed to the grievances of that race, and designed to remedy them as the fifteenth.” All three Reconstruction Amendments were thus limited to furthering the specific intent of protecting African Americans.

After discussing the intent behind the Reconstruction Amendments as a whole, Justice Miller turned to the intent behind the Equal Protection Clause in particular. He stated, “The existence of laws in the States where the newly emancipated negroes resided, which discriminated with gross injustice and hardship against them as a class, was the evil to be remedied by this clause, and by it such laws are forbidden.” The Equal Protection Clause was thus, like the Reconstruction Amendments as a whole, limited to protecting African Americans. Justice Miller, indeed, opined that the clause would probably never be construed more broadly:

We doubt very much whether any action of a State not directed by way of discrimination against the negroes as a class, or on account of their race, will ever be held to come within the purview of this provision. It is so clearly a provision for that race and that emergency, that a strong case would be necessary for its application to any other.

The Slaughter-House Cases, then, suggest that the Equal Protection Clause was permanently limited to protecting African Americans.

This exceedingly narrow initial interpretation of the Equal Protection Clause undoubtedly resulted from the factual background of the Slaughter-House Cases.

85 Slaughter-House Cases, 83 U.S. 36, 70 (1872).
86 Id. at 68–72.
87 Id. at 71.
88 Id. at 71–72.
89 Id. at 81.
90 Id.
91 See also Strauder v. West Virginia, 100 U.S. 303, 306 (1879) (stating the Fourteenth Amendment “was designed to assure to the colored race the enjoyment of all the civil rights that under the law are enjoyed by white persons . . . ”).
A group of butchers from New Orleans had invoked the Equal Protection Clause (along with a number of other Reconstruction provisions) to block the creation of a municipal monopoly over the livestock-landing and slaughterhouse business. The butchers had hoped that the Court would view the monopoly, which economically disadvantaged them, as a deprivation of equal protection. It was in response to their claim that the Court articulated this exceedingly narrow interpretation of the Equal Protection Clause.

Had a different group—that more closely resembled African Americans—litigated the first equal protection case, the Court might have responded differently. Consider the difference between African Americans and butchers in terms of our earlier discussion of groups. African Americans, marked by their “color and [history of] slavery,” clearly constitute an identity group. Butchers, however, united only by their shared practice of butchering (which is “external” to identity), would more closely fit Young’s definition of an association. Had an identity group, rather than an association, litigated the first equal protection case, the Court might have been less hostile. Thus, even this first equal protection decision can, with some elaboration, be read as establishing a doctrine structured around identity groups.

Obviously, the narrow interpretation of the Equal Protection Clause as limited to protecting African Americans no longer governs. It does, however, continue to motivate the present doctrine. As Jed Rubenfeld has observed, “The foundational application of the equal protection clause—its abolition of the black codes—plays a special, structuring doctrinal role, serving as a paradigm case and anchoring everything the Court has ever said about ‘suspect classes,’ the organizing concept for all modern constitutional antidiscrimination doctrine.” The Thirty-ninth Congress’s specific intent to abolish the Black Codes, thus, continues to constrain equal protection doctrine.

2. Phase Two—Protecting Multiple Identity Groups

In the second phase, from World War II until the late twentieth century, equal protection doctrine expanded to protect a variety of identity groups. Using African Americans as the paradigm, the Court developed a set of criteria for identifying other identity groups—or, in the Court’s terminology, other “suspect classes”—that similarly warranted special protection. By the end of the second phase, the Court had recognized a number of racial identity groups (e.g., Japanese

---

92 Slaughter-House Cases, 83 U.S. at 59, 66.
93 Id. at 66.
94 Id. at 72.
95 See supra Part III.A.1.
96 RUBENFELD, supra note 73, at 16.
97 See infra Part III.C.2.a.
Americans) as “suspect classes”\textsuperscript{98} and a number of nonracial identity groups (e.g.,
women) as “quasi-suspect classes.”\textsuperscript{99}

The Court justified the protection of these additional identity groups through a
new rhetoric of intent. Where it had previously expounded on the narrow intent to
abolish the Black Codes, the Court now expounded on a broader intent to abolish
racial discrimination.\textsuperscript{100} This rhetorical move was complete by the time the Court
decided \textit{McLaughlin v. Florida}, in which it described “the central purpose of the
Fourteenth Amendment” as the “elimination of racial discrimination emanating
from official sources in the States.”\textsuperscript{101} It was further solidified by the time the
Court decided \textit{Palmore v. Sidoti}, in which it described a “core purpose” of the
Fourteenth Amendment as “doing away with all governmentally imposed
discrimination based on race.”\textsuperscript{102} The rhetorical move from the Black Codes to
racial discrimination justified the parallel practical move from protecting African
Americans to protecting a variety of racial identity groups.

These rhetorical and practical moves were not, however, accompanied by a
conceptual move. Regardless of whether equal protection doctrine protected only a
single identity group, as in the first phase, or multiple identity groups, as in the
second, it remained firmly organized around identity groups. It, thus, evolved
rhetorically and practically, but not conceptually. Because this second phase is
quite lengthy, constituting a large part of the twentieth century, its discussion is
divided into three subparts: the first introduces the Court’s suspect class doctrine;
the second and third trace the Court’s application of that doctrine to racial and
nonracial identity groups, respectively.

\textbf{(a) African Americans as the Paradigm Suspect Class}

This subsection introduces the Court’s suspect class doctrine, which is
generally understood as originating in the 1938 decision of \textit{United States v.
Carolene Products}.\textsuperscript{103} In the now famous footnote 4 of that decision, the Court
explained that “prejudice against discrete and insular minorities may be a special
condition, which tends seriously to curtail the operation of those political processes

\textsuperscript{98} See infra Part III.C.2.b.
\textsuperscript{99} See infra Part III.C.2.c.
\textsuperscript{100} Note, however, that the Court did not go so far as to acknowledge the even more
general intent to institute an equal-birth principle. That would, indeed, have taken the
doctrine far beyond what the Court was comfortable with.
\textsuperscript{101} 379 U.S. 184, 191–92 (1964) (“This strong policy renders racial classifications
‘constitutionally suspect,’ subject to the ‘most rigid scrutiny,’ and ‘in most circumstances
irrelevant’ to any constitutionally acceptable legislative purpose.”). The Court in
\textit{McLaughlin} deployed this “central purpose” to justify the invalidation of a law prohibiting
unmarried cohabitation by “[a]ny negro man and white woman, or any white man and
negro woman.” \textit{Id.} at 184.
\textsuperscript{102} 466 U.S. 429, 432 (1984). The Court in \textit{Palmore} deployed this “core purpose” to
justify the reversal of a racially motivated custody award. \textit{Id.} at 433.
\textsuperscript{103} 304 U.S. 144 (1938).
ordinarily to be relied upon to protect minorities, and which may call for a correspondingly more searching judicial inquiry.”

Although *Carolene Products* was a due process case, the Court has often cited it in equal protection cases dealing with suspect class doctrine. Indeed, Justice Blackmun once, in a concurring opinion, described *Carolene Products* as “the moment the Court began constructing modern equal protection doctrine.”

“Suspect classes,” roughly defined, are groups that exhibit three criteria: (1) past subordination, (2) an immutable attribute, and (3) present political powerlessness. The Court has described suspect classes in various terms. In a 1973 decision, it explained that suspect classes are “saddled with such disabilities, or subjected to such a history of purposeful unequal treatment, or relegated to such a position of political powerlessness as to command extraordinary protection from the majoritarian political process.” In a 1987 decision, it explained that suspect classes, “[a]s a historical matter, . . . have . . . been subjected to discrimination; . . . exhibit obvious, immutable, or distinguishing characteristics that define them as . . . discrete group[s]; and . . . [are] . . . minorit[jes] or politically powerless.”

All three suspect class criteria clearly derive from the African American experience. The “history of subordination” criterion derives from African Americans’ subjection to slavery and then the Black Codes; the “immutable attribute” criterion derives from African Americans’ distinctive skin tone; and the “present political subordination” criterion derives from African Americans’ persistent political powerlessness. Although in the first phase African Americans were the only suspect class, during the second phase they became the paradigm suspect class. The Court, that is, recognized other identity groups as suspect classes only if it perceived them as sufficiently similar to the African American paradigm suspect class.

Being a member of a suspect class was, during this second phase, highly significant—indeed, it seemed to be one of only two avenues to strict scrutiny. The Court ostensibly applied strict scrutiny only to classifications that either “operate[d] to the peculiar disadvantage of a suspect class” or “impermissibly interfere[d] with the exercise of a fundamental right.” There was, however, some doctrinal ambiguity surrounding the application of strict scrutiny to the former classifications. The Court, at times, stated strict scrutiny applied because the plaintiff belonged to a “suspect class”—but, at other times, suggested that strict scrutiny applied because the plaintiff was disadvantaged by a “suspect

---

104 *Id.* at 152–53 n.4.
105 See, e.g., *Graham v. Richardson*, 403 U.S. 365, 372 (1971) (citing *Carolene Products* in support of the proposition that “[a]liens as a class are a prime example of a ‘discrete and insular’ minority for whom such heightened judicial solicitude is appropriate”).
110 *Id.*
classification."\textsuperscript{111} Even within individual opinions, the Court often slid between criticizing statutes because they “target[ed] suspect class[es]” and “involv[ed] suspect classifications.”\textsuperscript{112} Although this ambiguity was of little significance during the second phase (when most plaintiffs were both from suspect classes and disadvantaged by suspect classifications), it took on great significance during the third phase (when many plaintiffs came from privileged classes).

Regardless of whether the Court was dealing with a suspect class or classification, it justified the application of strict scrutiny—which was generally, during this period, both “strict in theory and fatal in fact”\textsuperscript{113}—as follows:

[Suspect classifications] are so seldom relevant to the achievement of any legitimate state interest that laws grounded in such considerations are deemed to reflect prejudice and antipathy—a view that those in the burdened class are not as worthy or deserving as others. For these reasons and because such discrimination is unlikely to be soon rectified by legislative means, these laws are subjected to strict scrutiny and will be sustained only if they are suitably tailored to serve a compelling state interest.\textsuperscript{114}

Such rhetoric, of course, implicitly encouraged all plaintiffs to present themselves to the Court as members of suspect classes. More generally, it encouraged all citizens to view themselves—and the world—in terms of suspect classes (or in my terminology, identity groups).

Suspect class doctrine is, of course, firmly organized around identity groups. The first suspect class criterion of “past subordination” clearly describes identity groups—subordination, as previously stated, is a necessary condition of identity group formation\textsuperscript{115}. The second suspect class criterion of “an immutable trait” also describes identity groups—identity group members, as previously stated, share a functionally immutable anatomical or quasi-anatomical attribute.\textsuperscript{116} And the third suspect class criterion of “present political subordination” also describes identity groups—subordinating cultural norms, as previously stated, deprive identity group

\textsuperscript{111} See, e.g., City of Richmond v. J.A. Croson Co., 488 U.S. 469, 493–94 (1989) (stating that “[c]lassifications based on race carry a danger of stigmatic harm” and “that the standard of review under the Equal Protection Clause is not dependent on the race of those burdened or benefited by a particular classification”).

\textsuperscript{112} See, e.g., Vacco v. Quill, 521 U.S. 793, 799 (1997).


\textsuperscript{114} City of Cleburne v. Cleburne Living Ctr., 473 U.S. 432, 440 (1985) (applying rational basis review but holding that the City of Cleburne violated the Equal Protection Clause when it required Cleburne Living Center to obtain a special use permit prior to operating a group home for individuals with mental disabilities).

\textsuperscript{115} See supra Part III.A.

\textsuperscript{116} See supra Part II.C.
members of political power.\textsuperscript{[117]} The terms “suspect class” and “identity group” are, thus, essentially interchangeable.

It should be noted that, although suspect class membership is not the only avenue to strict scrutiny under the Equal Protection Clause, it seems to be the primary avenue. While plaintiffs alleging interferences with fundamental rights could theoretically trigger strict scrutiny under the Equal Protection Clause,\textsuperscript{[118]} such allegations most often seem to trigger strict scrutiny under the Due Process Clause. While the results under either clause are from a practical perspective roughly equivalent,\textsuperscript{[119]} they are from a conceptual perspective quite different. Equal protection claims influence our concept of equality and, when equal protection claims are limited to suspect class claims, our concept of equality is limited to equality among suspect classes.

The availability of relief for fundamental rights claims outside the equal protection context is irrelevant to our concept of equality. Neither the frequent recognition of fundamental rights claims under the Due Process Clause, nor the rare recognition of such claims under the Equal Protection Clause enables us to end our obsession with suspect classes. Only an equal protection doctrine that regularly addressed a variety of claims operating outside the narrow context of suspect classes—including, but not limited to fundamental rights claims—would enable us to end our obsession with suspect classes and ultimately transcend identity politics.

\textit{(b) Other Races as Suspect Classes}

Understanding African Americans as the paradigm suspect class is crucial to understanding the Court’s categorization of other classes as suspect or not suspect. Once the African American identity group became the paradigm suspect class, other racial identity groups were easily recognized as similarly suspect classes.

The term “race” was, during World War II, used quite broadly to include what we now refer to as national origin and alienage.\textsuperscript{[120]} The Court, therefore, spoke of

\begin{itemize}
  \item[117] See supra Part II.D.1.
  \item[118] See, e.g., Shapiro v. Thompson, 394 U.S. 618, 634 (1969) (applying strict scrutiny under the Equal Protection Clause to invalidate a statute that interfered with the right to interstate travel); Skinner v. Oklahoma, 316 U.S. 535, 541 (1942) (applying strict scrutiny under the Equal Protection Clause to invalidate a statute that interfered with the exercise of the fundamental right to procreate).
  \item[119] The results are equivalent inasmuch as all interferences with fundamental rights trigger strict scrutiny—regardless of the clause under which they are analyzed. The Court has, however, protected slightly different sets of “fundamental” rights under the Equal Protection and Due Process Clauses. ERWIN CHEMERINSKY, CONSTITUTIONAL LAW: PRINCIPLES AND POLICIES 793 (2006) (stating that the right to refuse medical care has only been recognized under the Due Process Clause, and the right to travel only under the Equal Protection Clause).
“race” as it recognized both Japanese Americans\textsuperscript{121} and Japanese immigrants\textsuperscript{122} as suspect classes. In \textit{Korematsu v. United States}, for example, the Court recognized the Japanese American “race” as a suspect class, stating that “all legal restrictions which curtail the civil rights of a single racial group are immediately suspect.”\textsuperscript{123} Although the acquisition of suspect class status did not translate into any practical protection for Japanese Americans in \textit{Korematsu}, the decision was nevertheless significant in illustrating the Court’s willingness to recognize new suspect classes.

There were, however, limits to the Court’s willingness to recognize new suspect classes. While the Court extended the set of suspect classes beyond African Americans to other \textit{racial} identity groups, it was, and presently remains, unwilling to extend suspect class status to \textit{nonracial} identity groups such as women and gays.\textsuperscript{124} In the 1948 case of \textit{Goesaert v. Cleary}, for example, the Court declined to recognize women as a suspect class.\textsuperscript{125} There, a female plaintiff’s challenge to a sex-based classification—prohibiting virtually all women from bartending—was subjected to only rational basis review.\textsuperscript{126} The Court reasoned:

The Constitution in enjoining the equal protection of the laws upon States precludes irrational discrimination as between persons or groups

\begin{itemize}
\item \textsuperscript{121} Korematsu v. United States, 323 U.S. 214, 223–24 (1944) (suggesting nationality as a basis for suspect status).
\item \textsuperscript{122} Takahashi v. Fish & Game Comm’n, 334 U.S. 410, 420 (1948) (suggesting alienage is suspect). \textit{But see} Foley v. Connellie, 435 U.S. 291, 294 (1978) (stating that the Court has not “held that all limitations on aliens are suspect”).
\item \textsuperscript{123} 323 U.S. at 216 (stating that such classifications are also subject “to the most rigid scrutiny”).
\item \textsuperscript{124} As will be discussed later, women have subsequently been recognized as a quasi-suspect class. \textit{See} Craig v. Boren, 429 U.S. 190, 197 (1976). Gays have not been recognized as a suspect class, but some argue they have received a more searching form of rational basis review. \textit{See}, e.g., Jeremy B. Smith, \textit{Note, The Flaws of Rational Basis with Bite: Why the Supreme Court Should Acknowledge Its Application of Heightened Scrutiny to Classifications Based on Sexual Orientation}, 73 \textit{Fordham L. Rev.} 2769, 2770 (2005) (arguing that the Court should acknowledge its application of “a more searching form of rational basis review” to sexual orientation classifications in both \textit{Romer v. Evans} and \textit{Lawrence v. Texas}).
\item \textsuperscript{125} 335 U.S. 464, 465–67 (1948) (upholding a law denying bartenders licenses to women who were not the wives or daughters of proprietors).
\item \textsuperscript{126} \textit{Id.} at 466.
\end{itemize}
of persons in the incidence of a law. But the Constitution does not require situations “which are different in fact or opinion to be treated in law as though they were the same.”

While the Court was willing to equate African Americans with other racial identity groups, it was not willing to insert women into the equation. Women—and, indeed, all nonracial identity groups—were perceived as too differently situated to merit equal protection.

During most of the second phase, the Court treated racial classifications as suspect and nonracial classifications as innocuous. Kenji Yoshino has gone so far as to describe the Court’s treatment of race as “talismanic.” He has asserted, “In considering arguments that other classifications be accorded heightened scrutiny, the courts have required claimants to demonstrate the similarities these classifications share with race.” This requirement, as Professor Yoshino has illustrated, encourages all identity groups to present themselves as similar to racial identity groups. William Eskridge has observed, “Once race was established as a suspect classification, the women’s, lesbian and gay, and disabled people’s movements all sought to establish their defining traits as similarly suspect.”

To date, the Court has recognized a variety of racial identity groups as suspect classes, while emphatically declining to recognize any nonracial identity groups as suspect classes. It has, however, recognized certain nonracial identity groups as quasi-suspect classes.

(c) Nonraces as Quasi-Suspect Classes

Beginning in the early 1970s, the Court further broadened its interpretation of the Equal Protection Clause. Still using African Americans as the paradigm suspect class, the Court recognized certain nonracial identity groups, like women, as quasi-suspect classes. Although this broader interpretation might have been supported by a broader rhetoric of intent, it was not. Rather than making another generalization—from the intent to abolish racial discrimination to the intent to

---

127 Id.
129 Id. Although Yoshino discusses race and sex as talismanic classifications, I will address sex separately later.
130 Id. (“In considering arguments that . . . classifications [other than race and sex] be accorded heightened scrutiny, the courts have required claimants to demonstrate the similarities these classifications share with race and sex. Commonalities between the two paradigm classifications thus play a powerful gatekeeping role.”).
abolish all identity-group discrimination—and rather than recognizing the forgotten intent to institute an equal-birth principle, the Court simply neglected to discuss intent. It seemed to accept that equal protection doctrine had departed from, and could no longer be justified by, any of the original intents of the Thirty-ninth Congress.

Thus, without any claim that it was promoting original intent, the Court extended protection to certain nonracial identity groups. Its move from protecting racial to protecting nonracial identity groups was, like its predecessor, practical but not conceptual. Even as the Court extended protection to a larger set of groups—which ultimately included women and nonmarital children—equal protection doctrine remained firmly organized around the concept of identity groups.

Although women fought long and hard to become the first nonracial identity group recognized as a quasi-suspect class, their path to recognition was anything but linear. In the 1971 decision of Reed v. Reed, the Court for the first time invalidated a sex-based classification on the basis that it violated the Equal Protection Clause. Although the Court in Reed did not recognize women as a suspect class or formally apply any sort of heightened scrutiny, it signaled a new willingness to protect women—and, potentially, other nonracial identity groups—withing the context of the Equal Protection Clause.

Two years later, in Frontiero v. Richardson, a plurality of the Court actually recognized women as a full-fledged suspect class. Justice Brennan, writing for the Frontiero plurality, asserted that “classifications based upon sex, like classifications based upon race, alienage, and national origin, are inherently suspect.” He justified this assertion by illustrating the strong similarities between women and the always paradigmatic African Americans:

[T]hroughout much of the 19th century the position of women in our society was, in many respects, comparable to that of blacks under the pre-Civil War slave codes. Neither slaves nor women could hold office, serve on juries, or bring suit in their own names, and married women traditionally were denied the legal capacity to hold or convey property or to serve as legal guardians of their own children.

133 See Frontiero v. Richardson, 411 U.S. 677 (1973); Reed v. Reed, 404 U.S. 71 (1971).
134 Its predecessor was the move from protecting African Americans to protecting racial identity groups.
136 404 U.S. at 76.
137 Id. (applying rational basis review).
139 Id. at 682.
140 Id. at 685.
Women, then, were protected because of their similarity to African Americans. They, of course, met all three suspect class criteria. That they met the first criterion of “past subordination” is evident from the above-quoted passage. That they met the second criterion of “an immutable attribute” is evident from Justice Brennan’s statement that “sex, like race and national origin, is an immutable characteristic determined solely by the accident of birth.” That they met the third criterion of “present political powerlessness” is evident from Brennan’s observations that “women still face pervasive, although at times more subtle, discrimination in our educational institutions, in the job market and, perhaps most conspicuously, in the political arena”, and “in part because of past discrimination, women are vastly underrepresented in this Nation’s decisionmaking councils.”

Because Justice Brennan’s opinion in *Frontiero* commanded a plurality rather than a majority, women’s status remained uncertain for another three years, until the Court decided *Craig v. Boren*. In *Craig*, a majority finally agreed that sex-based classifications ought to be recognized as quasi-suspect and subjected to intermediate scrutiny, meaning that they would be invalid unless they were “substantially related” to achieving “important governmental objectives.” The decision was, in many ways, a harbinger of the doctrinal changes that demarcated the third phase. While the Court recognized sex-based classifications as quasi-suspect, the plaintiff was not actually a member of a quasi-suspect class. He was, in fact, a man who had been disadvantaged by a sex-based classification. The Court applied heightened scrutiny not because Curtis Craig belonged to an identity group, but because he had been disadvantaged by his lack of identity-group membership.

The Court has, to date, recognized two quasi-suspect classes—women and nonmarital children—while studiously avoiding the recognition of many others. The list of identity groups that have been denied suspect or quasi-suspect class status continually grows. In *City of Cleburne v. Cleburne Living Center*, for example, the Court declined to recognize mental disability as a suspect or quasi-suspect classification, while nevertheless invalidating a zoning ordinance that excluded group homes for people with mental disabilities. And, in both *Romer v.*
Evans\textsuperscript{149} and Lawrence v. Texas,\textsuperscript{150} the Court declined to say whether sexual orientation was a suspect or quasi-suspect classification by invalidating sexual-orientation classifications under rational basis review.\textsuperscript{151} The Court’s reluctance to recognize these additional identity groups as suspect or quasi-suspect classes seems to stem from its inability to equate them with African Americans.

The Court has not only been unwilling to recognize additional identity groups as suspect or quasi-suspect classes, it has also been unwilling to recognize non-identity-groups—such as the poor and the elderly—as suspect or quasi-suspect classes. In San Antonio Independent School District v. Rodriguez, the Court declined to recognize the poor as a suspect or quasi-suspect class.\textsuperscript{152} After considering a system that tied educational funding to local property taxes and thereby disadvantaged schools in impoverished areas, the Court refused to recognize that the system created a wealth-based classification—and made clear that, even if it did, such classifications had never been held suspect.\textsuperscript{153} In Massachusetts Board of Retirement v. Murgia, the Court rejected age as a suspect or quasi-suspect classification, stating:

While the treatment of the aged in this Nation has not been wholly free of discrimination, such persons, unlike, say, those who have been discriminated against on the basis of race or national origin, have not experienced a ‘history of purposeful unequal treatment’ or been subjected to unique disabilities on the basis of stereotyped characteristics not truly indicative of their abilities.\textsuperscript{155}

At present, the Court has recognized only three suspect classifications—race, alienage, and national origin—and only two quasi-suspect classifications—sex and illegitimacy.\textsuperscript{156} It has maintained that strict scrutiny is applied to suspect classifications; intermediate scrutiny, to quasi-suspect classifications; and rational

\textsuperscript{149} 517 U.S. 620 (1996).
\textsuperscript{150} 539 U.S. 558 (2003).
\textsuperscript{151} See Lawrence, 539 U.S. at 578; Romer, 517 U.S. at 632; Maria A. La Vita, When the Honeymoon is Over: How a Federal Court’s Denial of the Spousal Privilege to a Legally Married Same-Sex Couple Can Result in the Incarceration of a Spouse Who Refuses to Adversely Testify, 33 NEW ENG. J. ON CRIM. & CIV. CONFINEMENT 243, 275 (2007).
\textsuperscript{152} 411 U.S. 1, 2 (1972).
\textsuperscript{153} Id. at 22 (holding that “neither of the two distinguishing characteristics of wealth classifications can be found here”).
\textsuperscript{154} Id. at 29 (“[T]his Court has never heretofore held that wealth discrimination alone provides an adequate basis for invoking strict scrutiny.”).
\textsuperscript{155} 427 U.S. 307, 313 (1976).
basis review, to all other classifications. Although most classifications are supposedly relegated to rational basis review, many scholars have observed that the Court appears to apply a heightened form of rational basis review to classifications based on sexual orientation and disability. Indeed, in her concurring opinion in Lawrence, Justice O’Connor asserted that “[w]hen a law exhibits such a desire to harm a politically unpopular group, we have applied a more searching form of rational basis review . . . .”

3. Phase Three—Preventing Identity-Based Groupings

While equal protection doctrine first protected only African Americans and then expanded to protect other individuals disadvantaged by their identity group membership, it has further expanded to protect individuals disadvantaged by their lack of identity group membership. During phase three, any plaintiff, whether he or she is a member of a suspect or quasi-suspect class, can trigger strict scrutiny by demonstrating that he or she has been disadvantaged by a suspect or quasi-suspect classification.

This extension of protection from members of identity groups to individuals excluded from identity groups resembles the earlier extension of protection from members of racial identity groups to members of nonracial identity groups. The Court did not attempt to justify either extension by reference to a new rhetoric of intent. In both cases, it seemed to accept that its equal protection doctrine was simply no longer consistent with original intent. Thus, while the Court might have justified either extension by reference to the general intent of instituting an equal-birth principle, it has not—presumably, because recognizing this broader intent might lead to recognizing a broader set of claims than the Court is currently willing to entertain. Recognizing an equal-birth principle might, indeed, require the Court to entertain claims operating outside the context of identity groups.

The Court’s move from phase two, during which it protected only identity group members, to phase three, where it now protects individuals disadvantaged by

---

157 Cleburne, 473 U.S. at 440–41.
158 Id. at 440–41.
159 Cass R. Sunstein, Foreword: Leaving Things Undecided, 110 HARV. L. REV. 4, 61 (1996) (arguing that Romer, Moreno, and Cleburne establish a new tier of scrutiny in which rational relation review “is used to invalidate badly motivated laws without refining a new kind of scrutiny”).
161 Professor Darren Hutchinson describes this further expansion, noting that “while blacks or women might constitute suspect classes due to their socially disadvantaged statuses, whites and men [now] receive heightened scrutiny when they challenge laws that classify on the basis of race or gender.” Darren Lenard Hutchinson, “Unexplainable on Grounds Other than Race”: The Inversion of Privilege and Subordination in Equal Protection Jurisprudence, 2003 U. ILL. L. REV. 615, 639 (2003).
162 See id. at 638–39.
either identity group membership or lack thereof, has been described by Professor Rubenfeld as follows:

The interpretive framework governing discrimination claims under the equal protection clause used to be organized around “suspect classes.” [At present, it] is plainly organized around “suspect classifications.” Strict scrutiny is triggered not by the disadvantaging of suspect classes, but by the use of a suspect classification (no matter which group is advantaged or disadvantaged).\textsuperscript{163}

Some scholars have made much of this move, asserting that phases two and three “reflect fundamentally different theories of equality.”\textsuperscript{164} They have recognized that “the law of suspect classes is more consistent with the anti-caste or anti-subordination theory of equal protection, while the law of suspect classifications is more consistent with the anti-classification or anti-differentiation theory of equal protection.”\textsuperscript{165} Professor Rubenfeld has even gone so far as to classify the move from suspect classes to suspect classifications as a “paradigm shift.”\textsuperscript{166}

Although this move is undoubtedly significant, it is—like previous moves—practical rather than conceptual. As a practical matter, the Court in phase three protects a larger set of individuals. Both members and nonmembers of identity groups may now trigger strict scrutiny by challenging suspect classifications. As a conceptual matter, however, the Court in phase three continues to structure equal protection doctrine around identity groups. As Professor Rubenfeld has asserted, notwithstanding the fact that equal protection doctrine has become “profoundly nonoriginalist, the Fourteenth Amendment’s core applications continue to anchor and organize the doctrine.”\textsuperscript{167}

We observed the “law of suspect classifications” in our earlier discussion of \textit{Craig v. Boren}, where an individual who was not a member of a quasi-suspect class triggered intermediate scrutiny by challenging a quasi-suspect sex-based classification.\textsuperscript{168} The law of suspect classifications, applied in \textit{Craig} and other sex discrimination cases,\textsuperscript{169} is in phase three applied in numerous race discrimination

\textsuperscript{163} Rubenfeld, supra note 73, at 199.

\textsuperscript{164} Reginald C. Oh, \textit{A Critical Linguistic Analysis of Equal Protection Doctrine: Are Whites a Suspect Class?}, 13 TEMP. POL. & CIV. RTS. L. REV. 583, 588 (2004). See also Hutchinson, supra note 161, at 638 (characterizing the move from suspect classes to suspect classifications as at least a “doctrinal shift”).

\textsuperscript{165} Oh, supra note 164, at 588.

\textsuperscript{166} Rubenfeld, supra note 73, at 199.

\textsuperscript{167} Id. at 16.

\textsuperscript{168} See discussion supra Part III.C.2.c.

\textsuperscript{169} See, e.g., Miss. Univ. for Women v. Hogan, 458 U.S. 718, 724 (1982) (applying intermediate scrutiny to invalidate the women-only admissions policy at a state nursing school).
cases. Although the law of suspect classifications did not develop in an entirely linear manner, it was firmly established by the mid-1990s. In the 1995 decision of \textit{Adarand Constructors, Inc. v. Pena}, a nonminority subcontractor challenged a “government contract offering financial incentives to a prime contractor for hiring disadvantaged [i.e., racial-minority—] subcontractors.” That is, an individual who was not a member of an identity group challenged a suspect classification. The Court justified the application of strict scrutiny by citing previous statements that “[r]acial classifications [are] ‘constitutionally suspect’”—notwithstanding the fact that these statements came from cases involving \textit{suspect class} members challenging suspect classifications.

The Court has, in subsequent cases, repeatedly reaffirmed the law of suspect classifications. In the 2003 decision of \textit{Grutter v. Bollinger}, for example, the Court stated that “[a]ll government racial classifications imposed must be analyzed by a reviewing court under strict scrutiny.” Similarly, in the 2007 decision of \textit{Parents Involved in Community Schools v. Seattle School District Number 1}, the Court stated, “It is well established that when the government distributes burdens or benefits on the basis of individual racial classifications, that action is reviewed under strict scrutiny.” These cases reveal that heightened scrutiny is, during phase three, no longer triggered solely by identity group membership; it may now also be triggered by the lack of such membership.

Like the law of suspect classes that dominated phase two, the law of suspect classifications that dominates phase three is organized entirely around identity groups. While the Court has extended protection to all individuals disadvantaged by identity-based classifications, it has continued to withhold protection from individuals disadvantaged by other types of classifications—for example, wealth-, religion-, and age-based classifications.

All three of the doctrinal regimes described in Part III(C) are variations on a single understanding of our commitment to equality. From the initial protection of individuals disadvantaged by their membership in a single paradigmatic identity group; to the subsequent protection of individuals disadvantaged by their membership in one of several identity groups; to the present protection of individuals disadvantaged by their membership or lack thereof in one of several identity groups—our commitment to equality has been deployed through a doctrine that is centered around identity groups and is ill-equipped to address other sources of inequity. While the doctrine has abolished the Black Codes and a number of


\footnote{171 \textit{Id}. at 216.}

\footnote{172 \textit{Id}. at 216 (quoting \textit{McLaughlin v. Florida}, 379 U.S. 184, 192 (1964)); see also \textit{Korematsu v. United States}, 323 U.S. 214, 216 (1944) (“[A]ll legal restrictions which curtail the civil rights of a single racial group are immediately suspect.”).}

\footnote{173 \textit{Grutter}, 539 U.S. at 308.}

\footnote{174 \textit{Parents Involved}, 551 U.S. at 718.}
similarly subordinating institutions, it has not by any means instituted an equal-
birth principle.

IV. A SOLUTION: A NEW DOCTRINE AND A NEW POLITICS

Parts II and III, taken together, illustrate that the existing equal protection doctrine supports identity politics and thereby harms identity group members as individuals and society as a whole. Part IV, therefore, proposes that the existing doctrine be replaced with a new doctrine designed to liberate us from identity politics and its harms. Part IV(A) illustrates that the existing doctrine is fatally flawed when examined from pragmatic, originalist, and structuralist perspectives, in that it fails to honor our constitutional commitments to equality, autonomy, and democracy. Part IV(B) proposes criteria for the development of a new doctrine that would comport with pragmatic, originalist, and structuralist principles and honor the full range of our constitutional commitments. It suggests a new regime of ideology politics, which, free from the harms of identity politics, would provide individuals with autonomy and society with the ability to function as a true democracy.

A. Arguments Against Existing Equal Protection Doctrine

Part IV(A) illustrates that the existing doctrine is fatally flawed when examined from at least three different interpretive perspectives. Part IV(A)(1), drawing on Part II(D), makes the pragmatic argument that the existing doctrine is flawed in perpetuating identity politics and, in turn, harms both identity group members and society. Part IV(A)(2) makes the originalist argument that the existing doctrine is flawed in failing to honor our constitutional commitment to equality as originally conceived by the framers and ratifiers of the Fourteenth Amendment. Part IV(A)(3), finally, makes the structural argument that the existing doctrine is flawed in preventing the fulfillment of our corollary constitutional commitments to autonomy and democracy. Because the existing doctrine is fatally flawed, Part IV(A) ultimately concludes that it should be entirely abandoned.

1. From Pragmatism

From a pragmatic perspective, the existing doctrine is flawed in perpetuating identity politics and all its resulting harms. While pragmatism has been defined in many ways by many scholars, a general definition will suffice for present

175 For full discussion of the harms of identity politics, see supra Part II.D.
176 As early as 1988, Professor Daniel Farber commented, “An impressive array of recent legal commentary has suggested a movement away from grand theory toward something new, variously called ‘intuitionism,’” “pragmatism,” “prudence,” “institutionalism,” or “practical reason.” The difference among the views of these various scholars is mostly a matter of emphasis. . . . [W]hat they have in common is . . . their rejection of foundationalism and their emphasis on context, judgment, and community.”
purposes. Pragmatism is, at its core, the “solving [of] legal problems”—such as the meaning of the Fourteenth Amendment’s commitment to equality—“using every tool that comes to hand, including precedent, tradition, legal text, and social policy.” 177 The influence of pragmatism, and particularly of reasoning from social policy, in constitutional interpretation has been recognized in texts as venerated as *The Common Law*. 178 Indeed, Justice Holmes observed that “[t]he felt necessities of the time, the prevalent moral and political theories, intuitions of public policy, avowed or unconscious, even the prejudices which judges share with their fellow-men, have had a good deal more to do than the syllogism in determining the rules by which men should be governed.” 179

Because the existing equal protection doctrine is organized around identity groups, it perpetuates identity politics and thereby harms both identity group members and society as a whole. By protecting individuals only when they are subordinated by their identity group membership, or lack thereof, the existing doctrine reifies identity groups. It encourages individuals who share nothing more than an anatomical or quasi-anatomical attribute associated with a set of subordinating cultural norms to enter the public forum as a group. Although such individuals may have widely diverse ideologies, they are perceived as sharing a single meta-ideology. When identity group members acquiesce in their group’s meta-ideology, they are recognized; when they resist their group’s meta-ideology, they are marginalized.

The existing doctrine makes identity group membership the primary path to political power. It causes individuals to view identity groups—rather than ideology groups—as the primary political units. This, in turn, causes individuals to privilege the anatomical and quasi-anatomical attributes that define them as identity group members and to marginalize the non-anatomical attributes that set them apart from their identity groups. It causes them to view infringements on anatomical or quasi-anatomical attributes as grave harms, while viewing infringements on non-anatomical attributes as innocuous. The existing doctrine causes individuals to conform their ideologies to group norms without recognizing or complaining about the deprivation of autonomy that inevitably accompanies that conformity.

The existing doctrine not only causes individuals to define themselves by reference to identity groups, it also causes them to view the entire political landscape in terms of identity groups. Under an identity politics regime, individuals understand the spectrum of political views by reference to the spectrum of recognized identity group ideologies. Similarly, they understand equality as equality of identity groups. Equal protection doctrine, thus, distorts public forum

Daniel A. Farber, *Legal Pragmatism and the Constitution*, 72 MINN. L. REV. 1331, 1334–35 (1988). Since Professor Farber’s 1988 article, many new varieties of pragmatism have been identified.

177 Id. at 1332.
179 Id.
debate without creating the perception of any democratic deficiency. It is, in sum, a root cause of our lack of autonomy and democracy.

2. From Originalism

From an originalist perspective, the existing doctrine is flawed in failing to fulfill our commitment to equality as arguably conceived by some of the framers and ratifiers of the Fourteenth Amendment. While originalism, like pragmatism, has been defined in many ways by many scholars, the term is used here in its most general sense to refer to “[t]he theory that the U.S. Constitution should be interpreted according to the intent of those who drafted and adopted it.” It is unnecessary, for present purposes, to delve into the differences among the many forms of originalism.

Our original commitment to equality, expressed in the Fourteenth Amendment’s Equal Protection and Citizenship Clauses, was a commitment to abolishing the Black Codes as well as to instituting a broader equal-birth principle. While abolishing the Black Codes could be—and, indeed, was—achieved through a doctrine that addressed a narrow set of identity-based inequities, instituting a broader equal-birth principle would have required a doctrine capable of addressing a broad array of inequities.

Although we cannot know exactly what inequities the framers and ratifiers of the Equal Protection and Citizenship Clauses intended their equal-birth principle to address, we can hypothesize that some of them may have intended it to address some inequities beyond those experienced by identity groups—that is, beyond inequities arising solely from anatomical and quasi-anatomical differences. An equal-birth principle might, for example, reasonably have been expected to address inequities arising from non-anatomical differences, such as religious, philosophical, or other ideological differences, as well as non-identity-based differences, such as wealth- or age-based differences. As previously mentioned, Professor Amar suggests that some members of the Thirty-ninth Congress might have contemplated an equal protection doctrine addressing both religion- and wealth-based inequities.

While the existing doctrine is nominally capable of addressing inequities beyond those experienced by identity groups—for example, through its fundamental rights component—it does so only rarely and while retaining identity

---


181 BLACK’S LAW DICTIONARY 1126 (7th ed. 1999).

182 See supra Part III.B.

183 See supra Part III.C.

184 AMAR supra note 69, at 381–82.
groups as its organizing force. It therefore fails to institute a broader equal-birth principle, and thus arguably fails to fulfill our commitment to equality as originally conceived.

3. From Structuralism

From a structuralist perspective, the existing doctrine is flawed in preventing us from honoring our constitutional commitments to autonomy and democracy. Structuralism admonishes us to consider the full range of our constitutional commitments prior to deploying any one of them through a specific doctrine. A variety of scholars have advocated in favor of structuralism. Notably, Charles Black, in his lectures on *Structure and Relationship in Constitutional Law*, argued that “in dealing with questions of constitutional law, we [ought to engage not only in] exegesis of . . . particular textual passage[s, but also in] inference from the structures and relationships created by the constitution in all its parts . . .” John Hart Ely, in *Democracy and Distrust*, similarly argued that, in dealing with open-textured clauses like the Equal Protection Clause, we must transcend “clause-bound” textualism and infer meaning from the Constitution as a whole. Akhil Amar has echoed and developed upon these arguments with his method of “intra-textual” analysis, which involves the side-by-side comparison of similar clauses. These theories warn that, if we fail to appreciate the Constitution’s entire structure, we will fail to honor the full range of our constitutional commitments.

Structuralism, therefore, suggests that we ought to measure the efficacy of the existing equal protection doctrine in light of its ability to further the full range of our constitutional commitments. Although equal protection doctrine should ideally promote all—or at least not impede any—of our constitutional commitments, the existing doctrine impedes our commitments to equality, autonomy, and democracy. While Part IV(A)(2) illustrates that the existing doctrine fails to fully honor our commitment to equality, the following two subparts illustrate, respectively, how the doctrine also fails to honor our corollary commitments to autonomy and democracy.

(a) Our Constitutional Commitment to Autonomy

Our constitutional commitment to autonomy derives from at least four clauses—the Due Process Clauses of the Fifth and Fourteenth Amendments and the Freedom of Speech and Assembly Clauses of the First Amendment. The Due

---

185 See supra Part III.C.
188 See generally Amar, supra note 187, at 749.
189 It might also be manifest in the Ninth Amendment.
Process Clauses prohibit the federal and state governments from depriving any person “of life, liberty, or property, without due process of law.”\textsuperscript{190} While these clauses were not originally intended to protect substantive liberties,\textsuperscript{191} the Court has for most of the twentieth century deployed them to protect such liberties.\textsuperscript{192} According to current doctrine, “The Due Process Clause guarantees more than fair process, and the ‘liberty’ it protects includes more than the absence of physical restraint. The Clause also provides heightened protection against government interference with certain fundamental rights and liberty interests.”\textsuperscript{193} The Court applies strict scrutiny to state actions that infringe upon “fundamental rights.”\textsuperscript{194}

The Court has defined “fundamental rights” as those that are “implicit in the concept of ordered liberty” and “deeply rooted in this [nation’s] history and tradition.”\textsuperscript{195} It has, to date, recognized a number of fundamental rights, including the rights “to marry, to have children, to direct the education and upbringing of one’s children, to marital privacy, to use contraception, to bodily integrity, and to abortion.”\textsuperscript{196} An examination of the Court’s substantive due process doctrine reveals that underlying all of the rights that have been deemed “fundamental” is a general right to individual autonomy.

Although the Court has never directly articulated this general right to autonomy, it has in several cases come close. In \textit{Lawrence v. Texas}, for example, the Court stated that “[l]iberty presumes an autonomy of self that includes freedom of thought, belief, expression, and certain intimate conduct.”\textsuperscript{197} It continued, quoting the plurality opinion in \textit{Planned Parenthood v. Casey}, to assert that “‘[a]t the heart of liberty is the right to define one’s own concept of existence, of meaning, of the universe, and of the mystery of human life.’”\textsuperscript{198} Similarly, Justice Ginsburg, dissenting in \textit{Gonzales v. Carhart}, spoke of “a woman’s autonomy to determine her life’s course, and thus to enjoy equal citizenship stature.”\textsuperscript{199} This

\begin{footnotesize}
\begin{enumerate}
\item\textsuperscript{190} U.S. CONST. amends. V, XIV.
\item\textsuperscript{191} The framers of the Fourteenth Amendment intended to protect certain substantive liberties through the Privileges and Immunities Clause. See CHEMERINSKY, supra note 119, at 493.
\item\textsuperscript{192} Substantive due process doctrine has its origins in the \textit{Lochner} era, during which the Court recognized economic and other substantive liberties as protected under the Due Process Clauses. See, e.g., \textit{Lochner v. New York}, 198 U.S. 45, 53 (1905). Though modern substantive due process doctrine does not protect economic liberties, it protects a variety of other substantive liberties. See, e.g., \textit{Washington v. Glucksberg}, 521 U.S. 702, 720–21 (1997) (providing a list of substantive liberties that have been protected by modern substantive due process doctrine).
\item\textsuperscript{193} \textit{Glucksberg}, 521 U.S. at 719–20 (citations omitted).
\item\textsuperscript{194} \textit{Id.} at 721.
\item\textsuperscript{195} \textit{Id.} at 720–21.
\item\textsuperscript{196} \textit{Id.} at 720 (citations omitted).
\item\textsuperscript{197} 539 U.S. 558, 562 (2003) (invalidating Texas’s anti-sodomy law on due process grounds).
\item\textsuperscript{198} \textit{Id.} at 574 (quoting \textit{Casey}, 505 U.S. at 851).
\item\textsuperscript{199} 127 S. Ct. 1610, 1641 (2007) (Ginsburg, J., dissenting) (“Thus, legal challenges to undue restrictions on abortion procedures do not seek to vindicate some generalized notion
\end{enumerate}
\end{footnotesize}
language reveals the already existing overlap between the Court’s equal protection and substantive due process doctrines. Taken together, the Court’s substantive due process cases suggest the existence of a general right to autonomy.200

The Freedom of Speech201 and Assembly Clauses—which provide that “Congress shall make no law . . . abridging the freedom of speech . . . or the right of the people peaceably to assemble”202—also have been understood as securing a general right to individual autonomy.203 Indeed, Justice Brandeis, concurring in Whitney v. California, explained that the freedoms of speech and assembly are protected because “[t]heir who won our independence believed that the final end of the State was to make men free to develop their faculties.”204 Similarly, Justice Thurgood Marshall, concurring in Procunier v. Martinez, stated that “[t]he First Amendment serves not only the needs of the polity but also those of the human spirit—a spirit that demands self-expression.”205

Many scholars have also recognized that our freedoms of speech and assembly ought to be understood by reference to the purpose of promoting autonomy. Martin Redish, for example, has asserted that freedom of speech “ultimately serves only one true value—individual self-realization.”206 By individual self-realization, Professor Redish comprehends both “development of the individual’s powers and abilities—an individual ‘realizes’ his or her full potential—and the individual’s control of his or her own destiny through making life-affecting decisions—an individual ‘realizes’ the goals in life that he or she has set.”207 C. Edwin Baker has similarly explained that “[t]o engage voluntarily in a speech act is to engage in self-definition or expression” and argued that the “self-expressive use of speech” is important, “independent of any effective

____________________________________

200 That autonomy was highly valued is evident from the free speech decision of Whitney v. California, where Justice Brandeis said, “Those who won our independence . . . believed liberty to be the secret of happiness and courage to be the secret of liberty.” 274 U.S. 357, 375 (1927) (Brandeis & Holmes, J.J., concurring).

201 It should be noted at the outset of this discussion that, because the original intent behind the Freedom of Speech Clause is unclear, the Court’s First Amendment doctrine has long been relatively non-originalist.

202 U.S. CONST. amend. I.

203 It has also been understood to secure a number of other goals. See, e.g., Chemerinsky, supra note 119, at 925–30 (summarizing the goals of protecting speech as (1) self-governance; (2) discovering truth; (3) advancing autonomy; and (4) promoting tolerance). These four goals are widely accepted, as all have been manifest in Supreme Court decisions. See, e.g., Whitney, 274 U.S. at 375 (Brandeis & Holmes, J.J., concurring).

204 274 U.S. at 375 (Brandeis & Holmes, J.J., concurring) (holding that a criminal syndicalism statute did not violate the First Amendment rights to freedom of speech, assembly, or association).


207 Id.
communication to others, for self-fulfillment or self-realization.” The Freedom of Speech and Assembly Clauses, therefore, secure a right to autonomy.

The right to autonomy also derives from the unenumerated right to association, which arises from the Freedom of Speech and Assembly Clauses. In Roberts v. United States Jaycees, for example, the Court elaborated upon the right of association by saying that “the ability independently to define one’s identity . . . is central to any concept of liberty.” There are, thus, at least four clauses and one unenumerated right representing our commitment to autonomy.

(b) Our Constitutional Commitment to Democracy

Our constitutional commitment to democracy—and, more specifically, to democratic decision-making—pervades the entire Constitution, but it is particularly evident in the First Amendment’s Freedom of Speech Clause. That clause has, indeed, often been understood by reference to the purpose of promoting democratic debate. Justice Brandeis, in the previously mentioned Whitney v. California opinion, explained:

Those who won our independence . . . believed that freedom to think as you will and to speak as you think are means indispensable to the discovery and spread of political truth; . . . that it is hazardous to discourage thought, hope and imagination; . . . [and] that the path of safety lies in the opportunity to discuss freely supposed grievances and proposed remedies.

This illustrates that freedom of speech promotes democracy and that our commitments to autonomy and democracy are deployed in closely-related and, indeed, overlapping doctrines.

Inasmuch as the existing equal protection doctrine fails to honor the full range of our constitutional commitments—particularly, our commitments to equality, autonomy, and democracy—it is flawed from a structuralist perspective. The fact

---

209 See, e.g., NAACP v. Alabama, 357 U.S. 449, 460 (1958) ("Effective advocacy of both public and private points of view, particularly controversial ones, is undeniably enhanced by group association, as this Court has more than once recognized by remarking upon the close nexus between the freedoms of speech and assembly. It is beyond debate that freedom to engage in association for the advancement of beliefs and ideas is an inseparable aspect of the ‘liberty’ assured by the Due Process Clause of the Fourteenth Amendment, which embraces freedom of speech.") (citations omitted).
211 As Charles Black has stated, “[F]rom the very structure of the relation between the national representative and his constituency, there arises a compelling inference of some national constitutional protection of free utterance . . . .” BLACK, supra note 186, at 42.
212 274 U.S. 357, 375 (1927) (Brandeis & Holmes, J.J., concurring).
that the existing doctrine is, in fact, flawed from three highly disparate yet all well-accepted interpretive perspectives—pragmatism, originalism, and structuralism—suggests that it should be discarded and replaced with a new doctrine that better comports with those interpretive perspectives.

**B. Suggestions for a New Doctrine and a New Politics**

Part IV(B) begins by offering criteria for the development of a new doctrine that better comports with the pragmatic, originalist, and structuralist goals of equality, autonomy, and democracy. It proceeds by suggesting that this new doctrine might usher in a new “ideology politics” that would allow individuals to engage in activism based on not only identity groups but also ideology groups. Ideology politics, by encouraging alliances based on the non-anatomical attribute of ideology (while retaining alliances based on anatomical and quasi-anatomical attributes) would vitiate the harms of identity politics and promote our commitments to equality, autonomy, and democracy.

This Article does not attempt to elaborate the specifics of a new doctrine, but rather offers three criteria to guide its development. As previously mentioned, a new doctrine ought to improve upon the existing doctrine by better comporting with (1) pragmatic, (2) originalist, and (3) structuralist principles.213

First, from a pragmatic perspective, the new doctrine ought to vitiate the harms of identity politics. That is, it ought to promote both individual autonomy and democracy. While the existing doctrine diminishes individual autonomy by privileging anatomical and quasi-anatomical attributes and using them as proxies for ideology, the new doctrine ought to promote individual autonomy by equalizing all three types of attributes and preventing the use of any one of them as proxy for another.214 While the existing doctrine distorts democracy by

---

213 These principles are fully discussed in supra, Part IV.A.
214 Although this Article does not attempt to fully develop a new doctrine, one way of equalizing all three types of attributes would be to recognize claims based on not only anatomical and quasi-anatomical differences (in keeping with the current practice) but also non-anatomical differences (for example, religious and other ideological differences) and even non-identity-based differences (for example, wealth- and age-based differences). While some of these claims are currently recognized under other constitutional doctrines—such as those arising from the Due Process, Freedom of Speech, and Religion Clauses—their existence does not obviate the need for a reformed equal protection doctrine. Indeed, recognizing such claims under other doctrines can neither mitigate our obsession with identity groups nor broaden our identity-centric concept of equality—both of which arise from the existing equal protection doctrine.
215 Again, although this Article does not attempt to fully develop a new doctrine, one way of preventing the use of one type of attribute as proxy for another would be to provide greater protection against stereotyping. While the existing equal protection doctrine provides some protection against stereotyping—see, e.g., Weinberger v. Wiesenfeld, 420 U.S. 636, 645 (1975) (invalidating a law that provided more financial support for widows than widowers based on the sex stereotype that widows were unlikely to be employed while widowers were likely to be employed)—this protection could be more robust and
privileging identity group inequities and making identity groups the central agents of public-forum debate, the new doctrine ought to promote democracy by recognizing both identity and ideology group inequities and encouraging ideology groups to become more powerful and independent participants in public forum debate. A new doctrine, thus structured, would benefit individuals by allowing them to give equal weight—or, indeed, to give whatever weight they personally felt was due—to their identity and ideology group memberships; it would benefit society as a whole by allowing our policy choices to more accurately reflect the actual distribution of viewpoints.

Second, from an originalist perspective, the new doctrine ought to honor our constitutional commitment to equality as originally conceived by the framers and ratifiers of the Fourteenth Amendment. Whereas the existing doctrine is arguably too narrow to honor the full scope of our commitment to equality, the new doctrine ought to be broad enough to protect against the full range of inequities that the framers and ratifiers comprehended. This might, as previously discussed, require the institution of what Professor Amar has called an “equal-birth principle,” which would in turn require the recognition of claims based on not only anatomical and quasi-anatomical attributes but also non-anatomical attributes like ideology. A new doctrine, thus structured, would more fully honor our commitment to equality.

Third, from a structuralist perspective, the new doctrine ought to honor the full range of our constitutional commitments. While the existing doctrine impedes fulfillment of our commitments to equality, autonomy, and democracy, the new doctrine need not present such impediments. First, it could promote a broader understanding of equality, extending not only to anatomical and quasi-anatomical inequities but also to non-anatomical inequities such as those resting on ideology. Second, it could increase individual autonomy by removing the pressure on members of certain identity groups to adopt certain meta-ideologies. Finally, it could bring our society closer to true democracy by creating a public forum populated by a diverse array of identity and ideology groups, all composed of willing members and therefore reflective of the will of the people. A new

prevent presumed associations between any two attributes.

216 See supra, Part IV.A.2.

217 See supra, Part III.B.2, for a full discussion of the equal-birth principle. Although not directly relevant to the present discussion, it is worth noting that an equal-birth principle might require recognition of not only ideological differences but also non-identity-based differences such as wealth- and age-based differences.

218 See supra, Part IV.A.2–3.

219 Again, though not directly relevant to the present discussion, it is worth noting that a broader understanding of equality might encompass not only non-anatomical inequities but also non-identity-based inequities. It would not privilege anatomical and quasi-anatomical inequities by regarding their harms as greater than those inflicted by non-anatomical and non-identity-based inequities.
doctrine, thus structured, would represent significant progress toward honoring the full range of our constitutional commitments.220

A new doctrine, guided by these pragmatic, originalist, and structuralist principles, could enable us to transcend identity politics and inaugurate a new politics. While the current doctrine has allowed identity groups to be the central participants in public forum debate, the new doctrine could promote ideology groups as equally powerful participants. A new doctrine should not eliminate identity groups—identity groups will, indeed, remain relevant at least as long as our society continues to create and support anatomical and quasi-anatomical inequities—but rather should decentralize identity groups through the addition of a diverse array of new ideology groups.221 A public forum populated by truly independent identity and ideology groups would benefit individuals and society as a whole.

First, a public forum populated by equally powerful identity and ideology groups would benefit individuals by allowing them to independently express their anatomical, quasi-anatomical, and non-anatomical subordination. At present, individuals lack a certain measure of ideological independence—in order to challenge their anatomical and quasi-anatomical subordination, individuals are often forced to conform their true ideologies to their identity groups’ meta-ideologies. Adding truly independent ideology groups to the public forum would provide individuals with ideological independence. By disaggregating their identity group membership from their ideologies, it would allow them to simultaneously express their anatomical and quasi-anatomical subordination as well as their own ideologies.

Second, a public forum populated by equally powerful identity and ideology groups would benefit society as a whole by enabling us to make better policy choices based on better understandings of the actual distribution of viewpoints. While identity groups have espoused meta-ideologies, which allow individuals to either agree or disagree but not to insert a third perspective, ideology groups could (in the aggregate) espouse a more diverse array of more nuanced ideologies.222

220 There may be other commitments beyond those to equality, autonomy, and democracy that are impeded by the existing equal protection doctrine. Their discussion is, however, beyond the scope of this Article.

221 While there are certainly a number of ideology groups in the current public forum (see supra, Part II), many of them are not truly independent in the sense that they are intimately linked with identity groups. It is, indeed, often presumed that people from certain identity groups belong to certain ideology groups. I, therefore, advocate for the addition of ideology groups that are truly independent from identity groups—such that individuals are not pressured into them by their possession of certain anatomical and quasi-anatomical attributes.

222 It should be noted that the shift from a few meta-ideologies to many more diverse ideologies will inevitably change the nature of public forum debate. The proliferation of ideologies may, indeed, result in some loss of coherence. I believe, however, that some loss of coherence is a worthwhile trade-off for the resulting gain in authenticity of individual expression.
This greater diversity of ideologies would create the potential for better policy choices. Indeed, debates with more diverse viewpoints tend to generate more innovative and effective results. Our policy choices will, however, not only be better, they will also better reflect the will of the people. Once anatomical and quasi-anatomical attributes have been disaggregated from non-anatomical attributes—or, put another way, once identity and ideology groups have been recognized as distinct—we will be able to more accurately track ideological distributions. Rather than using identity groups as indirect indicators of their members’ ideologies, we will be able to use ideology groups as direct indicators and, thus, make policy choices that truly reflect the will of the people. A public forum populated by equally powerful identity and ideology groups would thus both increase individual autonomy and promote democracy.

V. CONCLUSION

This Article describes a social problem, identifies a legal cause, and proposes a legal solution. It therefore presupposes that the law has a profound influence on our social patterns—that the law has mired us in identity politics and can liberate us from identity politics. Although we can never be certain of the law’s actual influence on our social patterns, we can hope that it will be influential enough to mitigate our past transgressions. We can hope that, by aspiring to fulfill the aggregate of commitments expressed in our Constitution, we will eventually develop a set of doctrines capable of improving the lives of individuals and the operation of society as a whole.
EXTENDING THE SHADOW OF THE LAW: USING HYBRID MECHANISMS TO DEVELOP CONSTITUTIONAL NORMS IN SOCIOECONOMIC RIGHTS CASES

Brian Ray*

TABLE OF CONTENTS

I. INTRODUCTION ................................................................................................. 797
II. THE DEBATE OVER ADJUDICATION AND ALTERNATIVE DISPUTE RESOLUTION .......................................................................................... 801
   A. Fuller’s Forms, Functions, and Limits .............................................................. 802
   B. Fiss and the Public Function Critique ............................................................. 805
   C. Reconciling Fuller and Fiss ............................................................................ 808
III. PROMOTING LEGITIMACY THROUGH ADR PROCESSES ................................. 810
   A. Sturm’s Deliberative Model .......................................................................... 810
   B. A Norm-Creating ADR Process Independent of Litigation ......................... 812
IV. DEVELOPING A HYBRID MODEL THROUGH “ENGAGEMENT” ....................... 814
   A. The Socioeconomic Rights Debate ............................................................... 814
   B. City of Johannesburg and the Engagement Remedy .................................. 820
   C. Engagement as a Hybrid Mechanism for Developing Constitutional Norms .................................................................................... 825
      1. Successful Engagements ............................................................................. 827
      2. Bad-Faith Engagements ............................................................................. 831
   D. Establishing and Extending Engagement .................................................... 834
      1. Engagement in Eviction Cases .................................................................. 834
      2. The Limits of Engagement ........................................................................ 837
V. CONCLUSION ................................................................................................... 842

I. INTRODUCTION

One distinctive feature of constitutions developed in the twentieth century is their almost uniform inclusion of socioeconomic rights provisions—rights to basic human needs such as food, water, shelter, health care, and education.1 Despite the

* © 2009 Brian Ray, Assistant Professor of Law, Cleveland-Marshall College of Law. I would like to thank Lolita Buckner Inniss, Joan Flynn, Richard Goldstone, and participants in the South Africa Reading Group for their helpful comments and suggestions on earlier versions of this work. Angela Barstow and the staff of the Cleveland-Marshall Library, especially Jessica Mathewson, provided invaluable research assistance. The Cleveland-Marshall Fund provided generous research support. It was a pleasure working with the staff of the Utah Law Review in the editing and publication process.

relative ubiquity of these rights, however, judicial enforcement of them remains relatively controversial in theory and problematic in practice.2

While concerns over judicial review arguably are heightened in the socioeconomic rights context, the arguments over enforcement of these rights largely mirror the debate over judicial enforcement of constitutional rights more generally. In particular, both debates focus on the undemocratic nature of judicial review and, consequently, are concerned with defining (and confining) the judicial role in ways that maximize its legitimacy.3

Defenses of judicial review are connected to those of adjudication more generally. They often locate their legitimacy in a set of procedural characteristics—such as judicial independence, structured participation, and reasoned decisions—that promote objective results and serve the public interest. Lon Fuller in his famous essay The Forms and Limits of Adjudication defines “the distinguishing characteristic of adjudication” as “the fact that it confers on the affected party a peculiar form of participation in the decision, that of presenting proofs and reasoned arguments for a decision in his favor.”4 In a related vein, Owen Fiss argues that judicial review is rooted in a “conception of the judicial function [that] sees the judge as trying to give meaning to our constitutional values” and a “process through which that meaning is revealed or elaborated.”5

The more flexible, person-centered processes associated with alternative dispute resolution, or “ADR,” are often contrasted with adjudication. Both critics and proponents of alternative dispute resolution often assume that, unlike adjudication, these processes are inherently limited to solving particular disputes and thus are unable to establish precedents applicable beyond a single dispute.6

---


3 See Michelman, supra note 2, at 16 (“[I]t is clear that the debate [over socioeconomic rights] throughout has been centered on a concern about the place and work of the judiciary in the democratic political order. We seem to think the problem with constitutionalizing social rights comes down mainly, if not solely, to a matter of the separation of powers.”).

4 Lon L. Fuller, The Forms and Limits of Adjudication, 92 HARV. L. REV. 353, 368 (1978) [hereinafter Fuller, Forms and Limits].

5 Owen Fiss, Foreword: The Forms of Justice, 93 HARV. L. REV. 1, 12–13 (1979) [hereinafter Fiss, Foreword].

6 See, e.g., Harry T. Edwards, Alternative Dispute Resolution: Panacea or Anathema?, 99 HARV. L. REV. 668, 676–77 (1986) (“However, if ADR is extended to resolve difficult issues of constitutional or public law—making use of nonlegal values to resolve important social issues or allowing those the law seeks to regulate to delimit public rights and duties—there is real reason for concern.”); Owen M. Fiss, Against Settlement, 93 YALE L.J. 1073, 1085 (1984) [hereinafter Fiss, Against Settlement] (The role of adjudication “is not to maximize the ends of private parties, nor simply to secure the peace, but to explicate and give force to the values embodied in authoritative texts such as the Constitution and statutes: to interpret those values and to bring reality in accord with them.
This severely limits the capacity of ADR to generate public values and perform the norm-creation function usually associated with adjudication generally and constitutional adjudication in particular.

This Article challenges the general perception that ADR processes cannot develop public law norms. It follows a recent trend in ADR literature that seeks to define a public norm creation role for ADR in part by connecting these processes to other alternative legal and political problem-solving methods.\(^7\) This Article focuses on a recent South African Constitutional Court case, *Occupiers of 51 Olivia Road v City of Johannesburg*,\(^8\) in which the court interpreted the right to housing in the South African Constitution.\(^9\) The court held that municipalities must develop processes for negotiating—or, in the court’s language “engaging”—with citizens affected by redevelopment plans, to analyze how claims about the norm-creation potential of ADR processes could be developed in the context of constitutional adjudication of socioeconomic rights.

The heightened legitimacy and separation of powers concerns associated with socioeconomic rights\(^{10}\) mean that they have been a rich source for examining the use of alternative enforcement approaches. The South African Constitutional Court

---

\(^{7}\) *See*, e.g., Carrie Menkel-Meadow, *Deliberative Democracy and Conflict Resolution*, 12 NO. 2 DISP. RESOL. MAG. 18, 19 (2006) (arguing that there are strong connections between deliberative democracy theory and the ADR movement including a shared appreciation for “constitutional experimentalism . . . in which there are feedback mechanisms for sharing and coordinating local outcomes with the broader polity”); *see also* Amy J. Cohen, *Negotiation, Meet New Governance: Interests, Skills, and Selves*, 33 LAW & SOC. INQUIRY 503, 527–29 (2008) (examining connections between negotiation literature and new governance literature); Carrie Menkel-Meadow, *The Lawyer’s Role(s) in Deliberative Democracy*, 5 NEV. L.J. 347, 348 (2004) (exploring “the use of alternative legal, political and social problem solving institutions that draw on conflict resolution theory and practice”); Sturm & Gadlin, *supra* note 6, at 3 (arguing that ADR processes are capable of generating public law norms “when relevant institutional actors develop values or remedies through an accountable process of principled or participatory decision making, and then adapt those values and remedies to broader groups or situations”).

\(^{8}\) 2008 (5) BCLR 475 (CC) (S. Afr.).

\(^{9}\) S. AFR. CONST. 1996 § 26.

\(^{10}\) *See*, e.g., Dennis M. Davis, *Socioeconomic Rights: Do They Deliver the Goods?*, 6 INT’L J. CONST. 503, 678–89 (2008) (discussing objections to the inclusion of socioeconomic rights in the South African Constitution); *see generally In re Certification of the Constitution of S. Afr.* 1996 (4) SALR 744 (CC) at 793 (S. Afr.) (rejecting the argument that socioeconomic rights included in the new constitution are inconsistent with the separation of powers established by the constitution because they “would result in the courts dictating to the government how the budget should be allocated”).
is one of the most active courts in this area; other jurisdictions and academic literature cite its decisions as models for developing alternative approaches.\textsuperscript{11}

\textit{Occupiers of 51 Olivia Road v City of Johannesburg}\textsuperscript{12} portends a potentially important development in its approach to enforcement. The court adopted the term “engagement” to describe a unique remedy it developed—in essence, a permanent negotiation/mediation requirement in housing rights cases that may involve eviction.\textsuperscript{13} Properly implemented, the engagement remedy can be developed into a hybrid dispute resolution model. This model integrates ADR processes with formal adjudication in a manner that enhances the legitimacy of the resolution and makes possible extra-judicial interpretation and enforcement of socioeconomic rights. This hybrid process is particularly well-suited to enforcing socioeconomic rights because it is more democratic than formal adjudication and also more flexible and responsive to the practical concerns that socioeconomic rights raise.

Part II of this Article outlines two classic but competing accounts of the procedural justifications for adjudication and related assessments of the limitations of ADR processes by Fuller and Fiss.\textsuperscript{14} Despite their differences, the characteristics

\begin{itemize}
\item See, e.g., \textsc{Mark Kende, Constitutional Rights in Two Worlds: South Africa and the United States} 244 (2009) (“The South African Constitution’s socioeconomic rights provisions have been celebrated internationally.”); \textsc{Cass Sunstein, Designing Democracy: What Constitutions Do} 236 (2001) (arguing that the Constitutional Court’s enforcement approach “suggests . . . for the first time, the possibility of providing . . . protection [for socioeconomic rights] in a way that is respectful of democratic prerogatives and the simple fact of limited budgets”); \textsc{Tushnet, supra note 1, at xii (summarizing chapters 7 and 8 of the text, which draw “on South Africa’s developing jurisprudence of social welfare rights [to] show that the ‘capacity’ objection to judicial enforcement of social and economic rights rests on the assumption that such enforcement must take a strong form”); Michelman, \textit{supra} note 2, at 15 & n.8 (2003) (citing \textit{Minister of Health v. Treatment Action Campaign} 2002 (1) BCLR 1022 (CC) (S. Afr.), as “supporting evidence” that judges “can find both properly adjudicative standards for testing claims of social-rights violations and worthwhile, properly judicial remedies for violations when found”); Jeanne M. Woods, \textit{Justiciable Social Rights as a Critique of the Liberal Paradigm}, 38 Tex. Int’l L.J. 763, 766–67 (2003) (“The South African experience in the constitutional adjudication of social rights has profound implications for the international community at large”); Katharine G. Young, \textit{The Minimum Core of Economic and Social Rights: A Concept in Search of Content}, 33 Yale J. Int’l Law, 113, 158 (2008) (“South African constitutional law, a vanguard in many areas of constitutional rights, has inspired much commentary on the way that the minimum core concept might resolve the justiciability challenges of economic and social rights.”).
\item 2008 (5) BCLR 475 (CC) (S. Afr.).
\item See \textit{id.} at para. 5.
\item \textit{See generally Fiss, Foreword, supra note 5; Fiss, Against Settlement, supra note 6; Fuller, Forms and Limits, supra note 4; Lon L. Fuller, Mediation—Its Forms and Functions, 44 S. Cal. L. Rev. 305 (1971) [hereinafter Fuller, Mediation]. One commentator notes, “Fuller has become the paradigm of dispute resolution, just as Fiss is the paradigm of public law litigation.” Robert G. Bone, \textit{Lon Fuller’s Theory of Adjudication and the False Dichotomy Between Dispute Resolution and Public Law Models of Litigation}, 75 B.U. L. Rev. 1273, 1279–80 (1995). Bone cites a range of sources to support this
\end{itemize}
Fuller and Fiss identify as legitimizing adjudication share important similarities that Susan Sturm has argued ADR processes can promote and protect.\textsuperscript{15} Part III analyzes two related articles by Sturm and Gadlin that develop this argument. Sturm first identifies four key characteristics that Fuller’s and Fiss’s accounts share and argues that these characteristics can be advanced through mediation in the remedial phase of public law litigation.\textsuperscript{16} Sturm and Gadlin, in a more recent article, propose an ADR model that promotes those same values and therefore can be used to develop public norms outside of adjudication.\textsuperscript{17}

Part IV summarizes the debate over socioeconomic rights generally and the specific debate over the South African Constitutional Court’s enforcement approach and identifies important ways in which the arguments there track the debate over the relative roles and the legitimacy of adjudication and ADR. It then describes the Constitutional Court’s decision in \textit{City of Johannesburg} and argues that the court’s engagement remedy can be developed into a hybrid process incorporating aspects of adjudication and mediation/negotiation. It argues that this can be done in a way that retains the flexibility and responsiveness Fuller prizes in ADR, while still protecting the legitimacy norms both Fuller and Fiss associate with adjudication.

II. THE DEBATE OVER ADJUDICATION AND ALTERNATIVE DISPUTE RESOLUTION

Beginning in the mid-1970s, the idea of “conflict resolution” or “dispute resolution” outside of formal adjudication gained increased attention among courts, lawyers, and the general public.\textsuperscript{18} Proponents of ADR claim that its methods address many systemic problems in litigation and offer several benefits not available through traditional litigation. First, ADR could relieve congested court dockets while also offering expedited resolution to parties.\textsuperscript{19} Second, ADR techniques could give parties to disputes more control over the resolution characterization. \textit{Id.} at 1279 n.19 (citing William N. Eskridge, Jr., \textit{Metaprocedure}, 98 YALE L.J. 945, 955–56, 962–64 (1989) (book review); Richard L. Marcus, \textit{Public Law Litigation and Legal Scholarship}, 21 U. MICH. J.L. REFORM. 647, 684-85, 687 (1988); Arthur R. Miller, \textit{Confidentiality, Protective Orders, and Public Access to the Courts}, 105 HARV. L. REV. 427, 431 n.8 (1991)).


\textsuperscript{16} \textit{Id.}

\textsuperscript{17} See Sturm & Gadlin, \textit{supra} note 6.

\textsuperscript{18} See SARAH R. COLE ET AL., MEDIATION LAW § 5.2 (2d ed. 2007); MICHAEL FREEMAN, ALTERNATIVE DISPUTE RESOLUTION xi (1995).

The flexibility of ADR also creates opportunities for creative remedies that could more appropriately address underlying concerns in a dispute than could traditional remedies. By offering the opportunity for consensus-based resolution, ADR also is arguably better suited than litigation to preserving long-term relationships and solving community-based disputes.

One prominent metaphor that captures the distinction between ADR and traditional adjudication is the “shadow of the law” notion developed by Robert Mnookin and Lewis Kornhauser in the context of divorce. Under this view, the law acts “not as imposing order from above, but rather as a providing a framework within which [parties] can themselves determine their . . . rights and responsibilities.”

A. Fuller’s Forms, Functions, and Limits

Lon Fuller’s essay, Mediation—Its Forms and Functions, was one of the first attempts to theorize ADR processes. For Fuller “the central quality of mediation” is “its capacity to reorient the parties towards each other, not by imposing rules on them, but by helping them to achieve a new and shared perception of their relationship.” In dispute resolution processes like mediation, the goal, then, is not to get parties to accept formal rules to govern their

---


21 See, e.g., Carrie Menkel-Meadow, Whose Dispute Is It Anyway?: A Philosophical and Democratic Defense of Settlement (In Some Cases), 83 GEO. L.J. 2663, 2689–90 (1995) (“[P]arties (with the expert advice of lawyers) can decide how much ‘public discourse’ or confidentiality they need to resolve their dispute, how much direct confrontation or conversation they want with the other side, and how much flexibility they want to work out possible solutions that a court would not be authorized to award.”).

22 See, e.g., Deborah R. Hensler, Our Courts, Ourselves: How the Alternative Dispute Resolution Movement Is Re-Shaping Our Legal System, 108 PENN ST. L. REV. 165, 171 (2003) (“The adoption of mediation by community justice centers may have reflected a belief that mediation—often characterized by its supporters as antithetical to adversarial dispute resolution processes—was more likely to nurture positive relationships within the community.”).


24 Id. at 950.

25 Fuller, Mediation, supra note 14.

26 See Carrie Menkel-Meadow, Mothers and Fathers of Invention: The Intellectual Founders of ADR, 16 OHIO ST. J. DISP. RESOL. 1, 13 (2000) (“In many ways, Lon Fuller remains the only legal philosopher to take theorizing about dispute resolution processes seriously.”).

27 Fuller, Mediation, supra note 14, at 325.
relationship, but to help them “to free themselves from the encumbrance of rules” and to accept “a relationship of mutual respect, trust, and understanding that will enable them to meet shared contingencies without the aid of formal prescriptions laid down in advance.”

Fuller contrasts the mediative function with what he calls “the standard procedures of law,” central to which “is the concept of rules.” Drawing a sharp distinction between “acts” and “persons,” Fuller defines rules as “requiring, prohibiting or attaching specific consequences to acts” and places them in the realm of adjudication. By contrast, mediation is concerned principally with persons and relationships, and it deals with “precepts eliciting dispositions of the person, including a willingness to respond to somewhat shifting and indefinite ‘role expectations.”

In Fuller’s conception, mediation has no role to play in the interpretation and enforcement of laws; that is the role of courts and the function of adjudication: “[O]nce a law has been duly enacted its interpretation and enforcement is for the courts; courts have been instituted, not to mediate disputes, but to decide them.”

Thus, Fuller establishes a sharp dichotomy between the “rule of law” and ADR processes. Central to this dichotomy is his notion that rules “attribut[e] legal or social consequences to overt and specifically defined acts.” Rules (and laws) are established in advance and must be sufficiently precise both in terms of defining the conduct to which they apply and the consequences they will entail. By contrast, dispute resolution processes, in their focus on people and relationships, do not require “impersonal, act-prescribing rules” and therefore are particularly well-suited for dealing with the kinds of “shifting contingencies” inherent in ongoing and complex relationships.

Fuller further suggests that modern society creates an increasing number of people-dependent problems suitable for “mediative” approaches. He lists public welfare systems and public hospitals as prime examples in which the responsibility for distributing public goods “certainly needs to be at least ‘mediative’ (that is, as open-mindedly consultative)” as other mediated disputes.

Fuller’s essay The Forms and Limits of Adjudication analyzes the contrasting process of adjudication. For Fuller, a particular mode of participation defines adjudication: “the distinguishing characteristic of adjudication lies in the fact that it confers on the affected party a peculiar form of participation in the decision, that of

---

28 Id. at 325–26.
29 Id. at 327–28.
30 Id. at 329.
31 Id.
32 Id. at 328.
33 Id. at 329 (emphasis added).
34 Id. at 330–31.
35 Id. at 336–37.
36 Id.
37 Fuller, Forms and Limits, supra note 4.
presenting proofs and reasoned arguments for a decision in his favor.”38 The legitimacy of adjudication is dependent on the degree to which it maximizes this form of participation because adjudication is “a device which gives formal and institutional expression to the influence of reasoned argument in human affairs.”39

This rationality principle not only provides the justification for adjudication’s authority, but also limits the kinds of disputes appropriate for adjudication: “Wherever successful human association depends upon spontaneous and informal collaboration, shifting its forms with the task at hand, there adjudication is out of place except as it may declare certain ground rules applicable to a wide variety of activities.”40 Fuller describes these kinds of disputes as “polycentric,” meaning that they involve complex and intersecting sets of relationships.41 Adjudication “cannot encompass and take into account the complex repercussions” that result from resolution of a polycentric dispute.42 More importantly, in such disputes “it is simply impossible to afford each affected party a meaningful participation through proofs and arguments.”43

Socioeconomic rights present a paradigmatic example of polycentricity.44 Although he does not use the term “polycentric,” Frank Michelman’s description of the “raging indeterminacy” of socioeconomic rights captures their intense polycentric nature.45 Using a hypothetical right to “effective social citizenship” as an example, Michelman points out that determining whether such a right has been violated requires ascertaining the net effect of a range of government policies with uncertain and potentially conflicting effects.46

Thus for Fuller, polycentric disputes, such as those socioeconomic rights create, pose a real dilemma. The interrelated nature of the disputes’ issues is more susceptible to the give-and-take of ADR processes like mediation than to the

38 Id. at 368.
39 Id. at 366.
40 Id. at 371.
41 Id. at 395.
42 Id. at 394.
43 Id. at 394–95.
44 See, e.g., Bel Porto Sch. Governing Body v Premier of W. Cape Province 2002 (9) BCLR 891 (CC) at para. 51 (S. Afr.) (discussing the “polycentric” argument by respondents and citing Fuller, Adjudication, supra note 4); Kate O’Regan, Introduction to Socio-Economic Rights, 1 ESR REV. NO. 4 (1999), available at http://www.chr.up.ac.za/centre_projects/socio/esrvol1no4.html#2 (“Two main arguments are raised in relation to the institutional competence of courts to enforce socio-economic rights. The first is Lon Fuller’s argument that certain types of decisions are ‘polycentric’ and therefore unsuitable for adjudication.”); Craig Scott & Patrick Macklem, Constitutional Ropes of Sand or Justiciable Guarantees? Social Rights in a New South African Constitution, 141 U. PA. L. REV. 1, 24 (1992) (“The resistance to constitutionally entrenched social rights on the grounds of institutional competence is often summarized in the view that social rights are said to be . . . vague in terms of the obligations they mandate; and involving complex, polycentric, and diffuse interests in collective goods.”).
46 Id.
application of rules characteristic of adjudication. But modern society increasingly requires resolution of public disputes that are polycentric, and Fuller’s sharp distinction between the rule of law and ADR means that relegating such disputes to the realm of ADR comes at the cost of diminishing, if not eliminating, the ability of the resolution to establish any broader norm applicable across disputes.

Fuller never offers a solution to this apparent dilemma. Nor does he identify any process that is ideal for resolving polycentric disputes. But a close reading of *Mediation* and *Adjudication* together suggests that Fuller was, in fact, open to using hybrid forms of dispute resolution to deal with polycentric disputes.

In *Adjudication*, Fuller acknowledges that adjudication’s ability to deal with polycentricity depends primarily on the degree to which decisions have precedential force: “If judicial precedents are liberally interpreted and are subject to reformulation and clarification as problems not originally foreseen arise, the judicial process as a whole is enabled to absorb these covert polycentric elements.” Thus, more flexible forms of adjudication (in other words, adjudication that looks more like mediation) have greater capacity to deal with polycentricity.

Here, recall Fuller’s account of mediation and its characteristic ability to deal with shifting contingencies. Although he does not directly cite mediation as a mechanism for addressing polycentricity, the complex issues he describes as suitable for “mediative” approaches in his *Mediation* essay are plainly polycentric. And Fuller’s description of the interrelated issues typically present in mediation echoes his polycentric examples.

**B. Fiss and the Public Function Critique**

The emphasis on participant control over ADR processes, which is central to their claimed benefits, also forms the basis of one of the principal criticisms of ADR processes, i.e., that private resolution eliminates the public norm creation function of adjudication. Owen Fiss, in one of the earliest and most influential criticisms of the ADR movement, argues that adjudication is not merely a tool for resolving private disputes, but it is also “an institutional arrangement for using state power to bring a recalcitrant reality closer to our chosen ideals.” In other words, adjudication serves important public purposes that extend beyond the boundaries of a particular dispute and the interests of the parties to that dispute.

Fiss’s critique of ADR processes is rooted in his view that adjudication is a primarily public function that derives its legitimacy from a particular process. For

---

47 See generally Fuller, *Forms and Limits*, supra note 4.
48 Id.
49 Id. at 398.
51 Id. at 336 (describing distribution of “scarce public welfare funds” and “the problem of the crowded public hospital” as problems suitable for meditative approaches).
52 Id. at 317–18.
53 Fiss, *Against Settlement*, supra note 6, at 1089.
Fiss, a court’s power to resolve cases is based on a “conception of the judicial function [that] sees the judge as trying to give meaning to our constitutional values” and a view that “adjudication is the process through which that meaning is revealed or elaborated.”

Fiss develops his theory of adjudication in the context of a defense of court resolution of complex, institutional reform litigation, or what he calls “structural reform” litigation. As Fiss describes it,

“structural reform . . . is one type of adjudication, distinguished by the constitutional character of the public values, and even more importantly by the fact that it involves an encounter between the judiciary and the state bureaucracies. The judge tries to give meaning to our constitutional values in the operation of these organizations.”

These cases also implicate the same complex, interrelated issues that Fuller identifies as characteristic of polycentric disputes.

Fiss argues that two aspects of courts and the adjudicative process legitimize their resolution of public disputes: “one is the judge’s obligation to participate in a dialogue, and the second is his independence.” By “dialogue,” Fiss means the adversary process: judges do not pick their cases; they must listen to all parties, issue decisions, and articulate reasons for those decisions. Independence requires that the judge not identify with any of the parties; the judge’s decision must be impartial.

Fiss rejects the argument that courts lack the institutional competence to deal with complex public disputes as both empirically unsupported and inconsistent with this understanding of the judicial role. He argues that there is no convincing evidence that administrative agencies possess superior expertise in dealing with the problems raised in structural reform cases. But, Fiss argues, even accepting that courts have no claim to superior practical expertise, “[t]heir special competency lies elsewhere, in the domain of constitutional values, a special kind of substantive rationality, and that expertise is derived from the special quality of the judicial process—dialogue and independence.” Administrative agencies are too tied to the political process and therefore lack the necessary independence “that is so essential for giving expression to our constitutional values.”

---

54 Fiss, Foreword, supra note 5, at 12–13.
55 Id. at 2.
56 Id.
57 See supra note 42 and accompanying text.
58 Fiss, Foreword, supra note 5, at 13.
59 Id. at 13.
60 Id. at 14.
61 Id. at 32.
62 Id. at 33–34.
63 Id. at 34.
64 Id. at 35.
Fiss takes issue with what he views as Fuller’s cribbed conception of adjudication, focusing on Fuller’s “participation axiom,” i.e., that adjudication is both defined through and limited by the right of parties to participate in the process. Fiss notes that structural reform litigation would be impossible if adjudication were limited to cases in which individuals fully participated in the process, and he argues that such a requirement would eliminate adjudication’s ability to create public norms and, consequently, its ability to resolve a huge swath of constitutional and common law cases.

When it comes to remedies, however, Fiss implicitly acknowledges the limits of adjudication in terms strikingly similar to Fuller’s polycentricity concern. Fiss concedes that “[t]here is no likely connection between the core processes of adjudication, those that give the judge the special claim to competence, and the instrumental judgments necessarily entailed in fashioning the remedy.” In other words, the individual terms of the remedy cannot be justified by reasoned arguments—the same reason Fuller argues prevents adjudication from resolving polycentric disputes.

But Fiss argues that there is a “tight connection between meaning and remedy.” This connection “requires that the decision about remedy be vested in the judge, the agency assigned to the task of giving meaning to the value through declaration.” Delegating the remedial function to some other body “necessarily creates the risk that the remedy might distort the right, and leave us with something less than the true meaning of the constitutional value.”

While making the judge responsible for the remedy in complex cases creates a risk that she will lose some of the distance from the dispute that is central to independence, Fiss views that as a necessary compromise: “Independence is a critical element in the process that legitimatizes the judicial function, for having us believe that judges can articulate and elaborate the meaning of our constitutional values, and, yet, to fully discharge that function . . . judges are forced to surrender some of their independence.”

---

65 Id. at 42.
66 Id. at 43.
67 See id. at 52.
68 Id.
69 Fiss summarizes the “relative incapacity of adjudication to solve ‘polycentric’ problems” as rooted in “the incapacity of a given area of human activity to endure a pervasive delimitation of rights and wrongs [that] is also the measure of its incapacity to respond to a too exigent rationality, a rationality that demands an immediate and explicit reason for every step taken.” Fuller, Forms and Limits, supra note 4, at 371 (emphasis added). He goes on to explain that behind “both these incapacities lies the fundamental truth that certain kinds of human relations are not appropriate raw material for a process of decision that is institutionally committed to acting on the basis of reasoned argument.” Id.
70 Fiss, Foreword, supra note 5, at 52.
71 Id. at 52–53.
72 Id. at 53.
73 Id. at 57.
C. Reconciling Fuller and Fiss

By focusing on Fuller’s participation emphasis, Fiss ignores the strong similarities in their views of adjudication. Both see the process as distinctive in its reason-giving capacity. To be sure, Fuller emphasizes the role of the individual in that distinct process, but it is not participation for its own sake that gives adjudication its special character and legitimacy; rather, it is “participation through proofs and arguments.”74 In other words, the ability of individual participation in the structured setting of adjudication to produce a reasoned result is the basis of its legitimacy.75

Thus Fuller’s concern about polycentricity is “not merely a question of the huge number of possibly affected parties,” but instead relates to the lack of a “clear issue to which either side [of a polycentric dispute] could direct its proofs and contentions.”76 In addition, Fuller locates the fundamental problem with polycentric disputes in the fact that such disputes implicate “the incapacity of a given area of human activity to endure a pervasive delimitation of rights and wrongs [that] is also a measure of its incapacity to respond to a too exigent rationality, a rationality that demands an immediate and explicit reason for every step taken.”77 This is the same concern that Fiss raises regarding the court’s remedial function in a structural reform case: the specifics of the remedy are not susceptible to reasoned justification.78

In the end, then, Fiss’ and Fuller’s views of adjudication and its limits have significant similarities. Both agree that adjudication’s legitimacy is tied to its capacity to produce reasoned decisions through a structured process that emphasizes a particular mode of party participation and requires an independent adjudicator.79 In addition, Fiss and Fuller both believe that adjudication begins to lose legitimacy to the extent that its results cannot be justified by reasoned arguments80 and when the processes involved depart from the adversary model.81 They also acknowledge that complex disputes test the limits of that legitimacy because the remedies they entail cannot be justified solely by reason; the role of

74 Fuller, Forms and Limits, supra note 4, at 364.
75 Fuller answers “not necessarily” to the hypothetical question of whether a judge’s decision must be accompanied by reasons. Id. at 387. But he goes on to state that “[b]y and large it seems clear that the fairness and effectiveness of adjudication are promoted by reasoned opinions. Without such opinions the parties have to take it on faith that their participation in the decision has been real, that the arbiter has in fact understood and taken into account their proofs and arguments.” Id. at 388 (emphasis added).
76 Id. at 394–95.
77 Id. at 371.
78 Fiss, Foreword, supra note 5, at 52.
79 Compare id. at 12–14, with Fuller, Forms and Limits, supra note 4, at 365–69.
80 Compare Fiss, Foreword, supra note 5, at 42, with Fuller, Forms and Limits, supra note 4, at 368–69.
81 Compare Fiss, Foreword, supra note 5, at 29–30, with Fuller, Forms and Limits, supra note 4, at 386–88.
the parties is more complicated. And finally, both authors conclude that the practical need for a judge to assume a more direct role at the remedial phase compromises the judge’s independence.

The principal difference between the two approaches is that Fiss is willing to live with the compromise of permitting courts to fashion remedies in such cases, because he believes that the norm-creation capacity of adjudication requires courts to make remedial decisions. As a consequence, Fiss is critical of ADR processes out of concern that they will undermine the public values in adjudication.

Although Fuller appears to have a much stronger sense than Fiss of the limits of the ideal form of adjudication, he hints at a willingness to accept alternative forms of adjudication that fail to fully maximize individual participation. As noted above, Fuller recognizes that a liberal interpretation of precedent permits a more flexible view of adjudication “as a collaborative [process] projected through time” in which “an accommodation of legal doctrine to the complex aspects of a problem can be made as these aspects reveal themselves in successive cases.”

Fuller also cautions that his analysis of the “pure” form of adjudication that maximizes individual participation and judge neutrality should not be taken as a condemnation of every “mixed or ‘impure’ form of adjudication.” Fuller explains that he uses the term “parasitic” to describe such mixed forms in the neutral sense of a botanist, to imply that they draw “moral sustenance from another form of order.” In other words, Fuller recognizes that the real world often requires forms of adjudication that do not meet his described ideal. Thus, even for polycentric disputes, Fuller acknowledges that adjudication can set “ground rules” for a resolution that would better come from some other, more flexible form.

By contrast, Fiss rejects negotiated settlements, and other forms of private dispute resolution in public law cases because they lack the independence and reasoned decision-making procedures that legitimate adjudication. The central

---

82 Compare Fiss, Foreword, supra note 5, at 45–46, with Fuller, Forms and Limits supra note 4, at 393–95.
83 Compare Fiss, Foreword, supra note 5, at 46, with Fuller, Forms and Limits, supra note 4, at 394–98.
84 Fiss, Foreword, supra note 5, at 52–53.
85 See id. at 57–58.
86 See Fuller, Forms and Limits, supra note 4, at 371.
87 Id. at 398.
88 Id. at 405.
89 Id. at 406.
90 See Fuller, Forms and Limits supra note 4, at 371 (“Wherever successful human association depends upon spontaneous and informal collaboration, shifting its forms with the task at hand, there adjudication is out of place except as it may declare certain ground rules applicable to a wide variety of activities.”); see also Lon L. Fuller, Collective Bargaining and the Arbitrator, 1963 Wis. L. Rev. 3, 34 (“I am not here asserting that an agency called a ‘court’ should never under any circumstances undertake to solve a ‘polycentric’ problem. . . . All I am urging is that this sort of problem cannot be solved within the procedural restraints normally surrounding judicial office.”).
91 Fiss, Against Settlement, supra note 6, at 1085.
problem with settlement is that it is a purely private bargain with no necessary connection to the public values inherent in the laws at issue in the dispute. Judicial approval of a settlement thus fails “to explicate and give force to the values embodied in authoritative texts such as the Constitution and statutes: to interpret those statutes and give force to the values embodied in them.”

III. PROMOTING LEGITIMACY THROUGH ADR PROCESSES

A. Sturm’s Deliberative Model

Susan Sturm addresses the Fuller-Fiss debate, focusing specifically on the legitimacy concerns raised in the remedial phase of structural reform litigation. Sturm describes Fuller and Fiss as representative of two competing models of what she calls the process critique of adjudication.

Sturm largely agrees with Fiss’s criticisms of the limitations of Fuller’s model. But Sturm also finds fault with Fiss’s own attachment to the adversary process, noting that Fiss’s unease with the remedial stage of structural reform litigation illustrates that Fiss and Fuller “share many of the same concerns about the dangers of the court’s departure from the adversary model” of adjudication. She identifies “[t]hree shared norms of judicial legitimacy [that] underlie” Fiss’s and Fuller’s models: (1) participation, (2) judicial independence and impartiality, and (3) reasoned decision making.

Sturm rejects Fiss’s claim that judges must fashion remedies themselves to preserve the legitimacy of adjudication. While she agrees with Fuller and Fiss that participation plays an important role in legitimating the judicial function, Sturm argues that participation can be implemented in the remedial phase in a manner that departs from the traditional adversary model and also enhances the legitimacy of the process. Rather than relying on indirect representation through

---

92 Id.
93 Id. (arguing Fiss’s “structural reform model offers a powerful critique of the dispute resolution model and presents a normative theory intended to address and give legitimacy to the court’s role in public law litigation”).
94 Id. at 1359–60.
95 Id. at 1387.
96 Id. at 1390.
97 Id.
98 Id. at 1391–96.
legal professionals, Sturm advocates for maximizing the involvement of parties and other affected actors, and empowering those individuals and groups to come up with their own remedial plan. 101 Although this enhances the participation so valued by Fuller, it requires abandoning his ideal of the two-party dispute and the identification of clear-cut issues to which those parties can address their arguments. Similarly, while Sturm’s proposal embraces Fiss’s understanding that adjudication is norm-creating, it rejects his insistence that the judge must both determine liability and develop the remedy. 102

Sturm describes her approach as a “deliberative model of remedial decisionmaking.” 103 The deliberative model is essentially a proposal to use mediation to develop the remedy in a structural reform case with an enhanced role for the court. After determining liability, the court sets up a structured mediation process. The court first defines the parameters of the process, determined by “the liability norms that have been violated.” 104 At the prenegotiation stage, the judge assists in identifying stakeholders and appointing a mediator. The court also identifies characteristics of an effective consultation process, informs the participants of the standards it will use to evaluate the result, and sets deadlines. 105 This gives the court greater control over the process and requires specific attention to the underlying substantive norms.

After negotiations, the parties are required to present the court with a written agreement and explanation that the stakeholder groups have approved. The court then holds a public hearing and evaluates the remedy on three levels: (1) the adequacy of the process, (2) the responsiveness of the remedy to the concerns raised in the process, and (3) its capacity to address the underlying substantive norms. 106 This ensures that the process is fully participatory and that the result reflects reasoned decision making tied to the legal norms at stake. While the court is more directly involved than under stand-alone mediation, its impartiality and independence is preserved by limiting that role to structuring and evaluating—but not participating in—the remedial process.

For Sturm, then, adjudication’s legitimacy can be protected by using ADR processes at the remedial stage in a tightly controlled, judicially supervised setting. Like Fuller, she sees the flexibility, informality, and person-centered aspects of mediation as distinctly appropriate to addressing the remedial issues in complex public disputes. But she shares Fiss’s concern that using ADR processes risks compromise or elimination of the norms that adjudication is intended to promote. The answer is to enhance judicial control of the otherwise private process of mediation and specifically emphasize the norms at stake throughout the process.

---

101 Id.
102 See id. at 1431–32.
103 See id. at 1428.
104 Id. at 1428–29.
105 Id. at 1429–30.
106 Id. at 1431.
Sturm is careful to emphasize that her argument is limited to the remedial context and that extending the deliberative model to liability determinations would require careful consideration of the differences between the liability determination and remedial tasks of the court. Nonetheless, she notes that the model could serve a similar legitimating function in the consent decree context. If the negotiations leading to a consent decree were structured along the lines of the deliberative model, then judicial approval of that result could derive the same legitimating benefits despite the lack of a direct court role at the liability phase.

B. A Norm-Creating ADR Process Independent of Litigation

In a more recent article, Sturm, writing with Howard Gadlin, directly addresses the potential for ADR processes to serve a public norm creation function independent of adjudication. Echoing Sturm’s description of the deliberative model, Sturm and Gadlin contend that public norms emerge not only from formal adjudication but also “when relevant institutional actors develop values or remedies through an accountable process of principled and participatory decision making and then adapt these values and remedies to broader groups or situations.”

ADR processes have this norm-creation potential, provided that they are structured in a way that links “individual and systemic conflict resolution.” Sturm and Gadlin advocate a “combination of root cause analysis and multi-level remediation” to allow ADR to achieve this potential. Root-cause analysis makes implicit organizational norms explicit as part of the resolution of individual conflicts and in turn creates the opportunity for evaluating whether to reject or accept those norms. Multi-level remediation requires considering when and whether to apply the results of an individual resolution to others within an organization. As systemic problems are identified over time and solutions are applied system-wide, individual interventions “generate deliberations that produce an overarching governance structure built around principles, values and lessons” from individual resolutions. Formal law sets the outer bounds of possible

107 Id. at 1445.
108 Id. at 1446.
109 Sturm & Gadlin, supra note 6, at 3.
110 Id.
111 Id. at 4.
112 Id. at 53.
113 Id.; see also id. at 4 (“The linchpin of our approach is a form of root cause analysis, which enables intermediaries to identify and, where possible, address underlying problems as part of individual case work.”).
114 Id. at 54 (“Problems revealed through conflict resolution sometimes give rise to changes in policy, which apply to everyone similarly situated within the relevant domain.”).
115 Id.
resolutions and, at the same time, defines abstractly the values that individual resolutions must serve.\textsuperscript{116}

It is evident from Sturm and Gadlin’s description of this process that they believe it can incorporate the same legitimacy characteristics as the deliberative remedial model. But, rather than requiring direct and specific court involvement to protect public values, these informal processes are completely free of court involvement, whether at the liability or remedial stage, and yet they address all three legitimacy characteristics Sturm identifies with the deliberative model.

Thus, the authors note, “when linked to systemic change, non-adjudicative conflict resolution can foster the articulation of implicit norms [and] ‘reasoned elaboration and visible expression of public values.’”\textsuperscript{117} Also, “[c]onflict resolution thus institutionalizes principled decision making that can be generalized within the community of practice in which it operates.”\textsuperscript{118} Participation is protected through root-cause analysis, which “incorporates the participation of those affected by, responsible for and knowledgeable about, the problems at issue,” and because remedies “must emerge from this collective deliberation,” in the same manner as the deliberative model.\textsuperscript{119}

Independence and impartiality take on a more complex form. Rather than requiring the “‘detached neutrality’” characteristic of adjudication, Sturm and Gadlin argue that “‘multi-partiality’—critically analyzing a conflict from multiple vantage points”—can serve the same function.\textsuperscript{120} This, in turn, reinforces participation and reasoned decision making because it requires an “institutional design that builds in participatory accountability—ongoing examination and justification to participants and a community of practitioners.”\textsuperscript{121}

Sturm and Gadlin acknowledge limits to the norm-creation ability of stand-alone ADR processes, but they argue that the limits “focus attention on the interdependence of informal and formal conflict resolution systems.”\textsuperscript{122} The nature of the process and, more important, the identity of communities involved in the process will dictate the relative legitimacy and applicability of the norms developed through it.

Recognizing the complex relationship between formal and informal dispute resolution processes moves the debate over the relative capacities of ADR and formal adjudication beyond the sharp dichotomies reflected in the arguments of Fiss and Fuller, while recognizing the concerns of both. As noted above, Fuller’s apparent receptivity to hybrid forms of adjudication\textsuperscript{123} hints at a similar understanding that a combination of formal and informal processes can be the most effective at addressing complex public law issues. But Fuller was plainly

\textsuperscript{116} Id. at 54–55.
\textsuperscript{117} Id. at 55.
\textsuperscript{118} Id.
\textsuperscript{119} Id. at 56.
\textsuperscript{120} See id. at 4.
\textsuperscript{121} Id.
\textsuperscript{122} Id. at 56.
\textsuperscript{123} See id.; supra notes 50–52 and accompanying text.
concerned that this effectiveness would come at the cost of a loss of legitimacy. In the same way, Fiss recognizes the potential, at least at the remedial level, for other processes to displace adjudication, but is deeply concerned about compromising the legitimacy of the result.

Sturm’s deliberative model offers an initial step beyond the stark choices presented by Fiss and Fuller, but one that still emphasizes the need for direct oversight and significant court involvement. Sturm and Gadlin’s description of individual ADR processes, conscious of and linked to systemic problem solving, de-links norm creation from formal adjudication completely while still recognizing the relationship between formal and informal norm-creation processes.

Part IV analyzes a recent South African Constitutional Court case in which the court interpreted the right to housing in the South African Constitution to require that municipalities develop processes for “engaging” with citizens affected by redevelopment plans that may involve eviction. The case is used to consider how Sturm and Gadlin’s claims about ADR’s potential to create public norms could be extended further. Properly implemented, engagement can be developed into a hybrid dispute resolution model integrating ADR processes with formal adjudication in a manner that enhances the legitimacy characteristics identified by Sturm, while facilitating non-judicial development of public norms for socioeconomic rights.

IV. DEVELOPING A HYBRID MODEL THROUGH “ENGAGEMENT”

A. The Socioeconomic Rights Debate

Judicial enforcement of socioeconomic rights raises strong objections on institutional competence and separation of powers grounds. Critics of judicial enforcement argue that courts are simply not equipped to deal with the complex, interrelated issues these rights raise.124 Because enforcement has substantial and specific effects on the state budget, critics also argue that judicial enforcement raises insurmountable separation of powers concerns and creates significant practical problems by restricting the ability of the political branches to set budget priorities.125 Notably, Fuller’s argument against adjudication of “polycentric” disputes features prominently in the literature as an argument against

124 See, e.g., Michelman, supra note 2, at 15 (discussing theoretical objections to the constitutionalization of social and economic rights); see generally Kim Lane Schepple, Social Rights in Constitutional Courts: Strategies of Articulation and Strategies of Enforcement, 4–6 (Oct. 2008) (unpublished draft, copy on file with the author) (summarizing major criticisms of judicial enforcement of social rights).

125 See, e.g., Tushnet, supra note 1, at 231–33 (discussing the “conventional argument against judicial enforcement of social and economic rights”); Schepple, supra note 124, at 4–5 (“Many commentators believe that courts should not commit large segments of the state’s budget to particular purposes because that is a proper role for a democratically elected legislature, not for court judgment.”).
constitutionalizing socioeconomic rights and for limiting judicial enforcement when they are included.\textsuperscript{126}

These same objections, somewhat paradoxically, also emerge in the literature surrounding structural reform litigation. Fiss’s contention that adjudication is central to the elaboration of public norms\textsuperscript{127} suggests that structural reform litigation is a genre of disputes where adjudication is particularly important. Yet, judicial resolution of these disputes has instead prompted widespread criticism of the legitimacy and institutional capacity of courts to deal with the complex issues they raise.\textsuperscript{128} These are precisely the same objections to judicial enforcement of socioeconomic rights.\textsuperscript{129}

Recognizing the force of these objections in the socioeconomic rights context, the South African Constitutional Court has taken a leading role in developing innovative approaches to enforcement that attempt to mitigate institutional competence and separation of powers concerns. The most celebrated example of this innovation is the court’s first housing rights case, \textit{Government of the Republic of South Africa v Grootboom}.\textsuperscript{130}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image.png}
\caption{Image description}
\end{figure}

\begin{itemize}
\item \textsuperscript{126} See, e.g., supra note 44 and accompanying text.
\item \textsuperscript{127} See Fiss, Foreword, supra note 5, at 12–13.
\item \textsuperscript{128} See Sturm, supra note 15, at 1379. Sturm notes that “[t]he court’s dynamic and activist role in formulating public law remedies has triggered a heated debate among academics, judges, and politicians concerning the proper role of the court.” \textit{Id.} She identifies four major criticisms:
\begin{itemize}
\item 1. The courts’ public remedial activities fail to conform to the standards of legitimate judicial decisionmaking;
\item 2. The courts violate principles of federalism and separation of powers . . .;
\item 3. The courts are not competent to perform the role of public remedial formulation; and
\item 4. The courts are abusing their power and acting unfairly in the execution of their public remedial function.
\end{itemize}
\textit{Id.}
\item \textsuperscript{129} Compare \textit{id.} (listing the four objections described in the immediately preceding footnote), with Tushnet, supra note 1, at 231–33 (quoting Frank R. Cross’s summary of the arguments against socioeconomic rights in \textit{The Error of Positive Rights}, 48 UCLA L. Rev. 857, 887 (2001), which includes the arguments that enforcement of socioeconomic rights “raises the spectre of ‘the courts running everything—raising taxes and deciding how the money should be spent’” and that “[j]udges . . . are ill-suited for the evaluation and making of the trade-offs implied by many positive rights” (internal quotations omitted)), and Michelman, supra note 2, at 15 (“[I]t is clear that the debate [over socioeconomic rights] throughout has been centered on a concern about the place and work of the judiciary in the democratic political order. We seem to think that the problem with constitutionalizing social rights comes down mainly, if not solely to a matter of separation of powers.”).
\item \textsuperscript{130} 2000 (11) BCLR 1169 (CC) (S. Afr.); see also Sunstein, supra note 11, at 233–36 (analyzing \textit{Grootboom} and praising the court’s “adoption of a novel and highly promising approach to judicial protection of socio-economic rights”); Davis, supra note 10,
The plaintiffs in Grootboom were desperately poor members of an informal community who had established makeshift housing on a public sports field after their eviction from nearby private land. They brought suit against the City of Cape Town claiming that the city’s failure to provide housing violated their right to access to adequate housing under section 26(2) of the constitution, as well as the rights of the children in the community to shelter under section 28(1)(c).131

The court held that the city violated the general right to housing in section 26 because, although it had developed constitutionally adequate housing programs addressing medium- and long-term needs, the city’s plans lacked any provision for short-term, emergency needs like those of the Grootboom community.132 The court, however, limited its relief to a declaration that the state housing program in the Cape municipal region was unconstitutional in that it “failed to make reasonable provision within its available resources for people in the Cape Metropolitan area with no access to land, no roof over their heads, and who were living in intolerable conditions or crisis situations.”133 The court refused to issue more specific requirements and dissolved an injunction requiring the city to report back to court on its progress implementing a new housing plan.134 The city was thus left to determine for itself how best to cure the constitutional defect with no direct court oversight.

In large part due to the court’s limited remedy, Grootboom has been described variously as a “dialogic,”135 “weak form,”136 or “administrative law”137 approach. These labels all highlight the fact that the court’s use of a general declaration significantly limited the court’s role and largely left policy development to the political branches. For those same reasons, however, critics of the court’s approach in Grootboom have charged it with proceduralizing these rights by refusing to give them any substantive content.138


131 Grootboom (11) BCLR 1169 at para. 12.
132 Id. at paras. 95–96.
133 Id. at para. 99.
134 Id.
135 Rosalind Dixon, Creating Dialogue About Socioeconomic Rights: Strong-Form Versus Weak-Form Judicial Review Revisited, 5 INT’L CONST. L.J. 391, 417 (2007) (arguing that the court should adopt a “dialogue model,” which is “fully consistent with the approach the Court in TAC suggested might be appropriate in future cases”).
136 TUSHNET, supra note 1, at 242–43.
137 SUNSTEIN, supra note 11, at 234 (“What the South African Constitutional Court has basically done is to adopt an administrative law model of socioeconomic rights.”).
These critics focus on an important interpretive move made by the court in Grootboom and other cases. Most of the socioeconomic rights in the South African Constitution contain what is called an “internal limitations” provision. The right itself is stated in the first clause. For example, section 26 provides: “1. Everyone has the right to have access to adequate housing.” That right is then qualified in the second clause (the “internal limitation”), which states: “2. The state must take reasonable legislative and other measures, within its available resources, to achieve progressive realisation of this right.” The court has consistently held that these two provisions must be read together.

Critics argue that the court should interpret the right itself independent of the limitation. This would require a two-step process in which the court first declares what the right to housing requires in the abstract, and then considers cost-based justifications for any particular program that falls short of this ideal. As one critic describes it, this approach “view[s] resource scarcity not as qualifying the ambit [of the right] but rather as limiting the extent to which its implied benefits may be demanded at a given time.”

In addition to limiting the substantive scope of these rights, critics charge that the court’s refusal to interpret the right separate from the limitation fails to give adequate guidance to the political branches and potential claimants. They claim “there is a need for the Court to clarify the State’s obligations imposed by socio-economic rights.” Without such guidance, “the state is left with an amorphous standard by which to judge its own conduct” and is unable “to assess its conduct against clear benchmarks.”

Critics also argue that the court’s failure to give substantive content to these rights renders its decisions arbitrary and illegitimate. As one commentator puts it, the reasonableness standard adopted by the court “seems to stand for whatever the Court regards as desirable features of state policy. The problem with this approach is that it lacks a principled basis upon which to found decisions in socio-economic rights cases.”

Marius Pieterse, Resuscitating Socio-Economic Rights: Constitutional Entitlements to Health Care Services, 22 S. Afr. J. Hum. Rts. 473, 473–74 (2006) (“[T]he Court’s rejection of what can be called a ‘minimum core approach’ to the enforcement of ss 26(1) and 27(1) of the Constitution in favour of an administrative law-like ‘reasonableness approach’ . . . has been much lamented.”). Mark Kende provides a comprehensive survey of, and response to, these criticisms. Kende, supra note 11, at 243–85.

140 Id.
141 See, e.g., Gov’t of the Republic of S. Afr. v Grootboom 2000 (11) BCLR 1169 (CC) at para. 33 (S. Afr.).
142 See, e.g., Pieterse, supra note 138, at 481.
143 Id. at 480.
144 Bilchitz, supra note 138, at 10.
145 Id.
146 Id.
These criticisms of the court’s enforcement approach reflect Fiss’s focus on the public function of adjudication and the unique role of courts in that process. In effect, the court’s critics are saying, as does Fiss, that only the court has the power and responsibility to interpret these rights and establish constitutional values. Without an authoritative court interpretation, the government cannot determine its constitutional responsibilities and citizens cannot challenge government programs and actions that fall short. Likewise, Fiss’s related concern about the threatened legitimacy of adjudication when the court does not dictate the remedy—because we might be left “with something less than the true meaning of the constitutional value”\(^\text{147}\)—is reflected in the criticism of *Grootboom*’s limited remedy and the court’s broader refusal to define the meaning of the right and the remedy separately in each case.

Finally, the charge that the court’s refusal to declare the content of the right makes it impossible to test its decisions in each case against any objective standard reflects Fiss’s emphasis that the legitimacy of adjudication rests principally on the fact that a judicial decision is the result of reasoned argument (dialogue) and is “justif[ied] . . . in terms of the norms of the constitutional system.”\(^\text{148}\)

But Fiss’s recognition that a court’s legitimacy is diminished when it fashions the details of a remedy because those details represent a choice among competing options for implementing the constitutional norm\(^\text{149}\) points toward a flaw in the criticisms of the Constitutional Court’s approach as well. Fiss acknowledges that “[t]he task of discovering the meaning of constitutional values such as equality, due process, or property, is . . . quite different from choosing or fashioning the most effective strategy for actualizing those values.”\(^\text{150}\) This is because “there is no likely connection between the core processes of adjudication, those that give the judge the special claim of competence, and the instrumental judgments necessarily entailed in fashioning the remedy.”\(^\text{151}\)

This gap between the abstract right and the practical remedy collapses when it comes to socioeconomic rights in a way that makes it almost impossible to disentangle the two. For example, take the right to health care. In contrast to the equality example used by Fiss, where the right itself could be interpreted to mean “racial equality,” which would then require choosing among a range of policy

\(^{147}\) Fiss, *Foreword*, supra note 5, at 53.

\(^{148}\) *Id.* at 45.

\(^{149}\) *Id.* at 49 (“The judge must search for the ‘best’ remedy, but since his judgment must incorporate such open-ended considerations as effectiveness and fairness, and since the threat and constitutional value that occasions intervention can never be defined with great precision, the particular choice of remedy can never be defended with any certitude.” (emphasis added)).

\(^{150}\) *Id.* at 51.

\(^{151}\) *Id.* at 52.
choices to implement that interpretation, it is impossible to define the right to “health care” without reference to specific policies.

This is illustrated by criticisms of the Constitutional Court’s interpretation of the right to health care in Treatment Action Campaign. Tracking the interpretive debate just described, critics argue the court should have defined the right to health care independent of the limitation. This would mean deciding, for example, “[w]hat are the services to which one is entitled to claim access? Do these services involve preventative [sic] medicine, such as immunizations, or treatment for existing diseases or both? Does the right entitle one to primary, secondary, or tertiary health care services, or all of these?” In other words, defining the right necessarily entails prescribing particular policies.

Furthermore, as Fiss also emphasizes, courts are in no better position (and likely a worse one) than the political branches to make the instrumental assessments necessary to determine whether a particular health care policy is more or less likely to be effective. Like the decisions over allocation of public goods that Fuller cites, interpreting the right to health care necessarily requires courts to make contingent policy decisions that will almost certainly shift over time and demand an intimate, real-time understanding of the specific conditions on the ground to be effective.

On the one hand, the “raging indeterminacy” of socioeconomic rights lends support to the court’s decision to interpret the right and the limitation together. Defining the right in the abstract requires articulating substantive policies in light of specific conditions in particular contexts. More important, however, the indeterminacy heightens the legitimacy deficit that Fiss acknowledges the remedial function creates for adjudication in complex disputes like these, due to the loss of judicial independence, and the lack of a necessary connection between the constitutional value and the specifics of the remedy. Thus, accepting Fiss’s view of adjudication’s legitimacy, there is at least no greater legitimacy problem in the

---

152 Id. at 52–53.
153 Sandra Liebenberg makes a similar point when she observes that the Constitutional Court’s enforcement approach implies “that there is no bright-line boundary between law and policy, and that substantive evaluative choices will have to be made regarding when and how the courts should intervene in policy choices which impact on people’s socio-economic welfare.” Sandra Liebenberg, Socioeconomic Rights: Revisiting the Reasonableness Review/Minimum Core Debate, in CONSTITUTIONAL CONVERSATIONS 303, 308 (Stu Woolman & Michael Bishop eds., 2008).
154 2002 (10) BCLR 1033 (CC) (S. Afr.).
155 Bilchitz, supra note 138, at 6.
156 Id.
157 Id.
158 Others have recognized these same institutional-competence problems with judicial enforcement. See, e.g., CASS SUNSTEIN & STEVEN HOLMES, THE COST OF RIGHTS 126–27 (1999) (noting that judges “lack the fact-finding capacity . . . that would justify their making particular allocative decisions”).
159 See Michelman, supra note 2, at 30–31.
160 See Fiss, Foreword, supra note 5, at 46.
court’s current approach than if the court were to interpret the right before considering whether the limitations clause justified a departure from that ideal. Any declaration of the meaning of the right to housing would be a policy proposal, i.e., a remedy, and thus would entail the kind of instrumental judgments that diminish the legitimacy of the result.

Sturm’s proposal that mediation can be used in the remedial phase in a manner that still protects the essential legitimating features of adjudication, and Sturm and Gadlin’s extension of that argument to ADR processes independent of adjudication, suggest a way out of the legitimacy dilemma posed by socioeconomic rights. At the same time, using ADR processes offers a possible solution to the institutional competence concerns they raise.

The Constitutional Court’s most recent housing rights decision in Occupiers of 51 Olivia Road v City of Johannesburg involved an interim order requiring a form of negotiation/mediation, which the court calls “engagement.” The court constitutionalized the engagement requirement in all future housing rights cases. After a brief description of that decision and the litigation that led to it, the following subsections consider how this requirement might satisfy the characteristics for a norm-creating ADR process described by Sturm and Gadlin, and thereby enhance the legitimacy of the court’s role in enforcing these rights. The hybrid dispute resolution process that engagement could become is fully consistent with the legitimacy norms Sturm and Gadlin identify, and it reflects the kind of hybrid process Fuller implies is most suitable for resolving polycentric disputes of this kind.

B. City of Johannesburg and the Engagement Remedy

City of Johannesburg began as a series of emergency applications in the Witwatersrand High Court by the City of Johannesburg to evict over 300 people from six properties in inner-city Johannesburg. The city sought these evictions as part of a broader regeneration strategy, one aspect of which was the identification, clearance, and ultimate redevelopment of more than 200 “bad” buildings with some 67,000 occupants in the inner-city district.

The targeted buildings clearly created unsafe and unsanitary living conditions. Notwithstanding the legitimate concerns over the health and safety hazards the buildings presented, the city’s eviction program was outrageous. The city implemented the evictions by filing form applications with very little notice to

162 City of Johannesburg v Rand Props. (Pty) Ltd. 2006 (6) BCLR 728 (W) at para. 2 (S. Afr.).
163 See CENTRE ON HOUSING RIGHTS AND EVICTIONS, ANY ROOM FOR THE POOR? FORCED EVICTIONS IN JOHANNESBURG, SOUTH AFRICA (March 8, 2005), at 41–46, 60–64 (describing the regeneration plan and the practice of forced evictions) [hereinafter Any Room?].
actual occupants and in most cases gaining summary eviction with no hearing.\textsuperscript{165} The city would then send in teams of workers dressed in red, known as “red ants,” to forcibly evict residents. The Geneva-based Center on Housing Rights and Evictions (COHRE) published an extensive report describing the abuses and outlining legal and policy arguments against the program.\textsuperscript{166}

Several groups organized the residents of six targeted buildings to oppose the applications in their cases.\textsuperscript{167} COHRE had partnered with the Centre for Applied Legal Studies (CALS) in drafting the initial report criticizing the eviction program.\textsuperscript{168} CALS then coordinated the litigation strategy.\textsuperscript{169} Other groups were also active in the effort, including the Community Law Centre,\textsuperscript{170} a public interest research and advocacy group based at the University of the Western Cape. Several of these groups filed amicus curiae briefs in support of the residents.\textsuperscript{171}

The High Court (the trial-level court in the South African system) was extremely sympathetic to the residents’ arguments. The court rejected the city’s eviction application and issued a broad order holding that the city had violated section 26 by pursuing these evictions without a plan to house the evicted residents, as required by \textit{Grootboom} and related legislation.\textsuperscript{172} The court also enjoined the city from seeking to evict the residents, and it ordered the city to develop a plan for housing these and other residents. Notably, the order specifically required the city to relocate residents within the inner-city district.\textsuperscript{173}

The city appealed to the Supreme Court of Appeal (SCA). The SCA reversed the high court, finding that the evictions were constitutionally permissible but triggered a much more limited responsibility by the city to relocate the displaced residents.\textsuperscript{174}

\begin{paracol}{2}
\begin{paracolenum}{2}
\item Any Room?, \textit{supra} note 163, at 60–64 (describing in detail the city’s typical eviction procedures).
\item Id.
\item \textit{City of Johannesburg v Rand Props. (Pty) Ltd.} 2007 (6) BCLR 643 (SCA) at para. 13 (S. Afr.).
\item See \textit{Any Room?}, \textit{supra} note 163, at 5 n.1; see also Press Release, COHRE/CALS, Jo-Burg City Housing Policy Goes to Bloemfontein (Feb. 20, 2007) (“The plight of [the residents] was first brought to public attention in a May 2005 report co-authored by researchers from the Centre for Applied Legal Studies (CALS) and COHRE . . . .”), available at http://www.law.wits.ac.za/cals/.
\item Community Law Centre, \textit{http://www.communitylawcentre.org.za/Court-Interventions} (click “read more” under “Rand Properties – right to adequate housing and evictions”) (last visited Sept. 1, 2009).
\item Id.
\item See \textit{City of Johannesburg v Rand Props. (Pty) Ltd.} 2006 (6) BCLR 728 (W) at para. 65 (S. Afr.).
\item Id. at 67.
\item Id. at 67.
\item \textit{City of Johannesburg v Rand Props. (Pty) Ltd.} 2007 (6) BCLR 643 (SCA) at para. 5 (S. Afr.).
\end{paracolenum}
\end{paracol}
The residents appealed the SCA’s order to the Constitutional Court, which accepted the application in May 2007. The court heard oral argument on August 28, 2007, and two days later it issued a remarkable interim order requiring the parties to “engage with each other meaningfully . . . in an effort to resolve the differences and difficulties aired in this application in light of the values of the Constitution, the constitutional and statutory duties of the municipality and the rights and duties of the citizens concerned.” The court also ordered the parties to file affidavits reporting the results of the negotiations with the court approximately one month later, on October 3, 2007.

After requesting several deadline extensions, the parties eventually reached a partial settlement that included the following provisions: in the short-term, the city agreed to cease its eviction attempts and to take specific measures to make the existing buildings safer and more habitable by cleaning the buildings and providing sanitation services, access to water, and functioning toilets. Before relocating the residents from the buildings designated for redevelopment, the city agreed to refurbish several other buildings in inner-city Johannesburg to at least provide “security against eviction; access to sanitation; access to potable water; access to electricity for heating, lighting and cooking” and to limit any rental fees to no more than 25 percent of the occupants’ monthly income. Finally, the city agreed to consult with the residents on the “provision of suitable permanent housing solutions . . . having regard to applicable national, provincial and municipal housing policies.”

Despite agreeing to these remarkable terms, both sides pressed the court to decide the broader issue of whether the city was in compliance with section 26 and Grootboom’s mandate to develop a plan that addresses the emergency needs of individuals like the residents in this case. The city submitted a “Draft Inner City Housing Plan” along with its affidavit reporting the results of the negotiation, and it requested that the court find that the plan satisfied constitutional obligations under section 26. The residents filed a supplementary affidavit objecting to the city’s submission of the new plan in the context of an affidavit that was intended to

176 Id. at para. 1.
177 Id. at para. 3.
178 Settlement Agreement Between City of Johannesburg and the Occupiers of 51 Olivia Road, Berea Township and 197 Main Street, Johannesburg (Oct 29, 2007) at paras. 2-4 (copy on file with the author).
179 Id. at para. 6.
180 Id. at para. 7.
181 Id. at para. 18.
182 Id. at paras. 3, 10.
be only a report of the parties’ negotiated settlement.\textsuperscript{184} Nonetheless, the residents
continued to urge the court to determine whether the city was in compliance with
section 26 and asked for additional time to respond to the city’s new plan.\textsuperscript{185}

The Constitutional Court issued its final opinion and order on February 19,
2008. The court specifically refused to deal with the residents’ broader claim that
the city still lacked a comprehensive housing plan as required by \textit{Grootboom}.
Citing the city’s commitment in the settlement agreement to develop a long-term
housing plan in consultation with the residents, the court found that “[t]here is
every reason to believe that negotiations will continue in good faith.”\textsuperscript{186} The court
noted that the city’s position had evolved considerably as demonstrated by its
“willingness to engage,” and the court was optimistic that “[t]here is no reason to
think that future engagement will not be meaningful and will not lead to a
reasonable result.”\textsuperscript{187} The court also emphasized that court intervention remains an
enforcement option “if this course becomes necessary.”\textsuperscript{188}

The court then formalized the negotiation/mediation requirement, calling it
“engagement.”\textsuperscript{189} It noted that it had called for versions of engagement in earlier
cases.\textsuperscript{190} In particular, in another eviction case, \textit{Port Elizabeth Municipality v
Various Occupiers}, the court stated:

\begin{quote}
In seeking to resolve the above contradictions, the procedural and
substantive aspects of justice and equity cannot always be separated. The
managerial role of the courts may need to find expression in innovative
ways. Thus one potentially dignified and effective mode of achieving
sustainable reconciliations of the different interests involved is to
encourage and require the parties to engage with each other in a pro-
active and honest endeavour to find mutually acceptable solutions.
Wherever possible, respectful face-to-face engagement or mediation
through a third party should replace arms-length combat by intransigent
opponents.\textsuperscript{191}
\end{quote}

The court found that a range of constitutional provisions, including the state’s
obligation to “encourage the involvement of communities and community

\textsuperscript{184} Moray Hathorn Affidavit, \textit{available at} http://www.constitutionalcourt.org.za
(search “Olivia Road” within “Cases” and select “Applicant’s Supplementary Affidavit”
under “Pleadings and Documents”) (Oct. 2007).
\textsuperscript{185} \textit{Occupiers of 51 Olivia Road v City of Johannesburg} 2008 (5) BCLR 475 (CC) at
paras. 31–34 (S. Afr.).
\textsuperscript{186} \textit{Id.} at para. 34.
\textsuperscript{187} \textit{Id.}
\textsuperscript{188} \textit{Id.}
\textsuperscript{189} \textit{Id.} at paras. 9–23.
\textsuperscript{190} \textit{Id.} at paras. 10–12.
\textsuperscript{191} \textit{Port Elizabeth Municipality v Various Occupiers} 2004 (12) BCLR 1268 (CC) at
para. 39 (S. Afr.).
organizations in local government, as well as the rights to human dignity and life broadly require engagement with citizens affected by state policies. The court then described in more specific terms what engagement should entail.

First, in accordance with *Grootboom*, the court emphasized that “section 26(2) mandates that the response of any municipality to potentially homeless people with whom it engages must also be reasonable.” Reasonableness is context-specific and permits a range of substantive outcomes: “It may in some circumstances be reasonable to make permanent housing available and, in others, to provide no housing at all. The possibilities between these extremes are almost endless.”

Second, in most cases, and in particular where a large-scale program is involved, engagement must be more than a merely “ad hoc” process. Emphasizing that “[i]t must have been apparent [from the outset of the city’s regeneration strategy planning] that the eviction of a large number of people was inevitable,” the court also noted that “[i]f structures had been put in place with competent sensitive council workers skilled in engagement, the process could have begun when the strategy was adopted.” Thus, engagement must be incorporated at the outset of any long-term planning process and must involve a trained cadre of government employees.

Third, the court recognized that “[p]eople about to be evicted may be so vulnerable that they may not be able to understand the importance of engagement and may refuse to take part in the process.” These vulnerable groups “must not be regarded as a disempowered mass.” Instead, the state must make every effort to engage, and these groups may require assistance from civil society groups. For this reason, the court specifically recognized that “[c]ivil society organisations that support the peoples’ claims should preferably facilitate the engagement process in every possible way.”

Finally, the court established what amounts to a public reporting requirement for the government after any engagement process. Emphasizing that “secrecy is counter-productive to the process of engagement,” the court stated, “[T]he provision of a complete and accurate account of the process of engagement including at least the reasonable efforts of the municipality with that process would ordinarily be essential.” Courts are then required to consider “[w]hether there

---

192 *Occupiers of 51 Olivia Road* 2008 (5) BCLR 475 at para. 16.
194 *Id.* § 11.
195 *Occupiers of 51 Olivia Road* 2008 (5) BCLR 475 at para. 16.
196 *Id.* at para. 18.
197 *Id.*
198 *Id.*
199 *Id.* at para. 19.
200 *Id.* at para. 15.
201 *Id.* at para. 19.
202 *Id.*
203 *Id.* at para. 21.
had been meaningful engagement between a city and the resident about to be rendered homeless” when considering a challenge under section 26.  

C. Engagement as a Hybrid Mechanism for Developing Constitutional Norms

On the one hand, the court’s description of the engagement process looks much like Sturm and Gadlin’s model of a norm-creating ADR process. It pushes the responsibility for developing the substantive content of section 26 into the political sphere and adopts a largely private, party-directed mechanism for doing so. Unlike their model, which operates separate from the judicial system, engagement remains tied to the courts in ways that make it more of a hybrid between the pure ADR process they describe and pure adjudication. The hybrid nature of engagement enhances its legitimacy and its norm-creation capacity in several ways.

The Constitutional Court created a structure that is really an expanded version of Sturm’s deliberative model, but one that operates without a court liability determination. Therefore, it tracks Sturm’s suggestion that the model might be extended to consent decrees. Rather than setting specific guidelines in each case, as under Sturm’s model, the court instead sets more general guidelines for the engagement process, as it did in City of Johannesburg. The guidelines will, presumably, be refined and expanded in cases that come before the court where engagement was tried and failed. This gives courts the power to structure the engagement process in ways that ensure attention to the values these rights protect, as Sturm and Fiss emphasize. Although this structuring role operates over a longer period and across multiple cases rather than within a single one, it is nonetheless similar to the structuring role Sturm assigns to the court in the deliberative model, and thus allows the courts to refine the process to ensure attention to public values.

The engagement remedy also circumscribes courts’ initial role in most cases and provides the baseline-setting role to which Fuller argues courts should limit themselves when dealing with highly polycentric issues. Rather than setting direct policy through substantive interpretation of section 26, the courts instead establish the ground rules for a procedure by which the parties themselves, assisted by civil society, can develop the specific policies required to provide access to adequate housing.

---

204 Id.
205 Compare supra Part IV.B, with supra Part III.B.
206 Sturm and Gadlin, supra note 6, at 6.
207 See Sturm, supra note 15, at 1446 (suggesting that “[c]onsent decrees reached through processes that conform to the deliberative model may satisfy basic requirements of legitimate judicial intervention”).
208 Occupiers of 51 Olivia Road v City of Johannesburg (5) BCLR 475 (CC) at para. 10 (S. Afr.).
209 See Fuller, Forms and Limits, supra note 4, at 398.
210 See Bone, supra note 14, at 1318 (arguing that in a case challenging the constitutionality of the conditions at a special-needs school, “Fuller would have had little
The public reporting requirement plays several important roles. First, by requiring the state to develop a complete record that will be the basis for potential judicial review it ensures that, in failed engagements, courts will have the information necessary to develop the process itself in ways that protect public values.

Second, because the parties know that they must develop a record that a court may ultimately review for compliance with procedural and substantive obligations, the public reporting requirement creates a strong incentive for engagements to incorporate these public values throughout the process. Both sides will be looking toward a potential endgame that involves representations to a court and will want to be able to demonstrate that their actions and proposals serve the broader values of the right.

This requirement also emphasizes the role of reasoned arguments in the process that is a central legitimating characteristic for both Fuller and Fiss. Fuller asserts that adjudication gives institutional expression to reason because “a decision which is the product of reasoned argument must be prepared itself to meet the test of reason.”211 As a result, “issues tried before an adjudicator tend to become claims of right or accusations of fault.”212 The public reporting requirement gives the parties incentive to make reasoned arguments and claims of rights because they know those arguments may be assessed by the courts and certainly will be subject to analysis and critique by the public at the end of the process.

This aspect of the engagement process comes with the cost of eliminating the confidentiality that many argue is an important feature of ADR processes and is necessary to avoid position-based bargaining.213 The potential for position-based bargaining is real, but disclosure of the engagement process is critical to protecting the legitimacy norms associated with litigation. Confidentiality also may be less important in this setting than in others.

First, one of the principal benefits of confidentiality in private disputes is the opportunity to avoid public disclosure of the terms of the settlement itself. But the policies that result from successful engagement will be public in any event, thus eliminating this potential concern.

Second, the need for potential court oversight is crucial to ensuring that municipalities engage seriously and also to providing the opportunity for public critique of the results. Others have argued for limited disclosure in court-connected

difficulty with ordering new procedures [for the provision of new facilities and personnel], but he certainly would have worried about the judge deciding on the facilities and personnel . . . ”).

211 Fuller, Forms and Limits, supra note 4, at 366–67.
212 Id. at 369.
213 See, e.g., Maureen A. Weston, Checks on Participant Conduct in Compulsory ADR: Reconciling the Tension in the Need for Good-Faith Participation, Autonomy and Confidentiality, 76 IND. L.J. 591, 633 (2001) (“Confidentiality in ADR is popularly viewed as crucial to the effectiveness of ADR and to participants’ willingness to use such procedures.”).
mediation to protect against potential abuse of ADR processes. In the engagement context, those concerns are heightened. The potential for bargaining disparity between the parties is much greater. Moreover, a public process and the potential for judicial review of that process are necessary to make the remedy both effective and responsive to the public values at stake.

In the end, the hybrid nature of engagement—the oscillation between a party-controlled process and court direction—permits the process to operate across public and private spheres in a way that combines the flexibility of ADR methods that Fuller emphasizes and, at the same time, draws on the legitimizing characteristics of adjudication. Rather than creating a parallel process, as Sturm and Gadlin describe, engagement remains tied to the courts and permits periodic court intervention across multiple cases.

Successful engagements will never reach the courts, but many engagements certainly will. Those cases will allow courts to provide guidance on how the engagement process should be structured and what substantive outcomes are constitutionally permissible. These decisions will then serve as guidelines for future engagements, thus creating a multi-level remediation process that Sturm and Gadlin emphasize is necessary for individual resolutions to become precedents in other cases. Rather than operating independently of courts, however, engagement ties the process back to courts and creates the opportunity to address the results of failed engagements and tweak the process to deal with problems they raise.

1. Successful Engagements

But what happens with successful engagements? Cases where the process works and the settlement never reaches a court directly implicate Fiss’s concern that private interests may trump public values, because there is no independent review of the result. As an initial matter, in City of Johannesburg the court recognized Fiss’s concern and emphasized that “[i]t will not always be appropriate for a court to approve all agreements entered into consequent upon engagement.”

214 See id. at 594.

215 See Melvin Eisenberg, Participation, Responsiveness, and the Consultative Process, 92 Harv. L. Rev. 410, 430 (1978). In this essay, published as a commentary on (and in the same issue as) The Forms and Limits of Adjudication, Melvin Eisenberg suggests that a hybrid combination of “good faith negotiation based on . . . relevant legal principles” and connected to the possibility for court review may be “an optimum form of ordering” in public law cases with polycentric dimensions. Id. Although Eisenberg argues that the need for this hybrid approach risks undermining the moral force of adjudication as defined by Fuller, it can be argued that the hybridity can instead preserve the legitimacy of adjudication by relying on ADR processes to deal with the polycentric dimensions of the dispute.

216 Occupiers of 51 Olivia Road v City of Johannesburg (5) BCLR 475 (CC) at para. 30 (S. Afr.).
dispute comes to court. Each engagement will take place in the “shadow” of the possibility of litigation and court involvement, thus building in an incentive to incorporate these values at the start.

More important, several features of the process the Constitutional Court described track the requirements Sturm identifies as central to a legitimate outcome for Fuller and Fiss and allow for incorporation of the features of the public norm creating ADR process Sturm and Gadlin describe. To begin with, the court’s demand that the state involve civil society organizations in the engagement process protects the participation principle Sturm identified. Civil society organizations active on housing issues will have broader perspectives and will understand how the results of individual negotiations may affect the broader policy landscape. Thus, these groups can negotiate for policy changes that extend beyond the individuals involved in the specific engagement. At the same time, these groups can help alleviate the disparity in bargaining capacity between the municipality and vulnerable populations.

The public reporting requirement, even absent direct review by a court, also increases participation at a broader level by permitting any interested group or individual to assess (and criticize) the result of an individual engagement and even the process that led to it. Although public assessment is unlikely to affect the outcome of what would be at that point a complete engagement, it can serve the same refining function as court review in the longer term by offering suggestions for improving the process or reasons why the next engagement should be more protective of the values of the specific right. In this way, the public report generated in a successful engagement serves some of the same functions as the reasoned decision of a court: announcing the terms of the agreement and stating how those terms are consistent with the requirements of the right involved.

The broad participation that engagement entails—bringing outside groups into the process and permitting review and critique by the general public after the fact—creates the “multi-partiality” that Sturm and Gadlin argue can substitute for judicial independence in the ADR context. Rather than a single, independent judge assessing the outcome for consistency with public values, engagement encourages a principled result protecting those values through the incorporation of multiple, experienced actors, as well as the integration of a range of perspectives and a critique of the process through public reporting.

---


218 See discussion supra Part III.B.

219 See Sturm, supra note 15, at 1410.

220 Sturm and Gadlin suggest that an additional benefit to including repeat players like civil society organizations in the negotiation process is increased legitimacy: “[P]anels of independent physicians and community advocates operating as third party intermediaries carry substantial weight and bring legitimacy to the process of conflict resolution and systems intervention.” Sturm & Gadlin, supra note 6, at 48.

221 See id. at 56.
More important, the public reporting requirement, combined with another, previously underutilized constitutional provision, creates a powerful tool for using individual engagements to establish broad norms across engagements. In combination, these features create incentives for the municipality and the affected residents to pay consistent attention during the negotiation process to the public norms these rights enforce. They also can be used to make the results of single engagements repeatable, where appropriate, in the same way that Sturm and Gadlin argue individual conflict resolutions should be adapted to establish broader policies within an organization.222

Each of the socioeconomic rights, including section 26, requires “progressive realisation” of the right over time.223 Progressive realization requires, at a minimum, that the state cannot decrease the level of benefit provided without substantial reason.224 The public reporting requirement means municipalities must document their engagement efforts and results.

The reasonableness and resources limitations in section 26225 give the government flexibility to argue that a different level of benefit or modified program is more appropriate to address the circumstances in a dispute, but it must support those claims in the engagement process. During engagement, residents and civil society groups will have the opportunity to argue for alternatives not considered by the government that may adequately address the basis for that modified program or diminished benefit.

The City of Johannesburg’s inner-city regeneration project provides a simple example of this effect. Continuation of this project will require a significant number of additional evictions. Following the court’s decision, the city must develop a plan for engaging in a systematic way with the residents of the buildings targeted for redevelopment. Those residents can now point to the settlement agreement the city reached as a benchmark for their cases. The city will have to justify offering lesser accommodation to these other residents as a matter of good negotiation practice and because the city knows that the engagement process is subject to court review if it breaks down.

Up until this point, the court’s controversial decision to link the substantive right to the internal limitations clause when interpreting socioeconomic rights has

---

222 See id. at 54.
224 See Gov’t of the Republic of S. Afr. v Grootboom 2000 (11) BCLR 1169 (CC) at para. 45 (S. Afr.). The court adopted the interpretation of “progressive realisation” put forth in paragraph 9 of the general comment to Article 2.1 of the International Covenant on Economic, Social and Cultural Rights: “Moreover, any deliberately retrogressive measures . . . would require the most careful consideration and would need to be fully justified by reference to the totality of the rights provided for in the Covenant and in the context of the full use of the maximum available resources.” Id.
225 Id. at paras. 19, 21.
been largely criticized as diminishing its force.\footnote{See, e.g., Bilchitz, supra note 138, at 9 (arguing in the context of section 27’s right to health care that the court’s approach “is guilty of failing to integrate ss 27(2) and (1): it focuses the whole inquiry on s 27(2) without providing a role for s 27(1)”)} But through engagement, the progressive realization qualifier has the potential to act as a powerful tool for giving section 26 substantive content—content, however, that is developed by the state itself through negotiation with affected citizens and civil society groups rather than mandated by courts.

For this structured approach to engagement to work effectively, it must include not only a process for documenting the individual engagements, which the court recognized, but also a publicly accessible repository of the reports. This will give municipalities the longitudinal information they will require to make engagement more than a merely ad hoc process. More important, access to the results of engagements is necessary for individual engagements to serve as potential precedents for future engagements and also to allow for public assessment of the results.

Civil society can play a key role here as well. Groups that are consistently involved in engagements can help develop appropriate record-keeping guidelines for each engagement. Those same groups can then press the government to make those records publicly available and, in turn, use those same records as the basis for negotiations in subsequent engagements.

Over time, then, engagement can establish a generalizable, but still flexible, set of process norms and substantive requirements for section 26 that can be applied and modified in later cases. The public reporting requirement permits broad access to these norms. The involvement of civil society helps ensure the government does not depart from these norms in later cases, and that modifications are justified by the particular circumstances in each case. The continuing obligation to engage in socioeconomic rights cases creates an incentive for the government to incorporate these norms into social policy development more generally, and also to consult with civil society groups when developing those policies.

Equally important, the potential replication of engagements is largely controlled by the political branches themselves, thus enhancing the democratic legitimacy of the process. After City of Johannesburg, municipalities must develop structured, long-term approaches to engagement and build plans for engagement from the start of any redevelopment process. This forces municipalities to pay consistent attention to the requirements of section 26 because they must consider its implications from a long-term perspective in any development plan. It also gives municipalities control over the timing and circumstances of engagement.

Going back to the Johannesburg example, if the city takes seriously the obligations the court has described, it should develop something like an “engagement department”—or at least a structured engagement review process—that will consider what aspects of its redevelopment plans might require engagement under section 26. The city can then decide whether a particular
building or set of buildings requires redevelopment, and it can assess the potential cost of engaging with residents in light of the result in City of Johannesburg. Rather than responding in an ad hoc way to individual lawsuits over section 26, the city can choose which interventions to make in light of its overall budget and policy priorities.

2. Bad-Faith Engagements

There is, of course, always the possibility that a municipality will decide to engage in bad faith. This could happen in one of several ways. First, the city could simply refuse to engage. Second, it could go through the motions of engagement without offering any serious concessions to residents and without seriously justifying that refusal beyond simply saying it lacks the resources. A more subtle alternative is for the city to go through the motions of engagement using facially reasonable excuses for refusing to provide additional benefits, such as increased demands from other sectors or legitimate—but pretextual—differences in the situations between one set of residents and another.

Those cases likely will end up in court. The question then will be how a court should deal with this kind of recalcitrance once it is identified. In the first two scenarios—outright refusal or obvious bad faith—it makes sense for the court, at least initially, to order further engagement with additional court control. Exercising this option would create a process that tracks Sturm’s deliberative model even more directly.227 The court could find the municipality liable for violating section 26, not for the substantive reason that it failed to provide sufficient benefits, but on the procedural ground that it failed to engage in good faith. Because of the public reporting requirement, the court will have the benefit of a record from which it can assess the process itself and order the parties to return to the bargaining table with specific modifications. These could include a broad range of options, including appointing a specific civil society group (or some other person) to act essentially as a mediator. Or, less dramatically, the court could order more inclusive consultations with groups that were either excluded or not sufficiently included.

The option to modify the specific process also gives the court an opportunity, without directly interpreting section 26, to signal the parties in general terms through informal discussions or formal statements on the record what it thinks section 26 might require in a particular situation and in light of previous engagements. Just as in Sturm’s deliberative model, this would give the court an additional opportunity to reinforce the public values at stake.228 The parties would then return to the bargaining table but with a more specific, court-directed procedure and substantive guidelines.

227 See discussion supra Part III.A.
228 Sturm, supra note 15, at 1429–30 (describing the court’s role prior to negotiation as “outlining[ing] for the participants the characteristics of the process” and “inform[ing] the participants as to the standards it will use to assess the adequacy of the proposed remedy”).
There is good reason to think that this kind of repeated engagement under specific court pressure will work, even where a municipality initially ignores its obligations. This is precisely what happened in *City of Johannesburg*: the specific pressure the Constitutional Court exerted worked to force the city to engage more seriously. In a left-handed compliment, the court deliberately “commended [the city] for the fact that its position became more humane as the case proceeded through the different courts, and for its ultimate reasonable response to the engagement order.”229 Later in the judgment, the court itself highlighted the fact that direct court pressure to engage and report back was “the deciding factor” that moved the city to make concessions: “The deciding factor in this case in my view was that engagement was ordered by this Court, and the parties had been asked to report back on the process while the proceedings were pending before it.”230 But the court went on to emphasize its preference that engagement “take place before litigation commences unless it is not possible or reasonable to do so because of urgency or some other compelling reason.”

What happens in the most extreme cases where the process breaks down completely, or the municipality persists in refusing to offer a reasonable program? As an initial matter, the court has structured the engagement process to avoid this result. In particular, the possibility for more direct court control over the renewed engagement process just described will ensure that a complete breakdown is only possible if the municipality takes an extremely hard line over time.

Nonetheless, when faced with repeated refusals to engage seriously (or simply a good-faith impasse), the court may ultimately have to substantively interpret section 26 and order the municipality to take specific action. On the one hand, this looks like the failed result that Sturm argues forces the court to choose among the less legitimate alternatives in the structural reform litigation context.232 But the key

229 *Occupiers of 51 Olivia Road v City of Johannesburg* 2008 (5) BCLR 475 (CC) at para. 28 (S. Afr.). This quote echoes the court’s rhetorical move in *Minister of Health v. Treatment Action Campaign* 2002 (1) BCLR 1022 (CC) at para. 129 (S. Afr.), a health care rights case where, faced with public calls for defiance of any court order by senior government officials, the court noted in the judgment that the government had always complied with court orders in the past and there was no reason to expect a different result in that case. As in *Treatment Action Campaign*, the court in *City of Johannesburg* appears to be trying to coax the government into taking seriously its constitutional responsibilities in the future rather than truly commending it for having done so in the past. See generally Brian Ray, *Policentrism, Political Mobilization and the Promise of Socioeconomic Rights*, 45 *Stan. J. Int. Law* 151 (2008) (arguing that the court’s general approach in socioeconomic rights cases is targeted at accustoming the political branches to take seriously its obligations under these rights).

230 *Occupiers of 51 Olivia Road v City of Johannesburg* 2008 (5) BCLR 475 (CC) at para. 30 (S. Afr.).

231 *Id.*

232 Sturm, *supra* note 15, at 1439. Sturm contemplates the possibility of breakdown of the deliberative model and notes that “[a]lthough a court-imposed remedy may undermine norms of remedial legitimacy, the court’s adoption of this role derives support from the parties’ failure to reach agreement.” *Id.*
characteristics of engagement—its extended nature, the public reporting requirement, and the political control it creates—leave the court in a much better position to direct specific policy changes for several reasons.

First, the court now has a substantial record of proposals and counter-proposals, including detailed justifications by the municipality. This enhances the informational base from which it is making the substantive interpretation, thereby reducing the institutional competence concerns. In addition, once this process has developed over time, the court also will have the benefit of records of other engagements and can consider the similarities and differences of this particular situation.

Second, as discussed earlier, the city will have decided in advance to engage with this particular set of residents. Presumably, this choice will have taken into account the larger context of the city’s other responsibilities and priorities, and therefore ordering expenditures will not be as potentially disruptive as it would if the case were brought directly by residents. In other words, the political control that engagement creates should give the court more flexibility to order some substantive benefit where warranted, because it was a political decision to target these groups. For the same reason, any court-ordered relief to a specific set of residents will avoid the queue-jumping problem that the court has been concerned with—and has used as a reason to avoid ordering individual relief in other cases.

In this way, engagement opens the door to making section 26 individually

---

233 See id. (noting that, in the case of a court-imposed remedy following deliberation, “[t]he court’s remedial decision . . . will be informed by the data, diversity of perspective, and reasoning produced by the deliberations”).

234 These kinds of failed engagements also will give the court the opportunity to develop broad substantive guidelines for these rights to guide future engagements. Sandra Liebenberg has argued that the Constitutional Court’s reasonableness review could evolve to incorporate more substantive interpretation without abandoning the context-sensitivity and flexibility that are its key advantages. See Liebenberg, supra note 153, at 325, 328. The engagement remedy would permit periodic interventions of the kind Liebenberg advocates while still putting the emphasis on the development of standards through the political process. Engagement gives the court the opportunity to review those standards over time and, where it finds policy choices either sufficient or deficient, to articulate the constitutional values that either support the chosen policy or require change. In this way, the court retains the public value-creation role the Fiss argues it must play, Fiss, Foreword, supra note 5, at 29, while still relying on civil society and the political branches to come up with the policies to enforce those values.

235 See, e.g., Gov’t of the Republic of S. Afr. v Grootboom 2000 (11) BCLR 1169 (CC) at para. 92 (S. Afr.) (“This judgment must not be understood as approving any practice of land invasion for the purpose of coercing a state structure into providing housing on a preferential basis to those who participate in any exercise of this kind. Land invasion is inimical to the systematic provision of adequate housing on a planned basis.”); Modder E. Squatters v. Modderklip Boerdery 2004 (8) BCLR 821 (SCA) at para. 23 (S. Afr.) (discussing the “queue-jumping” problem identified in Grootboom).
D. Establishing and Extending Engagement

The success of engagement as a hybrid remedy for implementing socioeconomic rights depends on the willingness of the court and municipalities to apply it consistently over time and across different cases. Consistent application of engagement combined with refinement of the procedure is necessary for several reasons. First, as this Article argues, a critical mass of successful engagements is required to establish the precedents that create the potential for norm development outside of court decisions. Second, success of the process depends to a large degree on the development of structured mechanisms for engagement by government, civil society groups, and citizens. These mechanisms will develop only if the court remains committed to ordering engagement.

There are early indications that the Constitutional Court will continue to use engagement as a remedy, and also that it might extend it beyond eviction cases, even to rights other than housing. A few cases provide encouraging evidence that the court is willing to press for development of the remedy along the lines just described, and thus establish it as an important enforcement mechanism. But the prospect of extension also raises questions over the precise mechanisms necessary to make engagement effective and also the limitations of the remedy.

1. Engagement in Eviction Cases

The first case, *Residents of Joe Slovo Community Western Cape v Thubelisha Homes*, which the Constitutional Court decided as this Article goes to press, looks much like *City of Johannesburg*. Joe Slovo involved attempts by the City of Cape Town to evict and relocate thousands of residents of an informal community north of Cape Town along the N2 Highway, the major north-south corridor leading into Cape Town. The city targeted the residents for eviction and relocation as part of a broad redevelopment plan involving development of new housing to replace the existing informal settlements. Large numbers of residents protested the plan by demonstrating in and around the community and by opposing the plan in court.

---


238 *Id.* at para. 25.

239 *Id.* at paras. 25–26.

240 *Id.* at para. 34.
The high court issued a decision permitting the relocations to proceed and
denying the residents relief shortly after the Constitutional Court issued its opinion
in *City of Johannesburg*. The judge made passing reference to the engagement
requirement and, in a parenthetical aside, found that the numerous meetings the
City Council held with residents, “along with multiple averments in the court
papers of meetings and/or consultations that were held with the residents of Joe
Slovo indicate[d] that there was a sufficient amount of engagements . . . regarding
this matter.”

The residents appealed directly to the Constitutional Court, which, in a
somewhat surprising move, accepted the direct application rather than requiring
the residents to first go through the Supreme Court of Appeal. Several of
the same groups that were active in organizing the residents in *City of Johannesburg*
submitted amicus curiae briefs to the Constitutional Court, arguing specifically that
the City of Cape Town failed to adequately engage with the residents. The court
granted those groups permission to present this issue at oral argument. During
the hearing, Deputy Chief Justice Dikgang Moseneke, in a move reminiscent of
*City of Johannesburg*’s interim order, suggested the amici were correct by
“interven[ing] to suggest that the parties talk to each other and advise the court on
a ‘just and equitable’ solution.”

That attempt to use engagement in a similar fashion as the court had done in
*City of Johannesburg*—to resolve the substantive issues without direct court
involvement—failed. The court issued its decision on June 10, 2009. The
decision consists of five different decisions and spans 221 pages. All five opinions
agree that neither section 26 nor the Prevention of Illegal Eviction from and
Unlawful Occupation of Land Act 19 of 1998 (“PIE Act”) protects the residents
from eviction. And all five also concur in the final order that directs how the
evictions are to take place. Early reaction has been generally critical of the
court’s refusal to find in favor of the residents. Sandra Liebenberg describes the

---

241 *Thabelisha Homes v Various Occupants* Case No. 13189/07, at para. 85 (S. Afr.),
available at http://www.constitutionalcourt.org.za/Archimages/12328.PDF.

242 Id. at para. 24.

243 See Pearlie Joubert, ‘*It’s Our Duty Not to Be Silent,*’ MAIL & GUARDIAN ONLINE,
surprise move, the Constitutional Court gave the community permission to challenge the
ruling and approach it without going through the Supreme Court of Appeal.”).

244 See id. (noting that “[t]he community law centre of the University of the Western
Cape and the Centre on Housing Rights for Evictions were admitted as friends of the court,
in support of the resident’s [sic] right to be properly consulted before being evicted”).

245 See Further Directions by the Chief Justice 15 August 2008, available at
http://www.constitutionalcourt.org.za/Archimages/12788.PDF.


247 *Residents of Joe Slovo Cmty. W. Cape v Thabelisha Homes* CCT 22/08, [2009]
ZACC 16 (S. Afr.), available at http://www.constitutionalcourt.org.za/Archimages/
13625.PDF.

248 Id. at para. 4.

249 Id. at para. 5.
result as “the largest judicially sanctioned eviction of a community in South Africa’s post-apartheid period.” Pierre de Vos, while acknowledging that “the judgment shows a genuine concern for the plight of the Joe Slovo residents,” criticizes the substantive holdings as failing to expect “the state to act in an honest manner and to cater also for the most vulnerable and poor members of a well-established community whose area is to be upgraded.”

It is beyond the scope of this article to deal with the full implications of the court’s latest decision, but several preliminary observations are possible. First, it is evident that the court is committed to using engagement. While it was unsuccessful in convincing the parties to engage before judgment, the court incorporated engagement into the order itself. The unanimous order requires the parties to “engage meaningfully with each other” on the date for relocation of the residents, the timetable for that relocation and “any other relevant matter.” On the one hand, this confirms that the court meant what it said in City of Johannesburg: that engagement truly is a constitutional requirement.

More important, the court has begun to establish more specific guidelines in this case that will begin the broader project of defining the engagement procedure and developing specific precedents that might be applied in future eviction cases. Three points are of particular note. First, the court’s use of engagement as part of a remedial order after a substantive judgment is notable in itself as an extension of engagement to a new procedural situation. Pierre de Vos observes that this, combined with the court’s retention of supervisory jurisdiction in the case, may in fact be the court’s way of forcing the government to engage with the residents over a revised plan and could be a backdoor mechanism for creating the pre-move engagement that never occurred.

Second, in the engagement order, the court assumed much more control over the negotiation agenda than it did in City of Johannesburg or in the Mamba decision discussed below. Among other things, the court ordered the parties to determine “[t]he exact time, manner and conditions” of each relocation and also “the precise temporary residential accommodation” for each resident. This is the kind of ratcheting up of court control that takes advantage of the engagement’s iterative potential and can help make the remedy more effective over time.

Finally, the court has retained supervisory jurisdiction and ordered the parties to report back on the results of engagement on each issue. Like the more specific agenda, this also represents a potentially important innovation in the engagement

---

252 Joe Slovo CCT 22/08, at para. 7.
253 See Occupiers of 51 Olivia Road v City of Johannesburg 2008 (5) BCLR 475 (CC) at para. 15–16 (S. Afr.).
254 See de Vos, supra note 251.
255 Joe Slovo CCT 22/08 at para. 7.
256 Id.
process by demonstrating to lower courts the possibilities for enhancing court control while still giving parties the power to determine the substantive result.

Despite these procedural innovations, the ultimate result in *Joe Slovo*, like the *Mamba* decision discussed next, illustrates that engagement has real limits. It also highlights the risk in relying exclusively on this kind of indirect remedy without ever developing the substantive requirements of section 26 directly. Mark Tushnet (citing Cass Sunstein’s discussion of constitutional development in post-Soviet countries) argues that “[c]oupling strong rights with weak remedies, particularly when those remedies are rarely deployed . . . may be a formula for producing cynicism about the constitution.”

Both *City of Johannesburg* and *Joe Slovo* can be viewed as important opportunities for the court to develop the relatively modest enforcement of section 26 that it began in *Grootboom* into something more substantial. The success of engagement in *City of Johannesburg* deflected criticism about the court’s refusal to deal with the substance of section 26. But the critical reaction to the court’s approval of the evictions in *Joe Slovo* shows that the court must occasionally back up remedies such as engagement with more direct enforcement for those remedies to remain effective.

2. The Limits of Engagement

The second case, *Mamba v Minister of Social Development*, provides an example of the Constitutional Court extending engagement to a new context, specifically closure of refugee camps by the Gauteng government. The court’s actions in the case provide additional evidence of its commitment to this remedy and also illustrate a creative extension of engagement along the lines I have suggested. But the final result—complete refusal by the provincial government to meaningfully engage—demonstrates the limits of the remedy.

The wave of violent xenophobic protests that began in Johannesburg and extended to Durban and Cape Town in May 2008 displaced tens of thousands of people in South Africa. The Gauteng provincial government formally declared a state of emergency and used disaster relief funds to establish several temporary camps to provide security and shelter for victims of the violence. A range of national and international organizations provided logistical and financial support to the relief effort, and they began working with Gauteng and other provincial and local governments on medium- and long-term solutions for the camp residents.

---

257 See TUSHNET, supra note 1, at 252.
260 See id.
261 See id.
These efforts included repatriation of foreign nationals, asylum in another country, and reintegration in South Africa.\textsuperscript{262}

The Gauteng government initially set a deadline of July 31, 2008, for closure of the temporary camps, later extending that deadline to August 15, 2008.\textsuperscript{263} The Consortium for Refugees and Migrants in South Africa (CoRMSA), an umbrella organization that includes several of the organizations active in the relief efforts, pressed the government to delay the closures until it developed and published a reintegration plan.\textsuperscript{264} When the government ignored their request and publicly announced plans to move forward with the closures, CoRMSA and the Wits Law Clinic—housed at the University of Witwatersrand—sued in the Pretoria High Court seeking an injunction to prevent the closures and to require the government to develop and “communicate” a “comprehensive reintegration strategy that adequately protects the rights of all.”\textsuperscript{265}

On August 12, 2008, the High Court rejected the refugees’ arguments in a two-page order, finding that the government had not violated any rights and had no obligation to continue to provide accommodation.\textsuperscript{266} Two days later, the refugees sought direct access to the Constitutional Court, even though the court was in recess.\textsuperscript{267} Despite the recess, the court held an emergency hearing on the application the following Monday, August 18, and issued an interim order on August 21.

The order temporarily prohibited complete closure of the camps, subject to certain limitations, including the right to consolidate shelters and to deport illegal immigrants. But it also required, in language nearly identical to the City of Johannesburg engagement order, that the parties,

engage with each other meaningfully and with all other stakeholders as soon as it is possible for them to do so in order to resolve the differences and difficulties aired in the application in the light of the values of the

\textsuperscript{264} Press Release, CoRMSA, CoRMSA Calls, supra note 263.
\textsuperscript{265} Press Release, CoRMSA, Wits Law, supra note 263.
\textsuperscript{266} Mamba v Minister of Soc. Dev. No. 36573/08-rm, at 2 (S. Afr.), available at http://www.constitutionalcourt.org.za/Archimages/12791.PDF.
\textsuperscript{267} No Decision on Refugees’ Application, IOL ONLINE, Aug. 14, 2008, http://www.iol.co.za (enter title of article in the article search box).
Constitution, the constitutional and statutory obligations of the respondents and the rights and duties of the residents of the shelters.268

Paragraph 5 of the order specified that the engagement should include not only the refugees but also the United Nations High Commissioner for Refugees, the Jesuit Refugee Services (groups active in supporting the camps), and “[o]ther role players.”269 The order also required the refugees, the Gauteng government and the City of Johannesburg to report the results of the engagement by September 12 and set a hearing on the engagement for September 16.270

In spite of pressure from the National Parliament and facilitation offers by the South African Human Rights Commission,271 the Gauteng government adopted a narrow reading of the order and refused to negotiate a reintegration plan with CoRMSA and others. Instead, the provincial government read the August 21 order as requiring merely that it keep the refugees and the groups listed in the order apprised of its continued plans for closing the camps.272

After the September 16 hearing, the Constitutional Court postponed a hearing on the full application until November, but issued another interim order requiring the government to maintain the camps and ordering continued engagement under the guidelines of the August 21 order.273 CoRMSA hailed the court’s decision as an “opportunity for government, together with civil society and the broader humanitarian assistance community” to address the reintegration problem.274

This optimism proved unwarranted. The Gauteng government persisted in its narrow view of the court’s orders and began closing the camps without consulting...
on a reintegration plan. On October 16, recognizing that the case was effectively moot, CoRMSA withdrew the application and dismissed the case.

What does the failed engagement in Mamba suggest for the prospects of this remedy more generally? As an initial matter, it highlights the political nature of engagement and its dependence on the willingness of the political branches to take the process seriously. Put differently, the flexibility and enhanced legitimacy offered by engagement’s hybridity carries with it the cost of losing the direct court control and specificity of traditional remedies.

The history of more direct court interventions in the United States suggests that this cost may not be as great in practice as it appears in theory, because governments often find ways to resist even very specific court orders. But the ambiguity inherent in engagement arguably provides more opportunity for resistance, and potentially allows it to come at a lesser political cost. The Gauteng government’s narrow reading of what engagement required exemplifies this weakness.

By the same token, this “failed” engagement suggests that the remedy must be developed in ways that both encourage the political branches to take it seriously, and also permit stronger court intervention where appropriate. One potentially critical difference between City of Johannesburg and Mamba was the High Court’s initial injunction in City of Johannesburg, which forced the city to stop its eviction program. It was much easier for the Constitutional Court to order the parties to negotiate without the threat of imminent eviction, and the city had greater incentive to take the negotiations seriously, having already stopped the program. By contrast, the High Court in Mamba permitted the government to proceed with the closures, and, although the Constitutional Court’s August 21 order arguably required the government to maintain the camps, it was sufficiently qualified that the government could adopt the narrow reading that it did, thus eliminating any incentive to engage meaningfully.

These disparate results suggest two things. First, when ordering engagement in the context of an ongoing dispute, courts—at least in the short term—should be more willing to enjoin the challenged activity. Second, if, in the face of failed engagements, courts demonstrate a willingness to order substantive remedies, over

---

275 See South African Press Association, Refugees Turfed out of Their Tents, IOL Online, October 7, 2008, http://www.int.iol.co.za (enter the title of the article in the article search box). In a sadly ironic parallel, the government used the same “red ant” security groups to evict the refugees as the City of Johannesburg had used in the City of Johannesburg evictions. Id. (“‘They didn’t tell us they [the red ants] were going to do it,’ said Jane Senga, originally from Angola.”).


time that will create an additional incentive for the government to engage seriously or risk losing the political control the remedy creates.

More broadly, the Mamba result reinforces the need to develop engagement as a structured, long-term process, rather than relying on it solely as an ad hoc remedy during an ongoing case. As the court emphasized in City of Johannesburg, engagement will work best where it is built into the policy development process from the start.278

Pushing engagement back into the policy development process raises difficult questions, i.e., at what level and at what point in the process does it make sense to engage? It is relatively easy to imagine extending engagement to other cases, like Mamba or Joe Slovo. And the court’s willingness to order engagement in Mamba demonstrates that engagement is now an integral part of its enforcement arsenal. But Mamba arguably failed because the engagement came too late and with too little direct court involvement.

Assuming that the Mamba order can be read together with City of Johannesburg as requiring provincial governments to build engagement into broader immigration and refugee policy, what would that look like? This is an extremely complicated question to answer in the abstract with the camps now closed and the refugees dispersed. Is the Gauteng government required now to “engage” with the same NGO’s on emergency plans for responses to potential future crises? Must it engage with the refugees who remain in South Africa? Must the reports of those engagements be made public? All of these questions suggest difficult lines must be drawn for engagement to develop into an effective remedy.

It is beyond the scope of this Article to precisely describe where those lines should be drawn. But it is apparent that engagement should be viewed not as a single event limited to a particular point in time, but instead as a process that should be incorporated on several levels. To use the Mamba example, the Gauteng government should build engagement into whatever steps it takes to develop policies both to prevent the xenophobic violence that led to the Mamba camps and to respond to similar situations in the future. Doing so not only makes good policy sense, but also provides the government with a record to which it can point if the policies it ultimately adopts are challenged in the future.

This by itself suggests one (fairly weak) incentive for the government to engage outside of a pending or threatened lawsuit: the ability to create a record to which it can point if and when a challenge arises. But engagement should not be limited to broad consultation in the policy development process, or else it will become nothing more than the kind of good governance standard that the Constitutional Court’s critics argue diminishes the force of socioeconomic rights. If similar protests occur in the future, then the government should consider specific engagements when developing its disaster-relief plan and during the windup process that was at issue in Mamba.

278 See Occupiers of 51 Olivia Road v City of Johannesburg 2008 (5) BCLR 475 (CC) at paras. 14–15 (S. Afr.).
There are, of course, limits to how often and to what extent individual governments can incorporate engagement either into the policy development process or at the implementation phase. At some point, the government can legitimately decide that the process is complete—regardless of whether all parties are happy with the result. As discussed above, where there is disagreement, parties are free to seek court relief, and further engagement, if warranted, should be an option for the courts. But it is also conceivable that the court will simply determine that no further engagement is required and, based on the public record, that the policy adopted by the government is a reasonable one. The precise scope of the government’s responsibility to engage will never be completely clear, but some clarity in the form of general guidelines and precedents will develop over time. It is less important to establish clear guidelines for when and at what points to engage, than it is to emphasize the need for consistent engagement over time.

V. CONCLUSION

As a hybrid process that operates somewhere between pure ADR and pure adjudication—and, indeed oscillates between those extremes—engagement offers a novel and potentially important tool for enforcing socioeconomic rights. That tool falls somewhat short of the call by the Constitutional Courts’ critics for full-fledged judicial interpretation and enforcement, but the same features that make engagement something less than strong court enforcement also enhance its legitimacy.

Michelman’s description of the effects of a hypothetical “constitutionally declared right of everyone to the enjoyment of social citizenship” illustrates this point.279 Michelman points out that such an ambiguous right “would leave just about every major issue of public policy still to be decided.”280 But, he argues, such a right could still have important effects on democratic decision making:

Its maximum (but maybe not trivial) effect on democratic decision making (the courts being kept away) would be a certain pressure on the frame of mind in which citizens and their elected representatives would approach the sundry questions of public policy always waiting to be decided. In Rawlsian language, the point of naming social citizenship a constitutional right would be to give a certain inflection to political public reason. Across a very broad swathe of public issues, such a naming would amount to a demand that those issues be approached as occasions for exercises of judgment—which choice will be conducive to the social citizenship of everyone, on fair terms?—rather than as invitations to press and to vote one’s own naked interests and preferences.281

279 See Michelman, supra note 2, at 34.
280 Id.
281 Id.
Engagement operates in a similar, but somewhat more coercive, fashion to force the political branches to pay consistent attention to section 26 (and possibly other socioeconomic rights in the future) whenever they develop social policy. Rather than removing the courts (as in Michelman’s description), engagement gives them a specific, but, at least initially, limited role that incorporates the legitimacy norms emphasized by Fiss and Fuller while leaving substantive policy-making largely in the political realm.
DISABILITY-SELECTIVE ABORTION AND THE AMERICANS WITH DISABILITIES ACT

Dov Fox∗ and Christopher L. Griffin, Jr.†

This Article examines the influence of the Americans with Disabilities Act (ADA) on affective attitudes toward children with disabilities and on the incidence of disability-selective abortion. Applying regression analysis to U.S. natality data, we find that the birthrate of children with Down syndrome declined significantly in the years following the ADA’s passage. Controlling for technological, demographic, and cultural variables suggests that the ADA may have encouraged prospective parents to prevent the existence of the very class of people it was designed to protect. We explain this paradox by showing the way in which specific ADA provisions could have given rise to demeaning media depictions and social conditions that reinforced negative understandings and expectations among prospective parents about what it means to have a child with a disability. We discuss implications for antidiscrimination law and prenatal testing policy.

TABLE OF CONTENTS

I. INTRODUCTION ................................................................................................. 846
II. DISABILITY RIGHTS AND EXPRESSIVE EXTERNALITIES .................................. 848
    A. The ADA and Prenatal Testing .................................................................... 848
    B. Legal Theories of Expressivism ................................................................... 853
    C. Economic Theories of Expressivism ........................................................... 855
    D. An Expressivist Account of the ADA.......................................................... 857
III. THE ADA AND DISABILITY-SELECTIVE ABORTION....................................... 859
    A. The “Uplifting ADA” .................................................................................. 860
    B. The “Disappointing ADA” ......................................................................... 864

† © 2009 Christopher L. Griffin, Jr., Yale Law School, J.D. expected 2010; University of Oxford, M.Phil. 2004; Georgetown University, B.S. 2002. We thank Mary Ann Call for her excellent editorial suggestions. An earlier version of this Article was awarded the 2007–08 Margaret Gruter Prize by Yale Law School for the best paper on how ethology, biology, and related behavioral sciences deepen our understanding of law. For helpful comments and conversations in our work on this project, we are grateful to Bruce Ackerman, Jack M. Balkin, Melissa Bjelland, Robert Burt, Guido Calabresi, Glenn Cohen, Drew S. Days, III, Teun Dekker, John J. Donohue, III, Owen M. Fiss, Mitch Kaneda, Seth Lazar, Rosalind McDougall, Jennifer Nou, David Powell, Richard Re, Max Schmeiser, Reva B. Siegel, Michael Ashley Stein, Lucy Wang, David Wasserman, Victoria Weatherford, and Kenji Yoshino. A special debt of gratitude is owed to Christine Jolls, who read multiple drafts and provided especially constructive insights.
I. INTRODUCTION

Former Alaska Governor Sarah Palin’s decision to give birth to a child with Down syndrome put prenatal testing and children with disabilities in the national spotlight.¹ This Article examines the decision that thousands of other parents have made to continue a pregnancy after receiving a positive test for Down syndrome.² Our empirical and sociological analysis locates a likely feature of reproductive decision making in an unlikely place: antidiscrimination law. Specifically, we measure and explain what impact, if any, the Americans with Disabilities Act (ADA) had on the incidence of selective abortion on the basis of Down syndrome in the United States from 1989 to 2002.

The Article proceeds in four parts. Part II draws on expressive law theory to explain the connection among the ADA, social attitudes toward people with disabilities, and the practice of disability-selective abortion. Expressive law theory claims that law communicates messages about social norms in addition to assigning penalties for noncompliance.⁴ Where communities are connected by shared cultural understandings, legal statements can change social norms in ways

---

¹ Palin’s son Trig appeared on news broadcasts and the cover of People magazine. See People, Cover Story: Sarah Palin’s Family Drama: The Republican VP Candidate Is Raising a Baby with Down Syndrome and Coping with Her Teenage Daughter’s Pregnancy—While Running for Office, People, Sept. 22, 2008; see also Julie Bosman et al., In Palin, Families of Disabled Children See a Potential White House Friend, N.Y. Times, Nov. 3, 2008, at A19 (describing how Palin could be seen on her campaign plane “leaning over Trig, cooing and feeding him from a bottle”); John Fritze, A Spotlight for Special Needs Parents Hope Palin Lifts Awareness of Down Syndrome, USA Today, Sept. 8, 2008, at A6 (discussing the disability issues that parents hoped Palin would address as the mother of a child with Down syndrome); Jennifer Steinhauser & Amy Harmon, Parents of Special-Needs Children Divided over Palin’s Promise To Help, N.Y. Times, Sept. 7, 2008, at A25 (reporting how “the camera panned to her baby, Trig” during an early campaign appearance).

² See, e.g., Amy Harmon, Prenatal Test Puts Down Syndrome in Hard Focus; The DNA Age: In Their Shoes, N.Y. Times, May 9, 2007, at A1.


that are not reflected in the plain meaning of the text. These changes derive from the expressive effects of law. Specifically, antidiscrimination law can generate unexpected social costs for behavior that is completely unrelated to the law’s protections and regulations. We call these costs expressive externalities.

Part III presents twin hypotheses about how the ADA’s expressive externalities might have impacted the incidence of Down-selective abortion. Each theory relies on the process by which the ADA shaped parental attitudes toward people with disabilities. The first hypothesis suggests that the ADA strengthened parents’ understandings of people with disabilities as equal members of the political community. On this account, the ADA—by enhancing the rights and opportunities of people with disabilities—made having an affected child a less daunting prospect. In turn, parents would have been encouraged to continue a pregnancy after receiving a positive test for fetal disability. The second hypothesis suggests a less heartening expressive function. This theory would suggest that the ADA triggered collateral effects that increased negative exposure to people with Down syndrome. Specifically, off-putting social contact with and critical media coverage of people with Down syndrome after the ADA’s passage could have convinced parents that children with disabilities were burdensome or defective.

Part IV tests the relative strength of these hypotheses through regression analysis. Our empirical model suggests that the second hypothesis has greater relevance for Down syndrome than other prevalent conditions. The statistical results, which control for demographic characteristics and prenatal care histories, suggest that the birthrate of children with Down syndrome fell by between 13 and 18 per 100,000 relative to the pre-ADA period of 1989–1990.\(^5\) Additional data on amniocentesis screenings confirm that Down syndrome birthrates declined at the same time that prenatal testing rates remained constant.\(^6\) In fact, more children with Down syndrome were born when amniocentesis screening rates dropped.\(^7\) These findings suggest that parental choices, rather than other exogenous explanations, were responsible for the decline in Down syndrome births in the mid-1990s.

In Part V, we consider the plausibility of these findings by examining a range of legal, material, economic, technological, social, familial, and medical factors that might reasonably have informed prenatal testing and selective abortion for Down syndrome. The connection between civil rights legislation and private decision making yields insight into the far-reaching influence of antidiscrimination law on remote social behaviors. Scholars and lawmakers should consider not only the unintended social meanings that state action can convey, but also the possible effects these meanings can confer on those whom the law was designed to protect.\(^8\)

---

\(^5\) See infra Part IV.
\(^6\) See infra Part IV.
\(^7\) See infra Part IV.
\(^8\) For a discussion on the paradoxical effects of antidiscrimination law in the context of employment, see Ian Ayres & Peter Siegelman, The Q-Word as Red Herring: Why Disparate Impact Liability Does Not Induce Hiring Quotas, 74 Tex. L. Rev. 1487, 1497–98 (1996); Paul Oyer & Scott Schaefer, Sorting, Quotas, and the Civil Rights Act of 1991:
The Article concludes with policy recommendations and proposals for further research.

II. DISABILITY RIGHTS AND EXPRESSIVE EXTERNALITIES

A. The ADA and Prenatal Testing

The ADA is a sweeping federal antidiscrimination law that combines some of the most celebrated and influential provisions of earlier civil rights legislation, including Titles II and VII of the Civil Rights Act of 1964, several provisions of the Fair Housing Act of 1968, and section 504 of the Rehabilitation Act of 1973. The goals of the 101st Congress in enacting the ADA were no less than revolutionary. The ADA affirmed the rights of people with disabilities to work, seek redress for discrimination, and exercise the responsibilities of citizenship. The law has three parts: Title I prohibits disability discrimination in employment; Title II guarantees access to public services; and Title III guarantees access to public accommodations for persons who are (or are perceived to be) impaired in a legally recognized way.

---


10  Id. § 3601.
12  See Robert L. Burgdorf, Jr., The Americans with Disabilities Act: Analysis and Implications of a Second-Generation Civil Rights Statute, 26 Harv. C.R.-C.L. L. Rev. 413, 493 (1991) (“A prominent innovative feature of the ADA is the unprecedented scope of activities and entities it regulates.”).
16  For an itemized analysis and legislative history of the ADA’s various provisions, see generally Burgdorf, supra note 12.
18  Id. § 12131.
19  Id. § 12181.
20  Individuals can qualify for protection in three ways under the ADA’s definition of disability. See id. § 12102(2)(A)–(C). First, a person can show evidence of a known physical or mental condition or impairment that “substantially limits one or more . . . major life activities.” See id. § 12102(2)(A); 29 C.F.R. § 1630.2(i) (1997). Major life activities include “caring for oneself, performing manual tasks, walking, seeing, hearing, speaking,
The ADA’s antidiscrimination record has been mixed. On the one hand, ADA specifications for mass transit have led to widespread accommodations on breathing, learning, and working.” 29 C.F.R. § 1630.2(i). A “substantial” limitation on the major life activity of working prohibits the individual from performing a class of job activities compared with an average person with comparable skills and training. See id. § 1630.2(j)(3)(i). The second way people can qualify for ADA protection is to “have a record of” such a condition. 42 U.S.C. § 12102(2)(B). A record of disability means an individual “has a history of, or has been misclassified as having, a mental or physical impairment that substantially limits one or more major life activities.” 29 C.F.R. § 1630.2(k); see also id. pt. 1630, app. § 1630.2(k). The third way to qualify as “disabled” under the ADA is to give evidence of being “regarded as” having a substantially limiting condition. 42 U.S.C. § 12102(2)(C). This “regarded as” prong might include, for example, an asymptomatic individual who is denied an employment opportunity because of a supervisor’s or coworker’s negative attitudes or behavior toward that individual’s supposed psychiatric illness, predisposition for cancer, or human immunodeficiency virus (HIV). See Abbott v. Bragdon, 107 F.3d 934, 939 (1st Cir. 1997) (“We hold unhesitatingly that HIV-positive status, simpliciter, whether symptomatic or asymptomatic, comprises a physical impairment under the ADA.”), rev’d on other grounds, 524 U.S. 624 (1998); see generally Michael D. Moberly, Perception or Reality? Some Reflections on the Interpretation of Disability Discrimination Statutes, 13 Hofstra Lab. L.J. 345, 348 (1996) (reviewing perceived disability case law arising under state regulations).

The Americans with Disabilities Act Amendments Act of 2008 (ADAAA) was signed into law by President George W. Bush on September 25, 2008. 42 U.S.C.A. § 12101 (2008). The impact that this Act will have on courts’ interpretation of disability under the ADA, as of the time of this Article’s publication, remains to be seen. But there is reason for optimism. The ADAA affirms that the definition of disability should be construed in favor of broad coverage to the extent permitted by the terms of the ADA. The ADAAA retains the ADA’s basic definition of “disability” as an impairment that substantially limits one or more major life activities, a record of such an impairment, or being regarded as having such an impairment. Id. § 12102. But it changes the way that these statutory terms should be interpreted in several ways that are intended to make it easier for an individual seeking protection under the ADA to establish that he or she has a disability within the meaning of the ADA. See id. § 12101–02.

Among the most significant changes, the ADAAA states that: an episodic impairment or one in remission is a disability if it would substantially limit a major life activity when active; expands the definition of “major life activities” by including a range of activities (e.g., walking, reading, bending, and communicating) and bodily functions (e.g., “functions of the immune system, normal cell growth, digestive, bowel, bladder, neurological, brain, respiratory, circulatory, endocrine, and reproductive functions”); clarifies that mitigating measures other than “ordinary eyeglasses or contact lenses” shall not be considered in assessing whether an individual has a disability; changes the definition of “regarded as” so that it no longer requires a showing that the employer perceived the individual to be substantially limited in a major life activity; and provides that individuals covered only under the “regarded as” prong are not entitled to reasonable accommodation. See id. § 12102.

bus, rail, and other forms of public transportation. Employment provisions have effectively eliminated disability inquiries and pre-employment physical examinations. Telecommunications mandates have resulted in the establishment of a nationwide relay system enabling use of telephone services by those with hearing or speech impairments. On the other hand, voluntary noncompliance and private discrimination remain largely unchanged. Scholars disagree about the ADA’s impact on the employment and poverty rate among individuals with disabilities. This Article departs from traditional antidiscrimination analysis by focusing on heretofore unexamined effects of the ADA on selective abortion of fetuses with Down syndrome.

Prenatal testing can disclose single-gene mutations associated with more than 400 conditions. These conditions range from severe physical and mental

---


23 Id.

24 Id.


27 We discovered no previous scholarship exploring the effect of law on selective abortion. For empirical studies of selective abortion, see generally Allyson J. Peller et al., Trends in Congenital Malformations, 1974-1999: Effect of Prenatal Diagnosis and Elective Termination, 104 OBSTETRICS & GYNECOLOGY 957 (2004); Kenneth B. Schechtmann et al., Decision-Making for Termination of Pregnancies with Fetal Anomalies: Analysis of 53,000 Pregnancies, 99 OBSTETRICS & GYNECOLOGY 216 (2002).

28 Advances in reproductive biotechnology will bring tests for many more conditions and will reduce the time, inconvenience, invasiveness, medical risk, and cost for pregnant women. Wylie Burke, Genetic Testing, 347 NEW. ENG. J. MED. 1867, 1871 (2002).
impairments such as anencephaly and Tay-Sachs (untreatable diseases that result in childhood mortality), to relatively minor conditions such as color blindness or polydactyly, a heritable trait involving an extra finger. Some parents seek information about fetal disability for reassurance or to prepare for children with physical or mental impairments. However, prenatal tests are requested primarily to identify and abort fetuses that carry genetic mutations associated with disability. Since no cure or treatment is available for most anomalies identified at the prenatal stage, prospective parents who receive a positive test for fetal conditions such as spina bifida, muscular dystrophy, sickle cell anemia, fragile X, and Down syndrome are faced with the decision either to terminate the pregnancy or to bring to term a child with a known impairment.

Down syndrome, also called trisomy 21, is a genetic condition caused by the presence of an extra chromosome. The phenotypic characteristics associated with the condition include mild to severe mental retardation and distinctive physical features.

There are legal, social, and medical reasons why Down syndrome provides a fitting case to test the possible effects of the ADA on selective abortion. First,

---

33 For certain genetic disorders, such as those that cause high levels of cholesterol, detection before birth significantly increases the promise of effective treatment and prevention. In rare cases, early diagnosis can preclude years of unpleasant monitoring procedures, such as the regular colonoscopies required for diagnoses of familial adenomatous polyposis, a condition that frequently remains concealed for several years. See Norman M. Ford, *Ethical Aspects of Prenatal Screening and Diagnosis*, in Genetics and Ethics: An Interdisciplinary Study 197, 198–200 (Gerard Magill ed., 2004) (describing various procedures for prenatal screening and diagnosis).
35 In utero surgery may one day be used, albeit in an exceedingly small number of cases, to repair neural tube defects (myelomeningoceles). See A.J. Eggink et al., *In Utero Repair of an Experimental Neural Tube Defect in a Chronic Sheep Model Using Biomatrices*, 20 Fetal Diagnosis & Therapy 335, 339 (2005).
38 Id. at 794.
unlike many other conditions with a significant biological etiology.\(^39\) Down syndrome typically counts as a disability under the ADA.\(^40\) Second, Down syndrome is a condition with which most people are at least somewhat familiar, in that they have seen a person with Down syndrome, could identify the physical features that typically accompany the condition, or are aware the condition is associated with a certain level of mental retardation.\(^41\) Down syndrome is for this reason exceptional among genetic disorders such as hemophilia, color blindness, or Klinefelter syndrome, insofar as many people will have formed, or would have reason to form, attitudes and opinions about the lives of people with Down syndrome.\(^42\)

Third, persons with Down syndrome continue to encounter prejudice and discrimination because of their condition. For example, the legislative history of the ADA recounts the case of a child with Down syndrome being banned from a New Jersey zoo because a zookeeper feared that her presence would frighten the chimpanzees.\(^43\) Finally, among all genetic anomalies, Down syndrome is the genetic condition that occurs most frequently within the U.S. population,\(^44\) the condition for which Americans test most routinely before birth,\(^45\) and the condition


\(^{40}\) See, e.g., 28 C.F.R. pt. 36, app. B (2008) (“It would violate this section to establish exclusive or segregative eligibility criteria that would . . . limit the seating of individuals with Down’s syndrome to only particular areas of a restaurant.”). But see Littleton v. Wal-Mart Stores, Inc., No. 05-12770, 2007 WL 1379986, at *3–4 (11th Cir. May 11, 2007) (holding that individuals with intellectual and developmental disabilities—what the court calls “mental retardation”—do not count as “disabled” for purposes of the ADA).


on the basis of which fetuses are aborted at the highest rate. The American College of Obstetricians and Gynecologists (ACOG) set thirty-five as the recommended age at or above which pregnant women should receive an amniocentesis because, at that age, the average woman’s chance of having a baby with Down syndrome equals her chance of having a miscarriage associated with the procedure.

**B. Legal Theories of Expressivism**

The past decade has witnessed a renaissance in expressivist scholarship, both in traditional legal circles and in the law and economics community. We suggest a potential shortcoming in this literature—namely, a failure to appreciate the negative effects of law on social attitudes and practices orthogonally related to the law’s intended regulatory context. This subpart argues that antidiscrimination law exerts a more wide-ranging influence than expressivist scholars have recognized. We propose a broader understanding to account for these effects and introduce a concept called “expressive externalities,” which captures the social costs of behavior the law does not regulate. The concept of expressive externalities serves two functions in this Article. First, it informs our hypothesis about how the ADA influenced Down-selective abortion. Second, it bolsters the explanatory power of expressive law theory more generally.

Following the lead of Lawrence Lessig, legal scholars including Cass Sunstein, Dan Kahan, Elizabeth Anderson, and Richard Pildes have argued...
that laws may have subtle yet meaningful effects on public norms and associated behaviors. According to an expressivist account of law, regulations can discourage undesirable practices in ways that transcend the expected effects of punitive sanctions for noncompliance. Laws also influence the populace by “making statements” and “moralizing” through language “designed to affect social norms and . . . ultimately to affect both judgments and behavior.” Expressivist theories assume that legal discourse takes place in settings shaped by social norms and cultural meanings. Without these conditions, the only “messages” communicated would be the rules codified in the law. However, when a new law interacts with existing normative frameworks, it can generate changes in social behavior “because either the law induces [individuals] to change their tastes . . ., or [because the law] creates a fear of bearing social sanctions . . ., or because of pressure brought to bear upon them through societal sanction.” The point is that social responses to official sanctions are informed by more than the ordinal relationship between pain and gain.

The confluence of legal statements and social norms can drive a wedge between the intentions of lawmakers and the consequences of the laws they enact. Given shared understandings about the legal meaning within a particular

---


53 Id.


55 Sunstein, *supra* note 50, at 2024.


57 Sunstein, *supra* note 50, at 2025.


social context, laws can convey conceptions of social status even if they do not explicitly target a disfavored group. Nor need recipients of a law’s message accept it as valid for the law to have expressive power. Under the Jim Crow regime, for example, racial segregation of public facilities sent a clear message that “blacks [were] untouchable, a kind of social pollutant from which ‘pure’ whites must be protected.” Even if a majority of the white population rejected this notion, collective awareness “that segregation laws express contempt for blacks” was sufficient to ensure “these laws constitute blacks as a . . . stigmatized caste.” The historical and cultural meaning of racial segregation sent the message that blacks were not worthy of equal concern and respect.

Expressivist theories enrich our understanding of the interaction between law and social norms. However, expressivist scholars often fail to appreciate how law can transmit meanings that affect social attitudes and practices wholly unrelated to the law’s substantive provisions. Individuals often internalize messages at a tangent to the mandates embedded in the text. Such collateral effects are highly plausible given the complexity of modern social interaction. We draw on this ordinary observation to support a potentially extraordinary proposition: The Americans with Disabilities Act may have had a significant impact on the incidence of disability-selective abortion, a social practice completely beyond the scope of the text or legislative history.

C. Economic Theories of Expressivism

Expressivist theories sometimes draw on the language of microeconomics to articulate the relationship between legal intervention and preference formation. If we think of political and social interaction among citizens as a repeated game in which the players choose strategies over time, then collective norms and values emerge when a sufficient number of players pursue the same strategies. However,
modern advances in game theory stress that information and payoff structures can lead to multiple equilibria, in which case no unique convention emerges. In one of Cooter’s models, society achieves behavioral stability (in economic terms) when the relative proportions of individuals respecting norms and those flouting convention equalize payoffs between the two groups. He depicts a simple graph with a downward sloping payoff to “rightdoers” and “wrongdoers” as the percentage of wrongdoers in society increases. At a unique point, no individual can gain by altering her level of compliance; this is the economic definition of equilibrium. But self-enforcement mechanisms, through which individuals “punish” each other for deviant behavior, can raise this equilibrium payoff. Under conditions in which “some people respond [to a new law] by devoting more resources to upholding it,” the introduction of expressive law analysis can change our understanding of payoff structures and, in turn, change the incentive equilibrium for individual behavior.

Patricia Funk’s recent analysis of mandatory voting laws in Switzerland is one of the few empirical studies of the expressive function of law. She identifies the expressive effect of a legal change by detecting a time series break, an unmistakable, systematic shift that differentially affected Swiss geographical districts. However, Funk studies contributions to public goods rather than

---

68 Some theorists suggest that a particular equilibrium, known as a focal point, may eventually dominate. For a description of how focal points provide clues for coordinating behavior, see Thomas C. Schelling, The Strategy of Conflict 57–58 (1980).
70 Id.
71 Id. at 590.
72 Id. at 593.
74 Funk tests whether the “natural experiment” caused by the repeal of mandatory voting laws in four Swiss Cantons (districts) affected voter turnout. Id. at 138. The reported regression results indicate a 6 to 10 percentage point decrease in participation following abolition. Id. Given that the monetary penalty for noncompliance was less than one dollar, Funk interprets “the observed turnout drop . . . as support for a certain . . . ‘expressive effect of law’” because the “legal statement that citizens should vote apparently caused certain citizens to follow [it], most likely out of civic duty or fear of social sanctions.” Id. at 138–39. Funk ultimately fails, however, to identify an expressive effect from the change in voting law. Her empirical model does not follow the convention of comparing treatment and control groups before and after the legal change. Instead, she only estimates whether Cantons with mandatory voting laws at any time had higher turnout rates than those without such measures. The lack of a temporal element in her model means that Cantonal differences could have already existed before the laws changed. If they did, then the law had no effect at all, much less an expressive one.
antidiscrimination law, which should be of equal empirical interest to expressivist scholars.

D. An Expressivist Account of the ADA

A basic introduction to externalities is instructive for our discussion of expressive externalities. Economists have long recognized that private costs and benefits can diverge from those borne by society at-large. Because individuals often fail to account for the side effects of their activity, they externalize costs and benefits on other members of society. Classic examples in the marketplace are manufacturing waste that pollutes another’s land downstream (negative externality), and spillover effects from obtaining a college education (positive externality). This principle applies to legal activity as well; legislators may fail to consider the full range of behavioral influences transmitted by their statements and actions.

Based on extensive interviews with parents and genetic counselors, Dorothy Wertz identified eight factors that determine parents’ “revealed preferences” for childbirth rather than disability-selective abortion: (1) guilt over rejecting a child with a disability; (2) the quality of life from infancy through adulthood for a child with a disability; (3) whether the pregnancy is “wanted,” independent of fetal disability; (4) optimism that children with disabilities will be cured or treated of the disabilities with which they are born; (5) spousal compromises; (6) financial constraints; (7) risk; and (8) the effect of a child with disabilities on existing children. Each of these factors could be thought of as a component in the parents’ “reproductive utility function,” i.e., the mental calculus that determines the relative appeal of terminating a pregnancy as opposed to bringing a fetus with disabilities to term. Certainly none would be directly influenced by the ADA’s explicit provisions. Any relationship between the law and the reproductive utility function would have to emerge from something like the expressivist channels discussed in Parts III(A) and III(B).

Prospective parents might be aware of the ADA, for example, and internalize a message that is unintended by or even inimical to the law’s textual provisions. Or they may be unaware that an antidiscrimination law has passed, but the law’s effect on social expectations and interactions might manifest itself in one or more of the factors in the reproductive utility function. For example, when the ADA called national attention to the plight of people with disabilities, it may have led doctors—at least those engaged in genetic counseling sessions with prospective parents—


patients—to emphasize the long-standing but newly publicized stereotypes and discrimination faced by people with disabilities. The resulting effect on attitudes about the expected quality of life for children with disabilities (Wertz’s second factor) might be strong enough relative to the other seven factors to persuade many prospective parents to terminate the pregnancy.

Conventional expressivist accounts miss a critical element that this Article tries to correct. According to the traditional logic of expressivist legal theory, people tend to have minimal knowledge or awareness of a law’s mandates. A person might nevertheless internalize a message that is not stated in a statute’s text and have that message shape her norms or behavioral responses. Such awareness of the law, however, would be limited to those who have a stake in the law’s application and enforcement. In tort law, for example, an insurance company might transmit information to policyholders based on the company’s understanding of how the law shifts incentives, perhaps in ways that increase moral hazard. Our conceptual innovation here is to recognize that expressivist changes to social customs and attitudes can emerge without knowledge of a specific law but still because of the law’s existence. This scenario is most likely to occur when the affected social practices are completely unrelated to the substantive provisions of the law.

To understand the possible mechanism at work, consider a law intended to “clean up” professional baseball by forbidding the use of a new performance-enhancing drug. Reproachful congressional hearings that precede the law’s enactment cause fans to question the sport’s integrity. After a federal ban on the performance-enhancing drugs, major league ticket sales decline by 40 percent nationwide. Retail sales of products endorsed by high-profile athletes identified as drug users also fall dramatically. Without knowing more, one might suppose that the drop in attendance and merchandise sales was caused by an expressive effect of the anti-drug law. The government expressed a judgment about drug use in professional baseball, and the receipt of this message persuaded some fans not to attend professional games or buy merchandise associated with discredited athletes. These behavioral responses were caused not by the explicit threat of legal sanctions against drug use, but rather by the implicit message that baseball is a tainted activity.

Now imagine that, after the drug ban’s passage and the ensuing drop in attendance, parents who are altogether unaware of the law begin to steer their children away from Little League baseball. They may interpret the declining attendance at professional ballparks as a social verdict that baseball is boring or behind the times. Within a few years, reduced youth talent prevents some

---

78 See, e.g., Sunstein, supra note 50, at 2026 (arguing that “people support a law . . . because they believe that it is intrinsically valuable for the relevant ‘statement’ to be made”).

79 See, e.g., Joseph E. Stiglitz, The Contributions of the Economics of Information to Twentieth Century Economics, 115 Q.J. ECON. 1441, 1453 (2000) (describing moral hazard as a scenario in which, “if individuals are insured against a risk, they have inadequate incentives to take actions to avoid the risk”).
secondary schools from fielding baseball teams, and their districts’ athletic associations remove baseball from the schedule. This second scenario also fits within the expressivist paradigm, but a key feature differentiates the influence on youth sports from the decline in professional ticket and merchandise sales. The law was never intended to censure, regulate, or otherwise affect nonprofessional sports teams or athletes. Nevertheless, it is highly plausible that the professional drug ban is the “but-for” cause of a shift in attitudes about the sport generally and the accompanying decline in high school baseball.

One way to think about these related but distinct expressivist channels is the difference between direct expressivism (on ticket sales) and indirect expressivism (on youth participation). The basis for this distinction lies in the different kinds of public response a new law generates. Although the ban on performance-enhancing drugs did not prescribe a boycott of professional baseball, the public understood the law as direct government censure of the sport. In the youth sports arena, however, waning interest was an indirect by-product of the attitudinal changes that the law communicated with respect to professional baseball. Parents who removed their children from Little League teams did so not because of their understanding or interpretation of the drug ban itself, but because of the law’s predictable effects on public views about baseball.

When a law’s passage, enactment, and enforcement give rise to significant changes in social practices unrelated to the law’s mandates, the law generates what we call an expressive externality. Recall that the economic concept of externalities focuses on costs or benefits that accrue to third parties because of another agent’s activity (downstream pollution killing fish or improved education leading to a more productive workforce). Similarly, laws can “externalize” costs and benefits on individuals, social groups, or behavior wholly outside the law’s purview.

Returning to our analysis of disability rights, we might identify the ADA’s expressive externalities on the practice of disability-selective abortion using the reproductive utility function. The ADA may have caused a sufficiently large change in the eight factors in that function to shift parental preference orderings between abortion and childbirth. While positive expressive externalities can produce interesting welfare gains, we focus on negative expressive externalities (hereinafter “expressive externalities”). An expressive externality of the ADA relevant to disability-selective abortion would exist if some side effect of the law led prospective parents to choose abortion more often than childbirth after a positive fetal diagnosis for Down syndrome.80

III. THE ADA AND DISABILITY-SELECTIVE ABORTION

Part III presents competing hypotheses about expressive externalities the ADA could have generated with respect to the practice of disability-selective abortion. These hypotheses juxtapose the ADA’s exalted ideals against its often complicated social assimilation. The first hypothesis (the “uplifting ADA”)

80 See infra Part V.
suggests that the ADA’s affirmation of social equality and disability rights discouraged disability-selective abortion by tempering negative attitudes toward people with disabilities. According to this theory, by barring employment discrimination and assuring access to public services and accommodations, the ADA conveyed to parents the promise that prospective children with disabilities would lead happy and productive lives. The second hypothesis (the “disappointing ADA”) suggests that by increasing awareness of and exposure to people with disabilities, the ADA created feelings of discomfort or animus. When the ADA was passed, the media often reported unfavorable stories about people with disabilities, which might have convinced parents that children with disabilities would not be accepted into the world as social equals. Each hypothesis makes a rival claim about the expressivist effects that the ADA could have had on social attitudes and reproductive behavior in the 1990s.

For any law to have a systematic expressive effect on social behavior, there must be some mechanism by which people internalize the law’s meaning. The hypotheses we present therefore suggest potential pathways through which this transmission could have occurred. These mechanisms do not, however, supersede the ADA’s function as the source of expressive content. Consider the proposition advanced by John Donohue and Steven Levitt in their study of abortion and crime: the legalization of abortion prevented potential criminals from existing and eventually violating the law.81 This Article connects reproduction outcomes to a legal change, namely the ADA’s passage and enforcement. Similarly, Donohue and Levitt ascribe birth outcomes and falling crime rates to a revolutionary 7-2 Supreme Court decision that made no mention of abortion’s possible impact on public safety.82 The primary difference between the studies is that Justice Blackmun’s opinion in Roe was crafted to communicate an explicit constitutional right to private reproductive decision making, whereas, in drafting the ADA, lawmakers were silent on the issue of selective abortion. The link we propose between the ADA and disability-selective abortion follows the same basic logic as the proposition offered by Donohue and Levitt. Our argument, like theirs, suggests that legal transformations can reach social attitudes and behaviors that have no connection to the law’s content.

A. The “Uplifting ADA”

The sweeping goals of participatory83 and distributive84 justice that the 101st Congress identified in enacting the ADA indicate that the law aimed to combat

83 See IRIS MARION YOUNG, JUSTICE AND THE POLITICS OF DIFFERENCE 173 (1990) (defining social equality as requiring the “full participation and inclusion of everyone in a society’s major institutions”). The ADA facilitates participatory justice by requiring the redesign of practices that exclude individuals with disabilities. See 42 U.S.C. § 12101(a)(8) (2006) (stating that “[t]he Nation’s proper goals regarding individuals with disabilities are
attitudinal bias as much as it sought to eliminate physical barriers. Supreme Court jurisprudence affirms the view that implicit attitudes contribute significantly to the exclusion and devaluation of people with disabilities. In *Alexander v. Choate*, the Court recognized that discrimination against people with disabilities is “most often the product, not of invidious animus,” but rather of thoughtless or indifferent attitudes. Similarly, in *School Board of Nassau County v. Arline*, the Court noted that “society’s accumulated myths and fears about disability and disease are as handicapping as are the physical limitations that flow from actual impairment.” The ADA’s emphasis on social prejudice therefore might have improved perceptions about people with disabilities.

This first hypothesis suggests that positive effects on co-workers’ attitudes toward people with disabilities in the employment sphere could have collateral effects in the reproductive sphere. It might be that by prohibiting discrimination in the workplace, the ADA counteracted disability prejudice, which relies on the assumption that individuals with disabilities are less capable than able-bodied workers who perform similar functions. In their book-length study of workers with disabilities, David Engel and Frank Munger argue that the ADA has powerfully shaped attitudes about persons with disabilities. Engel and Munger provide testimonial evidence that the ADA produces discursive shifts altering both how individuals with disabilities think about their own capabilities and also how others regard them. The law does so by recasting people with disabilities as successful individuals who need only the opportunity to demonstrate their
abilities. Consider Engel and Munger’s discussion of Barry Swygert, a man paralyzed by a spinal tumor and who relied on knowledge about ADA protections to reclaim his “ambitious” career plans and “reconstitute his identity” as a self-conscious “rights-bearer.”

Causal mechanisms between attitudes in the employment and reproductive spheres might draw on three relevant psychological theories. First, intergroup cooperation theory posits that sustained contact with disabled persons in the workplace could reduce prejudice against others with disabilities. Positive attitudes emerge under conditions of equal social status, mutual goals, sustained intimate contact, and institutional support for equality. Second, cognitive dissonance theory predicts that people change their attitudes to reduce the tension they experience when those attitudes contradict some action. On this account, positive employment experiences among people with disabilities will cause co-workers to match positive attitudes to achieve cognitive consistency. Third, self-perception theory suggests that when people have weak or ambiguous attitudes, they develop stronger attitudes to match their own behavior and the circumstances under which it occurs. Self-perception theory predicts that when employment conditions support qualified employees with disabilities, a co-worker with weak or

90 Engel & Munger, supra note 89, at 116–22.
91 Id. at 98–102; cf. id. at 73–77 (describing Raymond Militello, a man with a learning disability who regards the ADA as giving disabled persons unjustifiable preferences).
ambiguous attitudes toward people with disabilities will come to believe that his favorable behaviors reflect positive attitudes toward individuals with disabilities.98

To the extent that prospective parents are included among employees who develop more positive attitudes about the capacities and opportunities of people with disabilities, perhaps they would be more willing to bring a child with a disability to term rather than seek an abortion. It could also be that, as compliance with the ADA became routine, non-discriminatory employment practices created a positive transformation in disability norms more generally. Positive attitudes toward people with disabilities could then have seeped into the public consciousness, with effects including more positive attitudes about the prospect of raising a child with such disabilities. We do not mean to suggest that this is a complete or correct story about the ADA and disability-selective abortion. We mention it here only as one possible theory to account for the ADA’s effect on social norms and practices that we discuss in the next Part.

Indeed, social science research describes how civil rights for blacks evolved in informal, extralegal ways after the school desegregation mandates.99 People who grow up in racist cultures often absorb stereotypes that reside in their psyches and influence behavior in subtle but pernicious ways. Similarly, federal rights laws can subconsciously affect public attitudes in ways that diminish implicit biases.100 Accordingly, scholars have argued that racial antidiscrimination law was accompanied by a reduction in racial prejudice.101 However, two salient differences emerge between the civil rights movement and the disability rights movement. First, each set of events generated drastically different cues for public observation. The 1964 Civil Rights Act “was preceded by images of courageous Freedom Riders, marches, bus boycotts, lynchings . . . and Martin Luther King, Jr. . . . delivering his ‘I Have a Dream’ speech.”102 The “disability rights movement,” by contrast, “[was] not powered by such compelling imagery.”103 The social movement preceding the enactment of the ADA had no charismatic social leader, and its most visible demonstration was described as “a pathetic event [where] crippled children and others crawl[ed] up the Capitol steps.”104 Consider, further,
that President Johnson addressed a joint session of Congress—and a national television audience—before introducing the Voting Rights Act of 1965, calling his proposal a fulfillment of the Emancipation Proclamation and a matter of human rights. President Bush held a signing ceremony for the ADA on the White House lawn and compared the Act to the dismantling of the Berlin Wall. Just one year earlier, however, he had vetoed the bill that would eventually become the Civil Rights Act of 1991. The difference between Johnson’s commitment to civil rights legislation and Bush’s tepid—or even contradictory—record on disability rights, could have contributed to different public responses. The social movement and executive actions that supported disability rights legislation in 1990 were far less emotionally resonant and politically potent than the events surrounding black civil rights in the mid-1960s. This is not to say that the uplifting account might not be correct, but only that there is reason to explore an alternative hypothesis about the possible effect of the ADA on selective abortion on the basis of disability, specifically Down syndrome.

B. The “Disappointing ADA”

A less optimistic theory suggests that the ADA transmitted or reinforced negative expressive content about people with disabilities. We discuss two potential ADA-induced sources for such outcomes: 1) an increased likelihood that prospective parents would be exposed to the impairments of persons with disabilities; and 2) popular media coverage of the ADA. The “disappointing ADA” hypothesis supposes a strong relationship between the ADA and public attitudes toward individuals with disabilities. The national spotlight shone brightly on persons with disabilities after the law’s enactment, and the ADA required important changes to the workplace and public accommodations. During this time, able-bodied citizens might have become more aware of the disability “problem” through greater exposure to people with disabilities. Hanoch Livneh has identified six channels through which negative attitudes about disability may arise, three of which are particularly relevant to changes brought on by the ADA’s passage. The first channel, what Livneh calls “conscious-unconscious,” includes causes that are fully known to the observer, or those of which he or she might be completely unaware. In this scenario, increased interaction with people with disabilities in

---

109 Id. at 193.
the context of public transportation or the workplace may lead to overt animus or implicit biases.

The second channel, “past experience-present situation,” connects early childhood experiences to present situations and interactions.\textsuperscript{110} If an able-bodied person attended school with a student in a wheelchair and observed or participated in the humiliation of that student, the able-bodied person might develop negative perceptions of persons with disabilities as an adult. Finally, “internally originated-externally originated” sources arise from interaction between an able-bodied person’s demographic or personality traits and those linked to a person with a disability or to the disability itself.\textsuperscript{111} This theory captures the complex ways in which persons without disabilities view differences in physical appearance, mental capacity, and economic potential as foreign to their own identity.\textsuperscript{112} Witnessing a colleague with disabilities use a special restroom or waiting on a bus while someone in a wheelchair uses an electric ramp might generate unfavorable perceptions about how those with disabilities cause social inconvenience or how they struggle to adjust to ordinary activities. These observations may not have occurred as publicly or frequently before the ADA’s employment discrimination,\textsuperscript{113} public services,\textsuperscript{114} and public accommodations\textsuperscript{115} mandates came into effect.

Empirical work in psychology and education quantified the extent to which negative attitudes arose, especially through workplace contact. Survey data, for example, show that Americans consider mental retardation a less stable condition than depression or psychosis, where stability measures the likelihood for positive response to counseling and medication.\textsuperscript{116} As workplace dynamics shifted after the ADA’s passage, formal integration did not assure social acceptance. A 1995 study found that “although persons with mental retardation were accepted within the workplace, few were befriended outside of the work setting.”\textsuperscript{117} Thus, in direct contrast to the uplifting account in Part III(A), workplace interaction could have exacerbated already exclusionary opinions about persons with disabilities. One study published shortly after the ADA’s passage concluded that overt animus was not a necessary condition for these adverse social consequences: “[W]orkers without mental retardation initiated interactions with nondisabled co-workers three

\begin{flushright}
\textsuperscript{110} Id.
\textsuperscript{111} Id.
\textsuperscript{112} Id.
\textsuperscript{114} Id. § 12131.
\textsuperscript{115} Id. § 12181.
\textsuperscript{116} See Patrick W. Corrigan et al., Stigmatizing Attributions About Mental Illness, 28 J. COMMUNITY PSYCHOL. 91, 98 (2000).
\end{flushright}
times more often than with co-workers with mental retardation.”¹¹十八 Negative social views “were not necessarily derogatory but also were not those of equals.”¹¹十九 So-called “interaction strain” occurs when non-disabled individuals struggle with self-consciousness and appropriate reactions to persons with disabilities.¹²零 This phenomenon could also have shaped negative perceptions about disability as a result of the ADA and thereby enhanced the perceived difficulty of raising a child with disabilities.

Consider the plausible reflections of a prospective parent who has received a prenatal diagnosis for a genetic disability such as Down syndrome. Regardless of their precise origins, unfavorable cognitive associations with disability brought on by the ADA would very likely affect one’s forecast about life with or for a child with a disability. Enhanced exposure to individuals with disabilities through ADA-related mandates and accommodations could prompt unenthusiastic attitudes toward people with disabilities generally, and toward the specific prospects of one’s own child. The diminished physical or mental capacity of a child with a disability could also remind a parent of his or her own mortality and that “[a]nyone can become a person with a disability virtually in a matter of seconds.”¹²一 Such attitudes might “reflect an awareness that persons without disabilities are vulnerable to death, injury, and disease—a vulnerability most . . . are eager to forget.”¹²二 Prospective parents who can see past the disability itself may still worry about external perceptions: “[W]ill [the parents] be considered second rate by association?”¹²三 Parents may be more likely to terminate a pregnancy on the basis of disability when they fear social reproach from others who regard the choice to bear children with disabilities as negligent or irresponsible.

An empirical study of attitudes toward mothers of children with Down syndrome sheds light on the social pressure that parents in other countries face to avoid having a child with Down syndrome. British researchers Theresa Marteau and Harriet

¹¹八 Id. (citing B. Ferguson et al., Type and Frequency of Social Interaction Among Workers with and Without Mental Retardation, 97 AM. J. ON MENTAL RETARDATION 530 (1993)).
¹¹九 Id.
¹²零 See id. (citing Fred Davis, Deviance Disavowal: The Management of Strained Interaction by the Visibly Handicapped, 9 SOC. PROBS. 120 (1961); Joseph H. Evans, Changing Attitudes Towards Disabled Persons: An Experimental Study, 19 REHABILITATION COUNSELING BULL. 572 (1976)).
¹²二 Id.
¹²三 Id.
¹²五 Id. An empirical study of attitudes toward mothers of children with Down syndrome sheds light on the social pressure that parents in other countries face to avoid having a child with Down syndrome. British researchers Theresa Marteau and Harriet
A crucial feature of this social interaction narrative is that prospective parents need not even know of the ADA’s existence for the law to affect reproductive decisions. This point reflects precisely where our concept of expressive externalities departs from the traditional theory of expressive law. According to the latter, individuals must recognize specific legislative provisions before they can ascribe expressive content to them. Expressive externalities, however, arise when laws generate social meaning in domains upon which the law has no bearing and not necessarily with awareness of the causal chain. In the pollution example, a fisherman downstream from the factory need not know the identity of the polluting company or that a polluting factory was constructed miles up the river. He still incurs the cost of contaminated water that the factory owner externalized. In our baseball hypothetical, parents who removed their children from youth leagues did not have to be aware of the drug ban. They need only have sensed public rejection of professional teams. Similarly, if there is reason to believe that increased contact with people with disabilities afforded by the ADA exacerbated adverse affective attitudes toward them, then prospective parents need not have learned about the ADA for the law to have influenced their reproductive decisions. The law’s positive effect on employment and public accommodation norms could in this way “externalize” negative views about people with disabilities because the ADA led parents to encounter circumstances that gave rise to feelings of discomfort, pity, or abhorrence.

Another mechanism might account for the “disappointing ADA” hypothesis. This second explanation turns on the fact that ADA-related media stories tended to depict people with disabilities in an unfavorable light. Mainstream media outlets

Drake tracked survey participants in three countries over three years to determine how society ascribes responsibility for the birth of a child with Down syndrome. Theresa Marteau & Harriet Drake, Attributions for Disability: The Influence of Genetic Screening, 40 SOC. SCI. & MED. 1127, 1128–29 (1995); cf. Sue Hall et al., Parents’ Attributions of Blame for the Birth of a Child with Down Syndrome: A Pilot Study, 12 PSYCH. & HEALTH 579 (1997) (interviewing mothers and fathers of children with Down syndrome and finding that a significant minority blamed health professionals or the health care system for not preventing the birth of their children during the prenatal stage). Marteau and Drake interviewed more than 930 pregnant women, geneticists, obstetricians, and a general sample of men and women from Germany, Portugal, and England. Marteau & Drake, supra at 1128–29. Respondents were presented with two vignettes about the gestational histories of children who were born with Down syndrome. Id. at 1129. In the first, the mother’s screening history was the single most important factor influencing attributions of blame for the birth of a child with Down syndrome. Id. at 1129. Women declining the offer of testing were assigned overwhelmingly greater blame for the birth of a Down child than were women who were not offered tests and subsequently gave birth to a child with Down syndrome. Id. at 1129–30.

126 See supra Part III.C.
played a significant role in diffusing national and local news in the early 1990s. Editorial and journalistic decisions about whether to cover and how to frame disability rights developments shaped popular discourse and influenced public attitudes about what it means to have a disability in America. When landmark legislation such as the ADA appeared before Congress, it “needed a messenger to convey its intent to the public. That messenger was the news media.” Given extensive media coverage of the ADA’s passage and preceding legislative debates, widely circulating newspapers and magazines served as “norm entrepreneurs,” drawing on the ADA’s salience to construct new cognitive frames and public understanding about people with disabilities.

Evidence from the early 1990s suggests that the media often portrayed people with disabilities either as economic burdens or as “supercrips,” i.e., extraordinary because they function despite having a disability. The empirical work of media content analysts John Clogston and Beth Haller builds a plausible bridge between


128 See HALLER, supra note 127.


131 As Haller notes, “news stories about disability . . . can sway public opinion about . . . the cultural representations of people with disabilities in general.” HALLER, supra note 127.

the passage of the ADA and ensuing attitudes about persons with disabilities. Clogston’s model of disability coverage classifies news articles as either “traditional” or “progressive.” “Traditional” depictions describe individuals with disabilities “as dysfunctioning in a medical or economic way,” while “progressive” portrayals emphasize civil rights and cultural pluralism. Clogston’s assessment of 363 articles from the first quarter of 1990 indicated that “medical treatment and institutionalization, government and private support programs, and victimization” dominated 60 percent of disability-related stories. In addition, 55 percent of the headlines used language consistent with the traditional model, often “refer[ring] to persons with disabilities with adjectives substituting for nouns (disabled, blind, etc.).” One can glean the ADA’s influence on affective attitudes through the media from a Seattle Times profile published in 1992. That article lamented the story of a young woman with Down syndrome in light of the new law: “At 22, Cathleen Haight has entered adulthood. She yearns to do things she thinks a woman should do . . . yet her mind is that of a 10-year-old.” “The Americans with Disabilities Act . . . is expected to increase employment of the disabled,” the profile continued, “though it may be some time before the change is felt.”

Focusing more closely on the relationship between media sources and disability-related stories, Haller concluded that the business community’s prevailing view about people with disabilities colored news reporting in ways that strongly typecast them as costly to society.


134 Clogston, supra note 132, at 46.

135 Id.

136 Id. at 48.

137 Id. at 49.

138 Vanessa Ho, Cathleen Haight Wants To Live the Life of a 22-Year-Old, but She Faces Different Challenges—Special Hopes and Needs, SEATTLE TIMES, Aug. 6, 1992, at F1.

139 Id.


141 See, e.g., PHILLIP J. TICHENOR, GEORGE A. DONOHUE & CLARICE N. OLIEN, COMMUNITY CONFLICT & THE PRESS 79–80 (1980) (asserting that news media echo the views of a community’s power-holding elite); Pamela J. Shoemaker, The Communication of Deviance, in 8 PROGRESS IN COMMUNICATION SCIENCES 151, 172 (Brenda Dervin & Melvin J. Voigt eds., 1987) (stating that “journalists’ normative judgments . . . will draw and define the attention of those who control social change. The journalist acts as a surrogate judge of deviance for his or her audience members”).
disabilities. In a follow-up empirical study, Haller found that businesspeople or business groups accounted for more than half (55.2 percent) of all sources in ADA-related news and feature articles appearing in twelve major newspapers and magazines between 1988 and 1993. She also discovered that two of the top three reasons given for the ADA’s passage were architectural access (26.5 percent) and labor market opportunities (18.3 percent). Business sources argued that “[m]aking society accessible for disabled people is not really worth the cost and overburdens businesses. . . . Accessibility is not profitable.” Small and midsized operators, in particular, expressed dismay that “the government [was] largely out of touch with . . . their cost of doing business” and that “the Americans with Disabilities Act [would] cost them more and thus have a negative impact on their businesses.” Media backlash was not limited to the business community. Three years after the ADA’s passage, a national youth sports leader asked in the Chicago Tribune:

To accommodate a child in a wheelchair, are you now creating danger for the other kids? I think you are[,] . . . there’s nothing built into the Americans with Disabilities Act that protects the safety of those who aren’t disabled. The easy solution, if there is enough of a pool, is to create a separate league.

Negative media messages related to the ADA could have been the very messages received and internalized by prospective parents considering whether to give birth to a child with a disability. We offer this claim, as with the others presented in Part III, as no more than hypotheses, without further evidence or argument. Nor are the “uplifting” or “disappointing” hypotheses necessarily exhaustive or mutually exclusive accounts of the ADA’s expressivist function, if any exists. The ADA may have produced no discernible effect on reproductive behavior at all, or the uplifting and disappointing accounts could have operated simultaneously to shape attitudes and practices in complex ways that led to counterbalance or dominance by one of the competing narratives. We test these hypotheses through empirical examination in Part IV and through analytical reasoning in Part V.

142 Haller, supra note 140.
143 Haller, supra note 129, at 65.
144 Id. at 64.
145 Id. at 61.
147 Mary Hill & Andy Trees, Dream on Hold for Boy with Down’s, CHI. TRIB., May 21, 1993, at 1.
IV. THE EMPIRICS OF DISABILITY-SELECTIVE ABORTION

In Part IV, we analyze whether either of the two hypotheses in Part III finds support in empirical data on Down syndrome birthrates after a positive fetal diagnosis for the condition. Our econometric approach resembles a program evaluation for which it is standard to estimate regression coefficients from a reduced form equation. We do so here. This technique measures variation in critical variables over time and across states that might have generated variation in the Down syndrome birthrate. We study birthrate patterns not only because currently available abortion data contain little to no information about fetal characteristics, but also because medical treatment has not successfully and independently reduced prenatal disability rates. Any decline in the Down syndrome birthrate would more likely follow from increased termination rates rather than more miscarriages or medical interventions on behalf of the affected fetus. In Part IV(A), we describe the data used to test the hypotheses from Part III. Section IV(B) presents and interprets our empirical results.

Our analysis yields suggestive evidence that the ADA led to a short-term decline in the Down syndrome birthrate. By 1993, the birthrate fell by between 13 and 18 per 100,000 relative to the pre-ADA period when controlling for demographic and medical care variables. We do not find any significant effect for two lesser-known disabilities for which screening technologies also exist. When we compare birthrate changes for infants with spina bifida and cleft palate before and after the ADA’s passage, no coefficient in the fully specified model is significant. In addition, the signs switch from negative to positive between the unadjusted and fully specified regression models. The decline in Down syndrome birthrates coincided with steady amniocentesis rates and increased sharply at the same time screening became less frequent. These collective findings suggest persuasively that the ADA’s effect on reproductive decision making was in large part responsible for the mid-1990s decline in the birthrate of children with Down syndrome.

As Donohue and Levitt emphasized in their study of abortion and crime, the explanation we offer here likewise represents one of several plausible accounts. A skeptical reader might even accept our statistical findings but still not agree with our narrative premise. In fact, labor economist Daniel Hamermesh, who has used the Donohue and Levitt study in the classroom, stated: “I’ve gone over [the] paper in draft, in its printed version, at great length, and for the life of me I can’t see

---

148 Genetic tests for Down syndrome usually are performed fifteen to sixteen weeks into the gestation period. See Wylie Burke, Genetic Testing, 347 NEW ENG. J. MED. 1867, 1871 (2002). After that point, the likelihood of miscarriage is about 1 to 3 percent. See, e.g., Statistics, http://www.pregnancyloss.info/info-howcommon.htm (last visited Sept. 1, 2009).

149 Donohue & Levitt, supra note 81, at 380 (“While acknowledging that [a number of other factors] may have also served to dampen crime, we consider a novel explanation for the sudden crime drop of the 1990s: the decision to legalize abortion over a quarter century ago.”).
anything wrong with it[.] . . . On the other hand, I don’t believe a word of it.”  

We do not wish to equate what we demonstrate in this Article to what Donohue and Levitt achieved in their study, which has withstood several rounds of scholarly criticism and has been confirmed by alternative data sources over the seven years since its publication. To the extent that our empirical and conceptual analyses are sound, however, the endurance of Donohue and Levitt’s bold claim lends credibility to the facially improbable connection we propose between the ADA and selective abortion on the basis of disability.

A. Data

1. Sources

Beginning in 1968, the Centers for Disease Control and Prevention (CDC), through the National Center for Health Statistics (NCHS), collected samples of birth records from the fifty states and the District of Columbia. In the early years of the program, “[d]ata were obtained from a 50-percent sample of certificates.” However, “[s]tarting in 1972 all records were included for States that participated in the Vital Statistics Cooperative Program (VSCP).” “The number of States participating in the VSCP increased from 6 in 1972 to 46 in 1984,” and “beginning in 1985, all States and the District of Columbia participated.” The CDC data files contain information from every field on the birth certificate, including demographic information about the child’s parents and various indicators of


153 See, e.g., James M. Poterba, Steven D. Levitt: 2003 John Bates Clark Medalist, 19 J. ECON. PERSP. 181, 188–90 (2005) (citing the Donohue and Levitt study as a major factor in the American Economic Association’s awarding Levitt the John Bates Clark Medal, which is given every two years to one American economist under the age of forty who has made a significant contribution to economic thought).


156 Id.

157 Id.
maternal and newborn health. Of particular interest for our study is information on the obstetric practices of parents who chose to bring a child with a disability to term. If the ADA had some effect on the incidence of Down-selective abortion, we hypothesize that the empirical model will detect changes in the frequency of fetal diagnoses of Down syndrome. It should be noted that birth certificates before 1989 specified whether the newborn had a congenital anomaly; only after a 1989 change in the design of birth certificates did medical attendants record specific disabilities. For this reason, we use data from 1989—the first year for which prenatal screening data also appear—through 2002, the last year for which birth certificate data are available. The information contained on these birth certificates has been transcribed into electronic data files, which are accessible through the Inter-University Consortium for Political and Social Research (ICPSR).

One might object that using just one year of data to identify the pre-ADA period limits the value of any attempt at before-and-after econometric specification. There are two reasons why this objection should not be overstated. First, because it generally takes nine months to carry a child from fertilization until birth, any changes in reproductive decision making that could be attributed to the law’s passage would not appear in the Down syndrome birthrate until after approximately 1991, one year after the ADA was passed. Our analysis therefore assumes that the years 1989 and 1990 form the “pre-ADA” period. Second, and more important, any expressive externalities from the ADA would operate with some delay. The impact of “cascading” norms emerges only after sufficient social pressure builds toward a tipping point for conformity around emergent attitudes and behaviors. However, because we cannot claim with any certainty when the externality reached its peak effect, we examine changes relative to the 1989–90 period for each year from 1991 to 2002. We elaborate in Part IV(B)(1) on reasons for the break in the birthrate trend between 1990 and 1991.

2. Summary Statistics

The raw natality data between the years 1989 and 2002 include a wide array of demographic and physical characteristics for 56,068,370 births. As the summary statistics in Table 1 show, about 52 children with Down syndrome were born per 100,000 across the United States over the period 1989–2002. We also document, for comparison purposes, the spina bifida birthrate, which averaged about 25 per 100,000 over the same period. The full list of explanatory variables

---


160 In addition to recoding variables for missing or undocumented values, we remove all births to foreign residents.

161 See infra Part IV.B.2.
culled from the NCHS natality files (also appearing in Table 1) controls for factors that are unrelated to or mildly correlated with passage of the ADA and might have induced variation in reproductive outcomes.\textsuperscript{162} Research published soon after the ADA’s passage suggests potential relationships between these covariates and Down syndrome birthrates.\textsuperscript{163} First, we might expect the incidence of disability-selective abortion to be higher among white women relative to minorities because, according to one anthropological study, “about half of minority women [in New York City] [did] not keep their appointments for counseling or prenatal diagnosis.”\textsuperscript{164} In addition, some research found that “college-educated, upper-income career women [were] less willing to risk having a disabled child than women with less education and income.”\textsuperscript{165} Thus, we would expect the numerical sign on college education to be negative, implying a lower probability of Down syndrome births.

The remaining variables capture the fertility and medical histories of women choosing to give birth during our observation period. As with maternal and gestational age,\textsuperscript{166} one might expect that a greater number of preexisting children will reduce the probability that women carry a fetus with a disability to full term, especially if their living children are not impaired.\textsuperscript{167} One of the key indicators for

\textsuperscript{162} An ideal model would include information on news media reports given the empirical evidence reported by Clogston and Haller. See supra notes 127–145 and accompanying text. The utility of such data, however, seems rather low in a regression model given that the quantity of positive and negative news reports omits the essential, qualitative characteristics that Clogston and Haller also emphasize.

\textsuperscript{163} See Wertz, supra note 76, at 167–70.

\textsuperscript{164} Id. at 170 (citing Rayna Rapp, Constructing Amniocentesis: Maternal and Medical Discourses, in UNCERTAIN TERMS: NEGOTIATING GENDER IN AMERICAN CULTURE 28 (Faye Ginsburg & Anna Lowenhaupt Tsing eds., 1990); Rayna Rapp, The Power of “Positive” Diagnosis: Medical and Maternal Discourses on Amniocentesis, in CHILDBIRTH IN AMERICA: ANTHROPOLOGICAL PERSPECTIVES 103 (Karen L. Michaelson ed., 1988); Rayna Rapp, Chromosomes and Communication: The Discourse of Genetic Counseling, 2 MED. ANTHROPOLOGY Q. 143 (1988)).

\textsuperscript{165} See Wertz, supra note 76, at 170 (citing KRISTIN LUKER, ABORTION AND THE POLITICS OF MOTHERHOOD (1984); Diane Beeson & Mitchell S. Golbus, Decision Making: Whether or Not To Have Prenatal Diagnosis and Abortion for X-linked Conditions, 20 AM. J. MED. GENETICS 107 (1985)).

\textsuperscript{166} See Ralph L. Kramer et al., Determinants of Parental Decisions After the Prenatal Diagnosis of Down Syndrome, 79 AM. J. MED. GENETICS 172, 172–74 (1998) (analyzing all cases of Down syndrome at a tertiary care center from 1989–1997 with respect to maternal age, parity, gestational age, sonographic findings, insurance status, and race, and finding that when Down syndrome is diagnosed prenatally, the choice of termination is related to maternal age and gestational age, but only gestational age is a significant independent predictor of pregnancy termination).

\textsuperscript{167} See David T. Helm et al., Prenatal Diagnosis of Down Syndrome: Mothers’ Reflections on Supports Needed from Diagnosis to Birth, 36 MENTAL RETARDATION 55, 60 (1998) (discussing the perceived challenges of raising a child with a disability alongside other children).
assessing our central hypothesis is the percentage of births subject to amniocentesis. This captures trends for the most commonly used Down syndrome diagnostic test. Finally, we detail the medical care sought by women who chose to give birth to a child with Down syndrome. Presumably, the chances of detecting Down syndrome should increase (at a diminishing marginal rate) with the number of prenatal medical consultations.

B. Results

1. Graphical Evidence

Figure 1 displays two national time series: the birthrate of infants with Down syndrome and the number of births screened through amniocentesis (each per 100,000) from 1989 to 2002. The natality files include separate variables for each congenital anomaly listed on the birth certificate. These variables are multinomial, taking one of four values: “reported,” “not reported,” “not on certificate,” and “not classifiable.” We assume that the latter two categories are comparable to “no reported condition,” which appeared in the aggregated anomaly variable before 1989. Thus, we recast the Down syndrome indicator as a dichotomous one, coding only “reported” values as positive diagnoses.

168 See Nicholas J. Wald et al., Antenatal Screening for Down’s Syndrome, 4 J. Med. Screening 181, 223 (1997).

Two features of the time series stand out in Figure 1. First, the Down syndrome birthrate trend, measured on the left axis, breaks sharply after 1990, falling about 9 percent from 48 children per 100,000 to 43 in 1991. Despite small increases in 1992 and 1996, the birthrate falls to a low of 42 per 100,000 in 1995. Starting in 1998, however, the share of Down syndrome births rises and nearly reaches its pre-ADA level by 2002.

On the secondary vertical axis, we measure the number of births for which the mother underwent an amniotic screening (again, per 100,000). The trend in the late 1980s through the mid–1990s appears relatively constant, remaining at an average of 3,100 screenings per 100,000 live births. The amniocentesis rate, however, drops steadily from 1997 through 2002. This decline in prenatal screening accompanies a rise in Down syndrome births. Before 1997, however, the Down syndrome birthrate fell off in the absence of any change in screening patterns.

Figure 1 represents the foundation for the following empirical analysis. It will also help determine whether either of the hypotheses in Part III has explanatory power, and if so, which dominates. The sharp downward trend in the Down syndrome birthrate through the mid-1990s implies that the data potentially support the “disappointing ADA” theory. However, its influence was most likely a short-term phenomenon, since the birthrate rebounded between 1995 and 2002. Several factors could explain the trend reversal in 1995, including the eventual emergence of an “uplifting ADA” story or, more simply, a gradual weakening of the ADA’s
negative externality. Nevertheless, the final data point in Figure 1 lies below the 1989–1990 levels, which indicates that the initial negative expressive effect was not completely eliminated by 2002. Thus, the available data suggest that overall Down syndrome births remained lower than they would have been had the ADA not been enacted.

Second, the amniocentesis birthrate line in Figure 1 could be misleading in ways that crucially affect interpretation about selective abortion. We claim that the amniocentesis rate remained fairly constant through the mid-1990s, but the data, by definition, exclude screenings that resulted in terminations. Thus, if amniocentesis rates increased in the 1990s and abortions after Down syndrome diagnoses increased, these screenings would not appear in Figure 1 even though they occurred. However, “[a]bout 98 percent of the women who have amniocentesis will receive the . . . news that their fetuses are free of the conditions for which they have been tested. The other 2 percent will . . . confront the . . . news that their fetus has a disability.” 170 Thus, under the assumption that parents are more likely to bear children without prenatal disabilities, all else equal, deriving the amniocentesis birthrate from live births should not be confounded significantly by selective abortion practices. This assumption permits reasonable inference about changes in the overall amniocentesis rate with changes in the amniocentesis birthrate.

2. Regression Results

Having explored the general time trends of Down syndrome birthrates as well as the incidence of amniocentesis during the observation period, we now turn to multivariate regression analysis. The first set of models is based on the following equation:

\[
BR_{st} = \beta_0 + \sum_{j=1991}^{J} \beta_{j-1990} I_j + \delta Z_{st} + \sigma + \varepsilon_{st}
\]

(1).

\(BR_{st}\) is the Down syndrome birthrate in state \(s\) and year \(t\); \(I_j\) is an indicator variable for each year \(j\) (\(J = 2002\)); \(Z_{st}\) is a vector of control variables for each state-year pair, and \(\sigma\) represents state fixed effects. Since our observation period begins in 1989, the year effects are estimated relative to the pre-ADA period from 1989 to 1990. For example, the coefficient for the variable \(I_{1995}\) \((\beta_5)\) represents the change in the Down syndrome birthrate, all else constant, between 1995 and the pre-ADA years.

Before discussing the results, we explain the rationale for the empirical specification in (1). Analyses of legal interventions on social and economic

170 Rayna Rapp, The Power of “Positive” Diagnosis: Medical and Maternal Discourses on Amniocentesis, in REPRESENTATIONS OF MOTHERHOOD 204, 205 (Donna Bassin et al. eds., 1994).
outcomes often rely on difference-in-differences models. In these models, the first difference compares outcomes for “control” and “treated” observations, and the second difference usually compares the first difference before and after the law’s effective date. Had we been able to obtain individual-level data on reproductive decisions, including whether disability-selective abortion occurred, we naturally would have followed the difference-in-differences framework. The “treated” observations would include all mothers who underwent amniotic screening, and the pre-ADA period for the second difference would remain observations from the years 1989 and 1990. Because of our data limitations, which required aggregation to a larger unit of analysis (the state of residence), sensible difference-in-differences analysis options would generate sample selection bias. For example, we might restrict the data to Down syndrome births and use difference-in-differences to examine changes before and after the ADA for women who did and did not use amniocentesis. However, this method would eliminate the relative frequency of Down syndrome births to non-Down syndrome births—the very data we need to show whether the ADA affected the selective abortion rate in either direction.

Column 1 of Table 2 shows results from the most parsimonious possible model, i.e., one that calculates raw (unadjusted) differences across time. As shown in Figure 1, the Down syndrome birthrate declined immediately after passage of the ADA. Nevertheless, the year-by-year decline does not achieve statistical significance until 1993, and the magnitude of these decreases remains high through 1999. However we explain these findings, the point estimate for 1995 (-11) is astounding; such a change represents one half of one standard deviation. Columns 2 and 3 add demographic and medical controls, respectively, via the vector $Z_{st}$ as well as state fixed effects to capture unobserved heterogeneity (especially variation in religious affiliation and general health in the state population). In most cases, the absolute value of the annual estimates increases (implying larger declines) and remains statistically significant through the period 1993–97. Finally, Column 4 represents the most fully specified model captured by (1). In this version, all the coefficient estimates from 1993 through 1998 are between 13 and 18 in absolute value. Thus, relative to the pre-ADA period 1989–90, the declines in successive years represent substantial changes in the birthrates of children with Down syndrome.

Point estimates for all control variables used in the four variations on (1) are presented in the second half of Table 2. Column 4 indicates that maternal age, prenatal visits, and amniocentesis rates are all positively correlated with Down syndrome birthrates. None of these factors, however, is statistically significant. On average, married, white, and college-educated women, as well as women with more living children, are less likely to give birth to a child with Down syndrome. The negative sign on the two significantly estimated variables (race and birth

---

171 Even still, we would not have access to individual-level decisions whether to give birth or terminate a pregnancy over time. Without such information, an individual-level difference-in-differences analysis would be meaningless.
order) further validates our empirical model. That is, the negative coefficient on maternal race and birth order supports two findings: first, that white women are more likely to have access to and use screening technologies; and, second, that parents are more likely to abort a Down syndrome fetus the more children they already have.

How do these results, which examine patterns in birthrates, relate to our analysis of selective abortion practices? First, note that Columns 2–4 in Table 2 include the number of births (per 100,000) subjected to amniotic screening as a control variable. Given that this rate remained relatively constant during the years between 1992 and 1997, one would not expect a variable with minimal variation to affect estimates of the β coefficients. Once the rate of amniocentesis began to fall, though, the declines relative to 1989–90 no longer are statistically significant. We interpret this result as highly suggestive evidence that antidiscrimination law engendered an expressive effect on Down syndrome births. Controlling as well for a host of factors that likely correlate with reproductive decisions, we still find robust decreases in the incidence of newborns with disabilities after passage of the ADA while prenatal screening frequencies remained virtually invariant. Finally, if the law actually produced expressive externalities—as opposed to some other account that could better explain the observed birthrate drops—we would not expect to observe these collateral effects as soon as the law was passed.172

As Figure 1 suggested, and the regression results confirmed, the expressive externality (if it operated) was confined to the short term, from roughly 1993 through 1998. This appears not only through the lack of statistical significance after 1998 but also through the upswing in the Down syndrome birthrate beginning in the late 1990s. These findings can be explained in one of two ways, neither of which can be tested formally with the available data.173 First, the salience of negative attitudes about Down syndrome could have dissolved over time such that the expressive externality ceased to affect reproductive decisions systematically after a certain period. Second, it might be the case that positive externalities emerged and outweighed any negative expressive effects. Either way, although the Down syndrome birthrate rose after the mid-1990s, it remained below the pre-ADA levels observed in 1989 and 1990. Thus, even if the negative expressive externalities operate only in the short term, its effects may continue to have meaningful long-term effects on reproductive outcomes.

No advances in prenatal or postnatal medical therapy have exogenously reduced the rate at which children with disabilities are brought to term. Conditional upon a fetus developing Down syndrome, the only way that parents can prevent the birth of a child with disabilities is to terminate the pregnancy.174 This important fact rules out any variation in birthrates attributable to causes outside of parental choice, and controls for any preferences or genetic propensities that would

172 See supra note 159 and accompanying text.
173 We are grateful to Jack Balkin for bringing this point to our attention.
174 See Nancy J. Roizen & David Patterson, Down’s Syndrome, 361 LANCET 1281, 1282 (2003).
correlate with a higher probability of Down syndrome births. Abortion rates are not the only measure of revealed preference for raising children with Down syndrome. Parents could decide to bring a child with Down syndrome to term and then place the child in an adoption agency. Further insight can be gleaned from analysis of the rate at which biological parents chose to place children with Down syndrome up for adoption and the rate at which adoptive parents chose to adopt children with Down syndrome. We consider adoption figures in the context of socioeconomic changes, changes in adoption laws, and other confounding variables. The absence of adoption data in our study means that the observed decline in Down syndrome birthrates represents a lower bound on parental disinclinations to raise a child with this disability. The existence of the adoption alternative, however, does not complicate our empirical findings with respect to birthrates. Since adoption rates are conditional on birthrates, our data also capture some of the ADA’s effect on adoption. Parental decisions to place a child with Down syndrome up for adoption may have increased or decreased after 1990. Regardless, our findings indicate that there were significantly fewer children born with Down syndrome that could have been placed with adoption agencies.

We interpret these regression results as evidence of an expressive externality because statistically significant declines in the Down syndrome birthrate coincided with a relatively constant amniocentesis rate. Ruling out the effect of changing medical diagnostics and personal characteristics, the “residual” explanation for falling birthrates could be the cascading norm shift regarding selective abortion set off by the ADA. On the other hand, the eventual birthrate increase that occurred when amniocentesis rates fell was statistically insignificant; otherwise, there would be more support for the proposition that changes in amniocentesis rates drove changes in Down syndrome birthrates. The magnitudes of our estimated coefficients (up to 18 per 100,000) might appear insignificant in absolute terms, but declines of this size are quite substantial relative to the pre-ADA distribution of Down syndrome birthrates. The numerical significance of the post-ADA birthrate drop (13 to 18 per 100,000) relative to the pre-ADA mean (52 per 100,000) serves to underscore the importance of these results.

Although we controlled for demographic changes in race and class over the period we investigated, these factors surely influenced prenatal testing rates. In a 1980 study of genetic testing among prospective parents in the state of Georgia, for example, prenatal diagnosis was used by 60 percent of urban white women over forty, but by just 0.5 percent of rural black women over forty. See David C. Sokol et al., _Prenatal Chromosome Diagnosis: Racial and Geographic Variation for Older Women in Georgia_, 244 JAMA 1355, 1356–57 (1980); see also Rayna Rapp, _Refusing Prenatal Diagnosis: The Meanings of Bioscience in a Multicultural World_, 23 SCI., TECH. & HUM. VALUES 45, 67 (1998) (“With better access [to information about prenatal testing], middle-class women are also less able to achieve any distance from the biomedical discourse within which their own rationality is forged. Those without much privileged scientific education are most likely to reject testing altogether . . .”).

See infra tbl. 1.
If our narrative about expressive externalities and Down syndrome birthrates is plausible, then we should not expect to observe similar changes for other congenital anomalies that generated significantly less social anxiety and received much less media coverage.\footnote{As expected, Down syndrome received far more extensive media coverage during the relevant period than spina bifida, cleft lip/palate, or other genetic anomalies. A search of the New York Times, Los Angeles Times, Washington Post, and Chicago Tribune between 1988 and 1993 reveals that Down syndrome, spina bifida, and cleft lip/palate appeared in 1028, 436, and 156 articles, respectively.} In Tables 3 and 4, we amend the baseline model to account for variation in the birthrates of spina bifida\footnote{Spina bifida is a neural tube disorder involving an incompletely formed spinal cord. The condition results in varying degrees of paralysis and is sometimes associated with cognitive problems. Hope Northrup & Kelly Volcik, Spina Bifida and Other Neural Tube Defects, 30 CURRENT PROBS. PEDIATRICS 317, 327–29 (2000).} and cleft palate,\footnote{Cleft palate is a genetic abnormality in which the two plates that form the roof of the mouth are not fully connected. The condition is usually correctable through surgery, but may cause problems with feeding, speech, and socialization. Francesco Carinci et al., Genetics of Nonsyndromic Cleft Lip and Palate: A Review of International Studies and Data Regarding the Italian Population, 37 CLEFT PALATE-CRANIOFACIAL J. 33, 35–36 (2000).} two other high-frequency congenital abnormalities.\footnote{The remaining conditions reported in the natality data are anencephalus, hydrocephalus, microcephalus, “other” central nervous system anomalies, heart malformations, “other” circulatory conditions, rectal atresia, tracheo conditions, omphalocoele, “other” gastrointestinal anomalies, malformed genitalia, renal agenesis, “other” urogenital anomalies, polydactyly/syndactyly/adactyly, club foot, diaphragmatic hernia, “other” musculoskeletal anomalies, “other” chromosomal anomalies, and “other” congenital anomalies. See Natality Data, Public-Use Data Files, http://www.cdc.gov/nchs/products/elec_prods/subject/natality.htm (last visited Sept. 1, 2009).} These data provide a relatively clean comparative test for whether social or medical salience, which we posit is a necessary condition for expressive externalities to emerge, motivated the decline in the Down syndrome rate. In Table 3, Column 1, the unadjusted annual differences suggest that negative expressive externalities might have generated similar effects for children with spina bifida. Although the coefficient estimates are not as large (in absolute value) as with Down syndrome, the results between 1998 and 2002 are either statistically significant or marginally insignificant. The significance of these results vanishes, however, when we include the full set of control variables in Column 4. No annual point estimate is statistically significant in that specification. More important, the Column 1 estimates were so fragile that most negative coefficients became positive with the introduction of controls. In Table 4, no coefficient estimate for cleft palate is estimated with enough precision to echo the expressive effects detected for Down syndrome. Together, Tables 3 and 4 indicate that the case of Down syndrome was uniquely affected by the ADA’s passage.

It may be worth trying to distinguish parental attitudes about the prospect of children with mental disabilities as opposed to physical disabilities. One reason to think the reproductive calculus varies depending on whether the disability in
question is predominantly mental or physical in nature relates to the labor market. Children with predominantly physical disabilities might be expected to obtain competitive jobs more easily (relative to children with predominantly mental disabilities) as adults in a twenty-first century employment market that increasingly values information processing skills over physical talents or sensory abilities. The distinction between mental and physical disabilities is not altogether straightforward in the case of Down syndrome, however, which typically results in both varying levels of mental retardation and a range of physical impairments. At any rate, the cognitive limitations that people tend to associate with the condition buttress our expressivist interpretation of the ADA. Considering Wertz’s eight factors, prospective parents may believe that future employers will consider adults with Down syndrome too unproductive (Wertz’s second factor) or focus on the fact that medical treatment cannot reverse the physical or cognitive limitations associated with Down syndrome (her fourth factor). However, it seems plausible that the overall salience of a disability matters more for any effect that expressive externalities have on the reproductive utility function than whether the disability is primarily mental or physical. Our empirical analysis does not test this hypothesis directly.

Given the strong correlation between maternal age and the likelihood of a fetus developing Down syndrome, we divide all birth records into two groups with age thirty-five as a cutoff. We use thirty-five as a threshold because the medical community traditionally has recommended amniocentesis for women who are at least that age and also because “[m]ore than 90 percent of all children and 70 to 80 percent of children with Down’s syndrome are born to women younger than 35 years of age.”

Table 5 suggests that most of the negative expressive effects estimated for Down syndrome actually arose in the younger maternal cohort. In this table we present only the unadjusted and fully specified models for both groups. As Column 2 shows, declines in the Down syndrome birthrate were much steeper and lasted longer for women younger than thirty-five than among the entire set of births. On average, between 1993 and 1999 the birthrate drop was about 16 per 100,000. Column 4, describing the same changes in Down syndrome births for women age thirty-five and older, shows even higher point estimates over the same period. However, the noisiness of the data (due to the lower frequency of births among women older than thirty-five) renders all of the coefficients insignificant. Although this noise prevents us from drawing firm conclusions, Table 5 would seem to support our interpretation that the ADA produced a negative expressive effect. Since women younger than thirty-five give birth to the vast majority of Down

---


syndrome children, and about 90 percent of pregnant women carrying a fetus diagnosed with the condition choose termination, the substantial declines observed in Table 5 likely were due to selective abortion.

To be clear, we do not claim to have detected a causal statistical relationship between the ADA and changes in the Down syndrome birthrate. We have used the best available natality information simply to show that shifts in otherwise stable birthrate patterns could plausibly have followed from passage of the ADA. Our regression analysis shows significant declines in the Down syndrome birthrate while controlling for a multitude of critical variables correlated with reproductive decisions. The absence of data on actual Down-selective terminations and of comprehensive interview responses by prospective parents who face this decision caution against drawing broad conclusions. Nevertheless, these findings are sufficiently robust to suggest that the “disappointing” hypothesis we proposed in Section III(B) has greater independent plausibility or net influence than the “uplifting” hypothesis from Section III(A). Further investigation into other family planning spheres such as adoption or artificial reproductive technologies might weaken or reinforce this conclusion.

V. THE ANALYTICS OF DISABILITY-SELECTIVE ABORTION

A complex array of legal, material, economic, technological, social, familial, and medical factors might reasonably inform prenatal testing and selective abortion for Down syndrome. Technological factors include the predictive accuracy, procedural invasiveness, and medical risks of diagnostic techniques, as well as the gestational age for which effective prenatal testing is available. Legal factors

184 Harmon, supra note 2.
185 Another possibility is that amniocentesis rates by age changed over the 1990s such that women under thirty-five switched to less invasive procedures such as the triple screen/AFP test. If so, this technological shift might explain the relative decrease in Down syndrome births among younger women. However, our natality data show that amniocentesis rates remained steady among both the younger cohort (about 2 percent) and the older cohort (about 15 percent) during the early to mid-1990s.
186 See infra Part VI.
187 See D.I. Bromage, Prenatal Diagnosis and Selective Abortion: A Result of the Cultural Turn?, 32 MED. HUMAN. 38, 39–41 (2006) (describing a range of influences that may be brought on women who are carriers of genetic diseases).
might include informed-consent laws\textsuperscript{190} and other government regulations;\textsuperscript{191} public funding;\textsuperscript{192} or tort litigation\textsuperscript{193} with respect to genetic testing and abortion. Relevant material factors include the availability of institutions that provide genetic testing;\textsuperscript{194} the required waiting period for genetic testing services,\textsuperscript{195} and challenges with transportation or child care to allow for medical services.\textsuperscript{196} Economic factors might include social status;\textsuperscript{197} availability of insurance coverage, government funding, or personal funds to pay for genetic testing\textsuperscript{198} or abortion services;\textsuperscript{199} financial ability to support a child with a genetic disability or disease;\textsuperscript{200} and the anticipation of denied coverage or increased costs for health and life insurance for avoidable or preexisting conditions.\textsuperscript{201} Salient social issues include pressure by nonfamilial individuals;\textsuperscript{202} ethnic, cultural, and religious

\begin{itemize}
\item \textsuperscript{190} Jean Gekas et al., \textit{Informed Consent to Serum Screening for Down Syndrome: Are Women Given Adequate Information?}, 19 PRENATAL DIAGNOSIS 1, 3–6 (1999).
\item \textsuperscript{192} \textit{Id.} at 137.
\item \textsuperscript{193} \textit{Id.} at 151–52.
\item \textsuperscript{194} See Dorothy C. Wertz & John C. Fletcher, \textit{Feminist Criticism of Prenatal Diagnosis: A Response}, 36 \textit{CLINICAL OBSTETRICS & GYNECOLOGY} 541, 543–44 (1993).
\item \textsuperscript{195} Marilyn L. Poland et al., \textit{Barriers to Receiving Adequate Prenatal Care}, 157 \textit{AM. J. OBSTETRICS & GYNECOLOGY} 297, 302 (1987).
\item \textsuperscript{196} \textit{Id.} at 300.
\item \textsuperscript{197} Babak Khoshnood et al., \textit{Advances in Medical Technology and Creation of Disparities: The Case of Down Syndrome}, 96 \textit{AM. J. PUB. HEALTH} 2139, 2143 (2006) (concluding that Down syndrome risk does not vary according to socioeconomic status).
\item \textsuperscript{199} See R. Alta Charo & Karen H. Rothenberg, \textit{“The Good Mother”: The Limits of Reproductive Accountability and Genetic Choice}, in \textit{WOMEN AND PRENATAL TESTING}, at 195. The concern about unequal access to abortions was echoed by Justice Ginsburg in a recent interview. Emily Bazelon, \textit{The Place of Women on the Court}, N.Y. \textit{TIMES MAG.}, July 12, 2009, at 22, 47 (quoting Justice Ginsburg) (“The states that had changed their abortion laws before Roe [to make abortion legal] are not going to change back. So we have a policy that affects only poor women . . . .”).
\item \textsuperscript{200} See generally Babak Khoshnood et al., \textit{Socioeconomic and State-Level Differences in Prenatal Diagnosis and Live Birth Prevalence of Down’s Syndrome in the United States}, 51 REV. EPIDEMIOLOGY 617 (2003) (examining the impact of socioeconomic differences in prenatal testing on disparities in the live birth prevalence of congenital anomalies).
\item \textsuperscript{201} See Philip Jacobs & Suzanne McDermott, \textit{Family Caregiver Costs of Chronically Ill and Handicapped Children: Method and Literature Review}, 104 \textit{PUB. HEALTH REPS.} 158, 159–60 (1989).
\end{itemize}
differences in parental attitudes toward genetic testing or abortion; and knowledge about the possibility, benefits, and risks of genetic testing. Expectations imposed by family members, the prospective impact of a genetically affected child on the family unit, costs and burdens to primary caregivers of children with genetic disabilities may also play a role. Finally, medical factors might include the delivery of genetic information by physicians and counselors, maternal and gestational age, and personal anxiety about genetic testing or abortion.

At first glance, any of these potential influences seems capable of explaining changes in prenatal testing or selective abortion rates. On closer inspection, however, it appears less likely that these factors, either independently or in combination, fully account for the Down syndrome birthrate decline in the years after the ADA’s passage. We consider the potential influence of each factor in turn, beginning with prenatal diagnostic technology.

Although the technology of prenatal testing developed dramatically both in the decades prior to the enactment of the ADA and since the year 2000, the 1990s saw few technological advances in fetal testing. The oldest and most common

---

203 See Rayna Rapp, Women’s Responses to Prenatal Diagnosis: A Sociocultural Perspective on Diversity, in WOMEN AND PRENATAL TESTING, supra note 191, at 219.


205 See James R. Sorensen & Dorothy C. Wertz, Couple Agreement Before and After Genetic Counseling, 25 AM. J. MED. GENETICS 549, 553–54 (1986) (suggesting that most couples confront disagreement with respect to reproductive plans).


208 See Amy S. Kaiser et al., The Effects of Prenatal Group Genetic Counselling on Knowledge, Anxiety and Decisional Conflict: Issues for Nuchal Translucency Screening, 22 J. OBSTETRICS & GYNECOLOGY 246, 252 (2002) (concluding that anxiety metrics among those studied did not change significantly before counseling compared with after counseling).

209 See Ralph L. Kramer et al., Determinants of Parental Decisions After the Prenatal Diagnosis of Down Syndrome, 79 AM. J. MED. GENETICS 172, 174 (1998) (determining that surveyed patients who chose abortion following a positive fetal test for Down syndrome tended to be older and earlier in their pregnancy than those electing to continue their pregnancy).


211 Before prenatal testing, information about congenital disability in reproductive decision making was limited to carrier screening, which cannot diagnose particular defects in prenatal lives, but only identify recessive genes that would place individuals at an
fetal testing technique is amniocentesis.\textsuperscript{212} First developed in the 1950s, the procedure became routine by the 1970s, and since then it has undergone no significant technical refinements.\textsuperscript{213} Just as in the 1970s, amniocentesis is currently performed in the second trimester of pregnancy, at sixteen to eighteen weeks into the gestation period.\textsuperscript{214} The procedure involves the insertion of a hollow needle through the abdomen of a pregnant woman to remove a sample of the amniotic fluid surrounding the developing fetus.\textsuperscript{215} The fluid contains fetal cells that can be examined to determine fetal sex, the number of cellular chromosomes, and specific genetic disorders.\textsuperscript{216} The risk of fetal loss, approximately 0.5 percent, has remained roughly constant since the 1970s.\textsuperscript{217}

Another diagnostic procedure, chorionic villus sampling (CVS), was introduced in the early 1980s.\textsuperscript{218} In CVS, a small tube is inserted through the vagina and cervix and into the placenta, from which a small amount of tissue is removed.\textsuperscript{219} CVS can be performed several weeks earlier than amniocentesis, but it poses a higher risk of fetal loss, namely about 1 percent.\textsuperscript{220} CVS has advanced little in technical precision, invasiveness, or risk, however, since its pre-ADA inception,\textsuperscript{221} decreasing the likelihood that the advent of this procedure had a significant impact on amniocentesis or selective abortion rates in the mid-1990s.\textsuperscript{222} Similarly, a prenatal technique called the triple screen,\textsuperscript{223} a maternal blood increased risk of having a child born with a disability. RUTH SCHWARTZ COWAN, HEREDITY AND HOPE: THE CASE FOR GENETIC SCREENING 10 (2008).

\textsuperscript{212} See R. Douglas Wilson, Amniocentesis and Chorionic Villus Sampling, 12 CURRENT OPINION IN OBSTETRICS \\& GYNECOLOGY 81, 81 (2000) (“Invasive prenatal diagnosis continues to be the gold standard for pregnancies.”).

\textsuperscript{213} See COMMITTEE ON ASSESSING GENETIC RISKS, ASSESSING GENETIC RISKS: IMPLICATIONS FOR HEALTH AND SOCIAL POLICY 76 (Lori B. Andrews et al. eds., 1994).


\textsuperscript{215} See Andrews et al., supra note 213, at 83.

\textsuperscript{216} See id.

\textsuperscript{217} See id.


\textsuperscript{219} See id.

\textsuperscript{220} See ELENA O. NIGHTINGALE \\& MELISSA GOODMAN, BEFORE BIRTH: PREGNATAL TESTING FOR GENETIC DISEASE 35–36 (1990) (noting that the risk of miscarriage from this procedure is about the same as amniocentesis).

\textsuperscript{221} See THE MERCK MANUAL OF DIAGNOSIS AND THERAPY 2146 (Mark H. Beers et al. eds., 18th ed. 2006).

\textsuperscript{222} Boaz Weisz \\& Charles H. Rodeck, An Update on Antenatal Screening for Down’s Syndrome and Specific Implications for Assisted Reproduction Pregnanacies, 12 HUM. REPRODUCTION UPDATE 513, 518 (2006). Moreover, testing for the effects of CVS on disability-selective abortion—as done with amniocentesis—is not possible due to the lack of information on the test in the U.S. natality data.

\textsuperscript{223} This test is also known as the Triple Test, Multiple Marker Screening, or AFP Plus. See Triple Screen Test, American Pregnancy Association, http://www.american
screening test for fetal levels of alpha-fetoprotein, estriol hormone, and human chorionic gonadotropin hormone, had gained widespread use by the late-1990s, but it was not generally available in U.S. hospitals before 1998. It therefore seems unlikely that technological advances in diagnostic procedures can explain significant changes in reproductive decision making during the years between 1992 and 1998.

Legal changes may seem like a more plausible candidate, but reproduction law likewise saw few significant changes during the mid-1990s. Informed-consent laws and public funding of prenatal testing could have a range of plausible medical, material, and economic consequences tending to influence reproductive decision making. These factors may affect the availability of genetic testing providers, the required waiting period for genetic testing services, the delivery of genetic information to prospective parents by physicians and counselors, and the ability to pay for prenatal diagnosis. While amniocentesis is expensive, at a cost of a little more than $1,000 in the early 1990s, most insurance plans covered prenatal tests long before the enactment of the ADA, especially for women older than thirty-five.

In the years between 1990 and 1992, up to and until the Supreme Court reaffirmed the abortion right in Planned Parenthood of Southern Pennsylvania v. Casey, a small number of states passed laws making abortion more accessible or less expensive when the parent wished to avoid having a child with a disability. In the early 1990s, Maryland, Kansas, Texas, and Utah enacted statutes...
permitting late-term abortion only on narrow grounds, including diagnosis of fetal
disability. During those same years, Washington, Minnesota, and California passed laws subsidizing prenatal tests. Tennessee, Iowa, and Maryland fund abortions to prevent disability but not for other reasons. Among reproductive laws passed before 1992 that affect parents’ decisions to undergo prenatal testing or to continue a pregnancy following a positive test for fetal disability, few statutes have since been repealed or substantially amended, and few similar laws were enacted before 2000.

A plausible legal influence on selective abortion is the Family Medical Leave Act of 1993 (FMLA), which requires employers for most of the nation’s workers to provide up to twelve weeks of unpaid leave of absence to care for the birth of a child, among other familial needs. Some might speculate that the FMLA’s protections for women in the workforce could incentivize having children at an earlier age and that resulting changes in maternal age patterns could in turn affect

233 Texas permitted abortion after viability if the fetus “has a severe and irreversible abnormality, as identified through reliable diagnostic procedures.” TEX. REV. CIV. STAT. ANN. art. 44956 (1991).
235 Washington required insurers who provide maternity benefits to include prenatal tests within the benefits package. Washington also required maternity benefits to include prenatal tests. See WASH. REV. CODE ANN. § 48.44.344 (1990) (group health care services contract to cover prenatal diagnosis); Id. § 48.46.375 (1990) (HMOs to cover prenatal diagnosis).
236 Minnesota also required maternity benefits to include prenatal tests. See 1991 MINN. CH. LAWS 33, § 36 (some insurance benefits to cover prenatal diagnosis) (codified at MINN. STAT. ANN. ch. 62K).
237 California also required maternity benefits to include prenatal tests. See CAL. INS. CODE § 10123.9 (1991) (group policy coverage of prenatal diagnosis); id. § 11512.18 (1991) (nonprofit hospital service plan coverage of prenatal diagnosis).
238 Tennessee gave public funding for abortion in cases where the fetus is “medically determined to have severe physical deformities or abnormalities or severe mental retardation” 1992 TENN. PUB. ACTS 1018, § 10, Item 4(3).
239 Iowa funded abortion when the “fetus is physically deformed, mentally deficient, or afflicted with a congenital illness.” 1991 IOWA ACTS 267, §§ 103, b & 210, a(2).
240 Maryland subsidized abortion to avoid “congenital defect or serious deformity or abnormality.” 1992 MD. LAWS 64, § 1, 32.17.01.02.
rates of fetal testing and termination. The FMLA offers no special provisions for children with disabilities, however, and natality data indicate that average maternal age is, at any rate, a statistically insignificant predictor of Down syndrome birthrates, making the FMLA an unlikely influence on disability-selective abortion rates in the 1990s.

Some states recognize “wrongful life” or “wrongful birth” torts, which might be thought to influence parental decisions to continue a pregnancy following prenatal testing for Down syndrome. In three states—California, New Jersey, and Washington—a child born with a serious genetic disability may sue a physician if its parents show that they would have aborted had they been informed of the potential disability. In these lawsuits, genetic professionals can be held liable for birth defects if physicians or counselors did not meet the standard

---


244 See infra tbl.1.

245 Malpractice claims for wrongful life and wrongful birth are more common than claims for failure to inform of the risk of miscarriage from amniocentesis. See Wendy Fritzen Hensel, 40 Harv. C.R.-C.L. L. Rev. 141, 151 (2005). This is likely due to the care with which testing providers characteristically discuss and document the details and risks of amniocentesis with their patients. See, e.g., Bedel v. Univ. of Cinn. Hosp., 669 N.E.2d 9, 13–16 (Ohio Ct. App. 1995) (affirming the dismissal of an informed consent action alleging a physician’s failure to disclose the risks of miscarriage attending amniocentesis because the patient was verbally informed of the risk and signed three consent forms). Causes of action for misdiagnosing prenatal genetic disorder are also uncommon. See, e.g., Martinez v. Long Island Jewish Hillside Med. Ctr., 512 N.E.2d 538, 538–39 (N.Y. 1987) (reversing the lower court’s dismissal of a claim for emotional distress against a physician for incorrectly diagnosing a fetal brain abnormality and reasoning that the plaintiff’s decision to terminate the pregnancy caused her psychological injury flowing from her belief that abortion is a sin except in unusual circumstances).

246 See Curlender v. Bio-Science Labs., 165 Cal. Rptr. 477, 487 (Cal. Ct. App. 1980) (reasoning that disability-selective abortion is a desirable means of reducing the “increasingly large part of the overall health care burden” represented by children with congenital disabilities (quoting Note, Father and Mother Know Best: Defining the Liability of Physicians for Inadequate Genetic Counseling, 87 Yale L.J. 1488, 1496, 1499 (1978)). The Curlender court acknowledged in dicta that its reasoning would support a cause of action by a seriously disabled child against his parents for choosing to bring him into the world. Id. at 487–89. If parents make a conscious and informed choice to carry a seriously disabled child to term, the majority reasoned, nothing should “protect [them] from being answerable for the pain, suffering and misery which they have wrought upon their offspring.” Id. at 488.


249 The “adequate disclosure” requirement in medical informed consent law was established in Canterbury v. Spence, 464 F.2d 772, 782 (D.C. Cir. 1972) (involving a physician who failed to warn his patient of a risk of paralysis that could result from back surgery).
of care in offering prenatal testing to a high-risk patient, if they failed to diagnose a detectable defect through prenatal testing, or if they failed to inform the patient about a detected defect. Success in such suits requires parents to prove that they would have terminated the pregnancy had they learned of the genetic disorder in time to abort. In 2007, a Florida couple was awarded $21 million after a physician failed to prenatally diagnose their son with Smith-Lemli-Opitz syndrome, an obscure congenital disorder that inhibits the production of cholesterol. Recognition of such suits could affect reproductive decision making to the extent that expected liability encourages physicians to endorse prenatal testing more often. However, all such state tort law has been on the books since before 1985, so it is unlikely this would have had a large effect on reproductive decision making in the 1990s.

No federal agency weighed in on the practice of genetic testing between 1982 and 2007. The American College of Obstetricians and Gynecologists (ACOG), which receives federal funds from the National Institutes of Health, released a technical bulletin in 1982 stating that “routine maternal serum prenatal screening of all [pregnant women] is of uncertain value.” Maternal serum screening is indeed a prenatal diagnostic technique with limited predictive capability. While the Food and Drug Administration approved the use of maternal serum screening for neural tube defects in 1983, it has never approved the use of that particular procedure to screen prenatally for Down syndrome. The ACOG prominently

251 See id.
255 See American College of Obstetricians and Gynecologists (ACOG), Technical Bulletin No. 67, Prenatal Detection of Neural Tube Defects (1982). ACOG claimed that in the absence of high-quality laboratory, counseling, ultrasound, and amniocentesis services, the procedure “could simply increase cost and parental anxiety . . . and possibly lead to unnecessary abortions.” See id.
256 James E. Haddow et al., Prenatal Screening for Down’s Syndrome with Use of Maternal Serum Markers, 327 NEW ENG. J. MED. 588, 588–93 (1992) [hereinafter Haddow et al., Prenatal Screening]. The procedure consists of measuring four chemical markers present in a pregnant woman’s blood, abnormal levels of which may indicate increased risk for certain birth defects and genetic diseases. See James E. Haddow et al., Screening of Maternal Serum for Fetal Down’s Syndrome in the First Trimester, 338 NEW ENG. J. MED. 955, 955 (1998) [hereinafter Haddow et al., Screening of Maternal Serum].
258 See Haddow et al., Prenatal Screening, supra note 256, at 588.
revisited the issue of prenatal testing for fetal disability in December 2007, when the organization recommended that hospitals and physicians be required to expand their offer of prenatal testing for Down syndrome and other common genetic disorders to pregnant women of all ages, not just to those thirty-five and older. While these most recent ACOG recommendations are very likely to influence amniocentesis and selective abortion rates in the future, their release in 2007 makes it implausible that they would have affected parental decision making during the period from 1992 to 2002.

The legal doctrine of informed consent could have influenced testing and abortion rates, insofar as it affects the delivery of genetic information to prospective parents by physicians and counselors. A seminal medical ethics textbook notes that the “primary justification advanced for requirements of informed consent has been to protect autonomous choice.” Guided by the principle of autonomy, universities began offering degree programs in genetic counseling. These courses were designed to help aspiring professionals educate patients about their genetic risks and testing options and to assist prospective

259 See Press Release, American College of Obstetricians and Gynecologists, New Recommendations for Down Syndrome: Screening Should Be Offered to All Pregnant Women (Jan. 2, 2007), available at http://www.acog.org/from_home/publications/press_releases/nr01-02-07-1.cfm; see also ACOG Practice Bulletin No. 88: Invasive Prenatal Testing for Aneuploidy, 110 OBSTETRICS & GYNECOLOGY 1459, 1465 (2007). The Senate Health, Education, Labor, and Pensions Committee responded by proposing legislation that would require prospective parents who receive a diagnosis of Down syndrome or any other genetic condition be provided with information about the condition as well as support services and networks that offer assistance in raising a child with the condition. See Prenatally and Postnatally Diagnosed Conditions Awareness Act of 2008, S. 1810, 110th Cong. § 2(1) (“It is the purpose of this Act to . . . increase patient referrals to providers of key support services for women who have received a positive diagnosis for Down syndrome, or other prenatally or postnatally diagnosed conditions, as well as to provide up-to-date information on the range of outcomes for individuals living with the diagnosed condition, including physical, developmental, educational, and psychosocial outcomes . . . .”). The final bill, signed into law by President George W. Bush on October 8, 2008, provides for the establishment of “resource telephone hotlines” and “a national registry, or network of local registries, of families willing to adopt newborns with Down syndrome . . . .” Id. § 3(b)(1)(B)(i), (iv). The law importantly requires health care providers receiving grants and other funds from the federal government to offer “[u]p-to-date, evidence-based, written information concerning the range of outcomes for individuals living with the diagnosed condition” and “[c]ontact information regarding support services” such as the hotlines established in § 3(b)(1)(B)(i) and educational programs at the national and local levels. Id. § 3(c)(1)(A)–(B).

260 See Susan J. Hayflick & M. Patrice Eiff, Role of Primary Care Providers in the Delivery of Genetics Services, 1 COMMUNITY GENETICS 18, 20 (1998) (explaining that genetic counselors begin by obtaining information including family history, medical history, and pregnancy history).


262 Id.
parents in making reproductive decisions consistent with their values and goals.\(^{263}\)

Since its inception, the field of genetic counseling has sought to distinguish its ideals from those of twentieth-century eugenics.\(^{264}\) It is not surprising, therefore, that the profession quickly adopted a nondirective approach that rejects specific recommendations in favor of a balanced and impartial presentation of all relevant information.\(^{265}\) Yet neither the professional ideals of genetic counseling nor the legal principle of informed consent has undergone significant changes over the past quarter century.\(^{266}\) The basic goal of nondirectiveness has not changed since the 1980s.\(^{267}\)

The foregoing discussion suggests that legal, economic, medical, and technological factors cannot plausibly account for the large changes in Down syndrome birthrates observed during the 1990s.\(^{268}\) We cannot, of course, rule out these alternative explanations in the absence of considerably more rigorous research and analysis. Moreover, it may be that the social attitudes underlying changes in state reproductive laws themselves contributed to an increase in termination rates. One plausible explanation, however, is the “disappointing ADA” account we presented in Part III. We hypothesized that the ADA, through its influence on social interaction and media coverage, reinforced negative affective attitudes toward people with disabilities in general and toward those with Down syndrome in particular.\(^{269}\)

We have not attempted to prove that any specific mechanism causally links the ADA to declining birthrates among children with Down syndrome. We have argued that two accounts of expressive externalities merit further investigation. The first is the ADA’s influence on increased interpersonal contact with people with disabilities. The second is the media’s framing of stories about people with disabilities. Our quantitative and qualitative results suggest that the ADA, as filtered through complex social networks and the media, may have contributed to the decline in Down syndrome birthrates from 1993 to 1998.\(^{270}\)


\(^{267}\) See Beth A. Fine, The Evolution of Nondirectiveness in Genetic Counseling and Implications of the Human Genome Project, in PRESCRIBING OUR FUTURE: ETHICAL CHALLENGES IN GENETIC COUNSELING 101, 105–06 (Dianne M. Bartels et al. eds., 1993).

\(^{268}\) See supra Part V.B.I.

\(^{269}\) See supra Part IV.B.

\(^{270}\) See supra Part V.B.I.
We suggest that policymakers supplement the passage of certain civil rights laws with public education campaigns on behalf of the protected group in question. Where reproductive practices are affected, physicians should refer parents who receive an unexpected diagnosis to resources such as local support groups, national disability organizations, and parents who have children with disabilities. Future research might apply our expressive approach to explore the impact civil rights laws protecting women or homosexuals may have had on the incidence of prenatal selection based on sex or sexual orientation. Other studies might consider the expressive effects of antidiscrimination or immigration law on the rate at which biological parents place particular groups of children up for adoption, or the rate at which parents choose to adopt these children over others once they enter state custody.

VI. CONCLUSION

This Article has considered whether, why, and how the Americans with Disabilities Act influenced the birthrate of infants with Down syndrome. Part II drew on expressive law theory to explain the connection between the ADA, social attitudes toward people with disabilities, and the practice of disability-selective abortion. We introduced the concept of expressive externalities to capture the social cost of legal actions unrelated to behavior the law was designed to regulate. Part III presented competing hypotheses of the ADA’s expressive effects for disability-selective abortion. The first hypothesis—what we called “the uplifting ADA”—suggests that the ADA encouraged parents to bring fetuses with Down syndrome to term by promoting new rights and opportunities for people with disabilities. The second hypothesis—“the disappointing ADA”—conjectured that the ADA paradoxically might have promoted disability-selective abortion if social interactions reinforced negative attitudes toward people with disabilities or if the media portrayed people with disabilities as incurring undesirable costs for society.

The econometric analysis in Part IV showed that Down syndrome birthrates decreased significantly and steadily from 1993 to 1998. The decline of 13 to 18 Down syndrome births per 100,000 was robust, even with the inclusion of demographic and health-related control variables and against the backdrop of highly stable prenatal screening rates. We did not find supportive evidence for a similar effect among infants with spina bifida and cleft palate. Part IV provided support for our argument about the expressivist effects of the ADA on Down-selective abortion, by accounting for variables related to technology, law, and medicine.

These conclusions yield important implications for reproduction and antidiscrimination law. First, we should be clear that we do not believe our findings justify restrictions on a woman’s constitutional right to terminate a pregnancy for any reason, at least until the third trimester.271 One reason not to

restrict abortions that are sought for particular purposes is that such prohibition carries costs that may outweigh any benefits of prevention. For example, attempts to criminalize disability-selective abortion could authorize easily abused inquiries into the legitimate reasons that women have for seeking abortions. This is important because a woman’s right to avoid the obligations and expectations that attend childbearing and childrearing is properly protected as a matter of moral and legal equality. The short-term effects we observe on disability-selective abortion nevertheless may give reason to advance public education campaigns to correct misleading social perceptions about the eight components of the disability-based reproductive utility function.

Our expressive approach for analyzing the effects of disability law on social attitudes and reproductive behavior has fruitful and straightforward applications for other inquiries. Promising areas for additional research include adoption on the basis of race or disability, and prenatal selection for sex or for sexual orientation (to the extent that sexual orientation has a genetic basis). Future studies might investigate whether expressive externalities increase the rate at which biological parents place children with disabilities up for adoption, or whether expressive externalities decrease the rate at which adoptive parents choose to raise such children. Related projects might explore whether the second wave feminist movement had any impact on sex selection, or how the evolving legal and social

---

274 See Heidi Zhou, Doctors Urged to Rethink Options after Prenatal Down Syndrome Diagnosis, News 8 Austin, Oct. 12, 2007, available at http://www.news8austin.com/content/headlines/?ArID=193564&SecID=2 (discussing a campaign to educate health workers that was launched by the Down Syndrome Association of Central Texas).
275 See supra Part III.C.
276 A team of scientists at the National Institutes of Health reported discovering a “statistically significant correlation between the inheritance of genetic markers . . . and sexual orientation in a selected group of homosexual males.” Dean H. Hamer et al., A Linkage Between DNA Markers on the X Chromosome and Male Sexual Orientation, 261 Science 321, 321 (1993); see also Ivanka Savic & Per Lindström, PET and MRI Show Differences in Cerebral Asymmetry and Functional Connectivity Between Homosexual and Heterosexual Subjects, 105 Proc. Nat’l Acad. Sci. 9403, 9407 (finding that homosexual and heterosexual orientation appear to be linked to genetic brain structures rather than factors after birth).
status of homosexuals could influence selective abortion on the basis of sexual orientation, if such techniques became possible.277

Selective adoption patterns offer a fresh look at the impact of disability law on decisions to raise children with disabilities. A measurable increase in the rate at which children with disabilities are put up for adoption would presumably indicate that parents are less willing to rear them. Similarly, if adoptive parents tend to select healthy children over similarly situated children with disabilities, we would observe an ex post version of the fertility decisions examined between 1989 and 2002. Variation in preferences for male versus female children across countries and cultures is well-documented.278 In the last decade, fertility clinics have embraced a new technology that permits parents to choose a boy or girl prior to fertilization.279 This process, known as MicroSort, enjoys a 74 percent and 88 percent success rate in producing boys and girls, respectively.280 Expressive externalities could stimulate demand for one sex over the other if changes in law bring about changes in the relative value parents ascribe to male and female children. Even greater expressive externalities could attend the widespread passage of gay marriage laws in the wake of the U.S. Supreme Court decision in Lawrence v. Texas.281

This Article has explored the surprising ways that antidiscrimination law can change social behavior in spheres completely unrelated to those that the law regulates. We puzzled through this phenomenon by considering the relation between the Americans with Disabilities Act and selective abortion following a positive test for Down syndrome. We examined the paradoxical possibility that the ADA could have served to prevent the existence of people with disabilities, the very class of persons the law was meant to protect. Our analysis does not indict the ADA or antidiscrimination law generally. Instead, we have called attention to the way that antidiscrimination law can, at least in the short-term, transform social interaction and media portrayals in ways that reinforce “society’s accumulated myths and fears about disability and disease.”282 As the Nobel Prize-winning economist who popularized the concept of externalities observed almost fifty years ago:


278 See Dorothy C. Wertz & John C. Fletcher, Fatal Knowledge? Prenatal Diagnosis and Sex Selection, 19 HASTINGS CTR. REP. 21, 24 (1989); John A. Robertson, Preconception Gender Selection, 1 AM. J. BIOETHICS 1, 2 (2001).

279 See Keith L. Blauer, Human Sperm Sorting Method is Showing Success at Separating the Girls from the Boys, 16 GENETICS & IVF INST. NEWSL. 2 (2002).


[I]n choosing between social arrangements within the context of which individual decisions are made, we have to bear in mind that a change in the existing system which will lead to an improvement in some decisions may well lead to a worsening of others.283

Heeding Coase’s insight does not mean that antidiscrimination law—or any other legal provision—will necessarily generate controversial trade-offs. Our analysis suggests, however, that we would do well to contemplate the attitudinal and behavioral changes implicated by major legal changes. Our growing power over genetic testing and reproductive biotechnology gives special reason to attend to such collateral effects. Accounting for expressive externalities can enhance our understanding of the complex ways that people make decisions and help secure the promise of civil rights and self-respect for those in need of the law’s protection.

283 Coase, supra note 75, at 44.
### Appendix

**Table 1: Summary Statistics (1989–2002)**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Down syndrome births (per 100,000)</td>
<td>52.20</td>
<td>24.57</td>
</tr>
<tr>
<td>Spina bifida births (per 100,000)</td>
<td>25.48</td>
<td>15.79</td>
</tr>
<tr>
<td>White mothers (%)</td>
<td>79.67</td>
<td>15.65</td>
</tr>
<tr>
<td>Married mothers (%)</td>
<td>69.08</td>
<td>7.66</td>
</tr>
<tr>
<td>College-educated mothers (%)</td>
<td>21.64</td>
<td>6.34</td>
</tr>
<tr>
<td>Mother’s age at birth</td>
<td>26.71</td>
<td>1.03</td>
</tr>
<tr>
<td>Birth order (including newborn child)</td>
<td>2.02</td>
<td>0.11</td>
</tr>
<tr>
<td>Births subject to amniocentesis (%)</td>
<td>2.91</td>
<td>1.56</td>
</tr>
<tr>
<td>Number of prenatal visits</td>
<td>11.35</td>
<td>0.69</td>
</tr>
</tbody>
</table>

*Source:* National Center for Health Statistics Annual Natality Detail Files.  
*Notes:* The summary statistics are derived from the set of state-year observations such that \( N = 714 \).
Table 2: Estimated Annual Changes in Down Syndrome Birthrates (1989–2002)

<table>
<thead>
<tr>
<th>Year</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>-3.69</td>
<td>-4.93</td>
<td>-3.67</td>
<td>-5.17</td>
</tr>
<tr>
<td>1992</td>
<td>-3.80</td>
<td>-5.96</td>
<td>-4.00</td>
<td>-6.57</td>
</tr>
<tr>
<td>1995</td>
<td>-11.00***</td>
<td>-14.36**</td>
<td>-12.46***</td>
<td>-18.21***</td>
</tr>
<tr>
<td>1996</td>
<td>-6.24</td>
<td>-10.25</td>
<td>-7.45</td>
<td>-13.76*</td>
</tr>
<tr>
<td>1997</td>
<td>-9.92*</td>
<td>-14.80**</td>
<td>-10.47**</td>
<td>-17.67**</td>
</tr>
<tr>
<td>1998</td>
<td>-10.32***</td>
<td>-15.62**</td>
<td>-9.94**</td>
<td>-17.51**</td>
</tr>
<tr>
<td>2000</td>
<td>-1.41</td>
<td>-8.23</td>
<td>0.84</td>
<td>-8.61</td>
</tr>
<tr>
<td>2001</td>
<td>-5.24</td>
<td>-12.50</td>
<td>-2.15</td>
<td>-12.01</td>
</tr>
<tr>
<td>2002</td>
<td>-4.10</td>
<td>-11.97</td>
<td>-0.11</td>
<td>-10.88</td>
</tr>
</tbody>
</table>

Demographic Controls? | No | Yes | No | Yes
Health Controls? | No | No | Yes | Yes
N | 714 | 714 | 714 | 714

Source: National Center for Health Statistics Annual Natality Detail Files.
Notes: The dependent variable in each OLS regression is the birthrate for infants with Down syndrome in each state and year. All regressions except Column 1 include state fixed effects. Robust standard errors are clustered at the state level. Demographic controls include: (a) average age at childbirth of mother in state s, year t; (b) share of women who are white giving birth in state s, year t; (c) share of women with some college education or less, at childbirth in state s, year t; and (d) share of married women at childbirth in state s, year t. Medical and health controls include: (a) average live birth order (including the current birth) in state s, year t;
(b) average number of prenatal visits in state \( s \), year \( t \); and (c) share of live births subject to amniocentesis in state \( s \), year \( t \). ** = significance at the 5% level and *** = significance at the 1% level.

**Table 2 (Cont.): Estimated Annual Changes in Down Syndrome Birthrates (1989–2002)**

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Maternal Age</td>
<td>-</td>
<td>6.18</td>
<td></td>
<td>9.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(6.29)</td>
<td></td>
<td>(6.70)</td>
</tr>
<tr>
<td>Percentage Married</td>
<td>-</td>
<td>-8.79</td>
<td>-24.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(49.56)</td>
<td>(52.19)</td>
<td></td>
</tr>
<tr>
<td>Percentage White</td>
<td>-</td>
<td>-141.26**</td>
<td>-156.79***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(58.01)</td>
<td>(55.42)</td>
<td></td>
</tr>
<tr>
<td>Share with College Education</td>
<td>-</td>
<td>11.05</td>
<td>-3.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(83.60)</td>
<td>(73.39)</td>
<td></td>
</tr>
<tr>
<td>Average Birth Order</td>
<td>-</td>
<td>-41.76</td>
<td>-71.03***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(23.58)</td>
<td>(25.26)</td>
<td></td>
</tr>
<tr>
<td>Average Number of Prenatal Visits</td>
<td>-</td>
<td>2.02</td>
<td>2.77</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3.12)</td>
<td>(3.17)</td>
<td></td>
</tr>
<tr>
<td>Share using Amniocentesis</td>
<td>-</td>
<td>312.95</td>
<td>252.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(170.21)</td>
<td>(180.87)</td>
<td></td>
</tr>
</tbody>
</table>

*Source: National Center for Health Statistics Annual Natality Detail Files.*

*Notes: The dependent variable in each OLS regression is the birthrate for infants with Down syndrome in each state and year. All regressions except Column 1 include state fixed effects. Robust standard errors are clustered at the state level. ** = significance at the 5% level and *** = significance at the 1% level.*

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>-1.97</td>
<td>1.03</td>
<td>-1.78</td>
<td>-1.27</td>
</tr>
<tr>
<td></td>
<td>(1.64)</td>
<td>(1.94)</td>
<td>(1.63)</td>
<td>(1.93)</td>
</tr>
<tr>
<td>1992</td>
<td>-2.38</td>
<td>-0.15</td>
<td>-2.01</td>
<td>-0.48</td>
</tr>
<tr>
<td></td>
<td>(2.77)</td>
<td>(2.83)</td>
<td>(2.87)</td>
<td>(2.77)</td>
</tr>
<tr>
<td>1993</td>
<td>2.04</td>
<td>5.34</td>
<td>2.48</td>
<td>5.11</td>
</tr>
<tr>
<td></td>
<td>(2.76)</td>
<td>(3.21)</td>
<td>(2.94)</td>
<td>(3.25)</td>
</tr>
<tr>
<td>1994</td>
<td>-0.74</td>
<td>3.83</td>
<td>-0.42</td>
<td>3.45</td>
</tr>
<tr>
<td></td>
<td>(2.01)</td>
<td>(3.61)</td>
<td>(2.16)</td>
<td>(3.83)</td>
</tr>
<tr>
<td>1995</td>
<td>2.23</td>
<td>8.15</td>
<td>2.59</td>
<td>7.89</td>
</tr>
<tr>
<td></td>
<td>(2.14)</td>
<td>(3.96)</td>
<td>(2.57)</td>
<td>(4.11)</td>
</tr>
<tr>
<td>1996</td>
<td>0.73</td>
<td>7.56</td>
<td>1.33</td>
<td>7.32</td>
</tr>
<tr>
<td></td>
<td>(2.55)</td>
<td>(4.25)</td>
<td>(2.79)</td>
<td>(4.50)</td>
</tr>
<tr>
<td>1997</td>
<td>-2.88</td>
<td>4.51</td>
<td>-1.62</td>
<td>4.89</td>
</tr>
<tr>
<td></td>
<td>(2.61)</td>
<td>(4.48)</td>
<td>(2.94)</td>
<td>(4.71)</td>
</tr>
<tr>
<td>1998</td>
<td>-6.93***</td>
<td>0.92</td>
<td>-5.07</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>(2.31)</td>
<td>(4.48)</td>
<td>(2.58)</td>
<td>(4.52)</td>
</tr>
<tr>
<td>1999</td>
<td>-5.36</td>
<td>3.18</td>
<td>-2.98</td>
<td>4.71</td>
</tr>
<tr>
<td></td>
<td>(2.75)</td>
<td>(4.80)</td>
<td>(3.42)</td>
<td>(4.83)</td>
</tr>
<tr>
<td>2000</td>
<td>-4.48</td>
<td>4.52</td>
<td>-1.39</td>
<td>6.82</td>
</tr>
<tr>
<td></td>
<td>(2.36)</td>
<td>(5.04)</td>
<td>(3.22)</td>
<td>(5.12)</td>
</tr>
<tr>
<td>2001</td>
<td>-5.03</td>
<td>4.74</td>
<td>-1.45</td>
<td>7.48</td>
</tr>
<tr>
<td></td>
<td>(2.83)</td>
<td>(5.05)</td>
<td>(3.62)</td>
<td>(4.99)</td>
</tr>
<tr>
<td>2002</td>
<td>-5.16</td>
<td>5.49</td>
<td>-0.96</td>
<td>8.90</td>
</tr>
<tr>
<td></td>
<td>(2.40)</td>
<td>(5.49)</td>
<td>(3.50)</td>
<td>(5.64)</td>
</tr>
</tbody>
</table>

Demographic Controls? No Yes No Yes
Health Controls? No No Yes Yes

Source: National Center for Health Statistics Annual Natality Detail Files.
Notes: The dependent variable in each OLS regression is the birthrate for infants with spina bifida in each state and year. All regressions except Column 1 include state fixed effects. Robust standard errors are clustered at the state level. Demographic controls include: (a) average age at childbirth of mother in state s, year t; (b) share of women who are white giving birth in state s, year t; (c) share of women with some college education or less, at childbirth in state s, year t; and (d) share of married women at childbirth in state s, year t. Medical and health controls include: (a) average live birth order (including the current birth) in state s, year t;
(b) average number of prenatal visits in state \( s \), year \( t \); and (c) share of live births subject to amniocentesis in state \( s \), year \( t \). ** = significance at the 5% level and *** = significance at the 1% level.

**Table 3 (Cont.): Estimated Annual Changes in Spina Bifida Birthrates (1989–2002)**

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Maternal Age</td>
<td>-</td>
<td>-4.45</td>
<td>-5.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3.83)</td>
<td>(5.82)</td>
<td></td>
</tr>
<tr>
<td>Percentage Married</td>
<td>-</td>
<td>16.99</td>
<td>1.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(52.16)</td>
<td>(52.74)</td>
<td></td>
</tr>
<tr>
<td>Percentage White</td>
<td>-</td>
<td>-46.44</td>
<td>-33.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(35.24)</td>
<td>(33.80)</td>
<td></td>
</tr>
<tr>
<td>Share with College Education</td>
<td>-</td>
<td>-62.95</td>
<td>-60.58</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(33.79)</td>
<td>(34.89)</td>
<td></td>
</tr>
<tr>
<td>Average Birth Order</td>
<td>-</td>
<td>-30.23</td>
<td>-16.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(18.25)</td>
<td>(23.46)</td>
<td></td>
</tr>
<tr>
<td>Average Number of Prenatal Visits</td>
<td>-</td>
<td>-1.95</td>
<td>-0.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2.34)</td>
<td>(2.59)</td>
<td></td>
</tr>
<tr>
<td>Share using Amniocentesis</td>
<td>-</td>
<td>218.52</td>
<td>261.18**</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(133.54)</td>
<td>(112.21)</td>
<td></td>
</tr>
</tbody>
</table>

Source: National Center for Health Statistics Annual Natality Detail Files.

Notes: The dependent variable in each OLS regression is the birthrate for infants with spina bifida in each state and year. All regressions except Column 1 include state fixed effects. Robust standard errors are clustered at the state level. ** = significance at the 5% level and *** = significance at the 1% level.
<table>
<thead>
<tr>
<th>Year</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>-2.26</td>
<td>0.35</td>
<td>-2.21</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>(4.10)</td>
<td>(4.17)</td>
<td>(4.04)</td>
<td>(3.89)</td>
</tr>
<tr>
<td>1992</td>
<td>-6.60</td>
<td>-2.44</td>
<td>-6.71</td>
<td>-2.92</td>
</tr>
<tr>
<td></td>
<td>(4.11)</td>
<td>(5.43)</td>
<td>(4.27)</td>
<td>(5.30)</td>
</tr>
<tr>
<td>1993</td>
<td>-0.07</td>
<td>5.77</td>
<td>-0.09</td>
<td>5.52</td>
</tr>
<tr>
<td></td>
<td>(4.02)</td>
<td>(5.96)</td>
<td>(4.18)</td>
<td>(6.19)</td>
</tr>
<tr>
<td>1994</td>
<td>1.25</td>
<td>9.13</td>
<td>0.79</td>
<td>8.53</td>
</tr>
<tr>
<td></td>
<td>(4.35)</td>
<td>(7.52)</td>
<td>(5.31)</td>
<td>(7.83)</td>
</tr>
<tr>
<td>1995</td>
<td>-2.83</td>
<td>6.46</td>
<td>-3.39</td>
<td>5.91</td>
</tr>
<tr>
<td></td>
<td>(4.76)</td>
<td>(7.66)</td>
<td>(5.70)</td>
<td>(8.15)</td>
</tr>
<tr>
<td>1996</td>
<td>0.30</td>
<td>10.72</td>
<td>0.02</td>
<td>10.45</td>
</tr>
<tr>
<td></td>
<td>(5.11)</td>
<td>(9.60)</td>
<td>(6.08)</td>
<td>(10.02)</td>
</tr>
<tr>
<td>1997</td>
<td>-2.57</td>
<td>7.90</td>
<td>-1.59</td>
<td>8.84</td>
</tr>
<tr>
<td></td>
<td>(5.48)</td>
<td>(8.94)</td>
<td>(6.71)</td>
<td>(9.14)</td>
</tr>
<tr>
<td>1998</td>
<td>-2.95</td>
<td>8.34</td>
<td>-0.51</td>
<td>10.66</td>
</tr>
<tr>
<td></td>
<td>(4.50)</td>
<td>(9.21)</td>
<td>(5.90)</td>
<td>(9.56)</td>
</tr>
<tr>
<td>1999</td>
<td>-7.15</td>
<td>4.77</td>
<td>-3.30</td>
<td>8.63</td>
</tr>
<tr>
<td></td>
<td>(4.49)</td>
<td>(9.22)</td>
<td>(6.82)</td>
<td>(9.57)</td>
</tr>
<tr>
<td>2000</td>
<td>-3.94</td>
<td>8.04</td>
<td>1.71</td>
<td>13.88</td>
</tr>
<tr>
<td></td>
<td>(4.71)</td>
<td>(10.55)</td>
<td>(7.08)</td>
<td>(11.18)</td>
</tr>
<tr>
<td>2001</td>
<td>-3.13</td>
<td>9.83</td>
<td>3.77</td>
<td>17.08</td>
</tr>
<tr>
<td></td>
<td>(4.92)</td>
<td>(10.63)</td>
<td>(7.33)</td>
<td>(10.88)</td>
</tr>
<tr>
<td>2002</td>
<td>-2.18</td>
<td>11.88</td>
<td>6.33</td>
<td>20.90</td>
</tr>
<tr>
<td></td>
<td>(5.14)</td>
<td>(12.56)</td>
<td>(7.75)</td>
<td>(12.98)</td>
</tr>
<tr>
<td>Demographic Controls?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Health Controls?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
</tr>
</tbody>
</table>

**Source:** National Center for Health Statistics Annual Natality Detail Files.

**Notes:** The dependent variable in each OLS regression is the birthrate for infants with cleft palate in each state and year. All regressions except Column 1 include state fixed effects. Robust standard errors are clustered at the state level. Demographic controls include: (a) average age at childbirth of mother in state $s$, year $t$; (b) share of women who are white giving birth in state $s$, year $t$; (c) share of women with some college education or less, at childbirth in state $s$, year $t$; and (d) share of married women at childbirth in state $s$, year $t$. Medical and health controls include: (a) average live birth order (including the current birth) in state $s$, year $t$;
(b) average number of prenatal visits in state $s$, year $t$; and (c) share of live births subject to amniocentesis in state $s$, year $t$. ** = significance at the 5% level and *** = significance at the 1% level.

**Table 4 (Cont.): Estimated Annual Changes in Cleft Palate Birthrates (1989–2002)**

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Maternal Age</td>
<td>-</td>
<td>-1.62</td>
<td>-5.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(7.65)</td>
<td>(10.51)</td>
<td></td>
</tr>
<tr>
<td>Percentage Married</td>
<td>-</td>
<td>119.71</td>
<td>109.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(85.74)</td>
<td>(98.74)</td>
<td></td>
</tr>
<tr>
<td>Percentage White</td>
<td>-</td>
<td>-138.50*</td>
<td>-136.38*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(66.94)</td>
<td>(67.50)</td>
<td></td>
</tr>
<tr>
<td>Share with College Education</td>
<td>-</td>
<td>-46.68</td>
<td>-36.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(97.44)</td>
<td>(99.81)</td>
<td></td>
</tr>
<tr>
<td>Average Birth Order</td>
<td>-</td>
<td>-54.16</td>
<td>-41.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(40.94)</td>
<td>(47.29)</td>
<td></td>
</tr>
<tr>
<td>Average Number of Prenatal Visits</td>
<td>-</td>
<td>-0.26</td>
<td>2.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4.36)</td>
<td>(5.03)</td>
<td></td>
</tr>
<tr>
<td>Share using Amniocentesis</td>
<td>-</td>
<td>584.71**</td>
<td>589.23**</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(243.03)</td>
<td>(248.52)</td>
<td></td>
</tr>
</tbody>
</table>

*Source: National Center for Health Statistics Annual Natality Detail Files.*

*Notes:* The dependent variable in each OLS regression is the birthrate for infants with cleft palate in each state and year. All regressions except Column 1 include state fixed effects. Robust standard errors are clustered at the state level. ** = significance at the 5% level and *** = significance at the 1% level.
Table 5: Annual Changes in Down Syndrome Birthrates by Maternal Age (1989–2002)

<table>
<thead>
<tr>
<th>Year</th>
<th>Under 35</th>
<th>At least 35</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>1991</td>
<td>-6.15**</td>
<td>-6.73**</td>
</tr>
<tr>
<td></td>
<td>(2.51)</td>
<td>(2.85)</td>
</tr>
<tr>
<td></td>
<td>(2.89)</td>
<td>(4.03)</td>
</tr>
<tr>
<td>1993</td>
<td>-11.32***</td>
<td>-13.61***</td>
</tr>
<tr>
<td></td>
<td>(3.31)</td>
<td>(4.50)</td>
</tr>
<tr>
<td>1994</td>
<td>-12.20**</td>
<td>-15.42**</td>
</tr>
<tr>
<td></td>
<td>(3.29)</td>
<td>(5.92)</td>
</tr>
<tr>
<td>1996</td>
<td>-11.08**</td>
<td>-14.66**</td>
</tr>
<tr>
<td></td>
<td>(3.64)</td>
<td>(7.24)</td>
</tr>
<tr>
<td>1997</td>
<td>-13.98***</td>
<td>-17.49**</td>
</tr>
<tr>
<td></td>
<td>(3.38)</td>
<td>(7.13)</td>
</tr>
<tr>
<td></td>
<td>(2.90)</td>
<td>(7.08)</td>
</tr>
<tr>
<td>1999</td>
<td>-15.86</td>
<td>-18.15</td>
</tr>
<tr>
<td></td>
<td>(3.32)</td>
<td>(8.03)</td>
</tr>
<tr>
<td>2000</td>
<td>-12.93***</td>
<td>-14.85</td>
</tr>
<tr>
<td></td>
<td>(2.94)</td>
<td>(8.77)</td>
</tr>
<tr>
<td>2001</td>
<td>-13.50**</td>
<td>-14.85</td>
</tr>
<tr>
<td></td>
<td>(3.27)</td>
<td>(9.68)</td>
</tr>
<tr>
<td>2002</td>
<td>-14.73**</td>
<td>-15.77</td>
</tr>
<tr>
<td></td>
<td>(3.17)</td>
<td>(9.66)</td>
</tr>
</tbody>
</table>

Demographic Controls? | No | Yes | No | Yes
Health Controls? | No | Yes | No | Yes

N: 714

Source: National Center for Health Statistics Annual Natality Detail Files.
Notes: The dependent variable in each OLS regression is the birthrate for infants with Down syndrome in each state and year. All regressions except Columns 1 and 3 include state fixed effects. Robust standard errors are clustered at the state level. Demographic controls include: (a) average age at childbirth of mother in state s, year t; (b) share of women who are white giving birth in state s, year t; (c) share of women with some college education or less, at childbirth in state s, year t; and (d) share of married women at childbirth in state s, year t. Medical and health controls...
include: (a) average live birth order (including the current birth) in state $s$, year $t$; (b) average number of prenatal visits in state $s$, year $t$; and (c) share of live births subject to amniocentesis in state $s$, year $t$. ** = significance at the 5% level and *** = significance at the 1% level.
HEIGHT DISCRIMINATION IN EMPLOYMENT

Isaac B. Rosenberg*

This Article looks critically at heightism, i.e., prejudice or discrimination against a person on the basis of his or her height. Although much scholarship has focused on other forms of trait-based discrimination—most notably weight and appearance discrimination, both of which indirectly involve height as a component—little has focused on “pure” height discrimination. Nevertheless, within the past five years courts, scholars, and legislatures have increasingly tackled these non-traditional forms of discrimination. As such, this Article endeavors to fill the gap in the existing scholarship.

This Article specifically focuses on heightism in the workplace, with an emphasis on prejudice against short people because of the unique disadvantages they face vis-à-vis their taller counterparts. It examines the ways that existing federal antidiscrimination laws—namely Title VII of the Civil Rights Act of 1964 and the Americans with Disabilities Act of 1990—do and do not protect against height-based prejudice in the workplace. Moreover, after briefly examining state and local remedies for height discrimination, including state antidiscrimination laws, this Article considers but ultimately rejects enacting a federal law that would flatly prohibit all height-based employment decisions. Although a comprehensive prohibition would be easiest to administer, such a prohibition would prove both gratuitous and unwise. Instead, this Article recommends modest changes to federal regulations and increased state and local enforcement.

TABLE OF CONTENTS

I. INTRODUCTION ................................................................................................. 908
II. THE REALITIES OF HEIGHT DISCRIMINATION ................................................. 910
   A. Heightism, Generally .................................................................................. 910
   B. The Impact of Heightism in Employment .................................................... 914
III. HEIGHT DISCRIMINATION AND TITLE VII ..................................................... 915
   A. The “Traditional” Approaches ..................................................................... 917

* © 2009 Isaac B. Rosenberg, law clerk to the Honorable Barry G. Silverman, United States Court of Appeals for the Ninth Circuit; J.D., Order of the Coif, William & Mary School of Law (2008); B.S., summa cum laude, B.A., magna cum laude, University of Maryland, College Park (2004). The views expressed in this Article are those of the author alone and not those of the United States Court of Appeals for the Ninth Circuit or its judges. The author thanks Mohsin Syed, Esq., for his thoughtful comments on earlier drafts of this Article; Professors William Van Alstyne and B. Glenn George for their help in developing this topic; and the Honorable John R. Padova, United States District Court for the Eastern District of Pennsylvania, and his staff for their solicitude and guidance during the writing of this Article.
I. INTRODUCTION

“Well, I don’t want no Short People
Don’t want no Short People
Don’t want no Short People
Round here.”

At first blush, the concept of real height discrimination is almost laughable. After all, we do not typically think of height when we discuss types of discrimination. Yet there is no denying that we place a high premium on height,

1 RANDY NEWMAN, SHORT PEOPLE, ON LITTLE CRIMINALS (Warner Bros. Records 1977).
2 See LESLIE F. MARTEL & HENRY B. BILLER, STATURE AND STIGMA: THE BIOPSYCHOSOCIAL DEVELOPMENT OF SHORT MALES 2 (1984). Despite the palpably strong relationship between height, behavior, and personality, researchers largely ignored the issue until the advent of synthetic growth hormone therapy. See id. at 3–4. As one researcher speculated, “I think the whole problem makes everybody nervous all around with short people themselves wishing the issue would just go away, [and] normal sized people wishing short people would just go away.” Id. at 3 (quoting RALPH KEYES, THE HEIGHT OF YOUR LIFE 92 (1980)) (alterations in original).
be it social, sexual, or economic, and our preference for height pervades almost every aspect of our lives. Economist John Kenneth Galbraith—who towered at six feet eight inches—described the favored treatment we afford taller people as one of the “most blatant and forgiven prejudices” in our society. If you do not believe it, consider whether you yourself would like to be taller and try putting your finger on the reason why, or why not.

This Article looks critically at heightism, i.e., prejudice or discrimination against a person on the basis of his or her height. It specifically focuses on heightism in the workplace, particularly prejudice against short people because of the unique disadvantages they face compared with their taller counterparts. Although much scholarship has focused on other forms of trait-based discrimination—most notably discrimination based on weight and appearance, both of which indirectly involve height as a component—little if any scholarship has focused on “pure” height discrimination. Thus, this Article aims to fill that

---


4 See id. (internal quotation marks omitted) (quoting Arthur Ungur, Galbraith: Turning Economics to Show Biz, CHRISTIAN SCIENCE MONITOR, May 18, 1977, at 22). Galbraith went on to observe: “We tall men, being higher than anybody else, are much more visible and thus more closely watched. Therefore it follows that our behavior is naturally superior. So the world instinctively and rightly trusts tall men.” Id.; see also Robert Fulford, It’s A Tall World, After All, THE GLOBE AND MAIL, Oct. 13, 1993, at C1. (“Height remains one of the last frontiers of unabashed prejudice.”).


6 Credit for “heightism” goes to sociologist Saul Feldman, who coined the phrase in the early 1970s. See Heightism, TIME, Oct. 4, 1971, at 64.


9 Paul Steven Miller, a former commissioner of the U.S. Equal Employment Opportunity Commission who now serves as both Henry M. Jackson Professor of Law and Director of the Disability Studies Program at the University of Washington, wrote a thoughtful piece more than twenty years ago about workplace discrimination against little people who suffer from no other physical impairments, emphasizing liability under the Rehabilitation Act of 1973. See Paul Steven Miller, Note, Coming Up Short: Employment Discrimination Against Little People, 22 HARV. C.R.-C.L. L. REV. 231 (1987).
gap by examining how existing federal antidiscrimination laws—namely Title VII and the Americans with Disabilities Act of 1990—do and do not protect against height-based prejudice in the workplace.

Part II explores the pervasiveness of heightism generally and its specific impact on hiring, wages, and other aspects of employment. Part III looks at the various ways plaintiffs have pursued height-based claims under Title VII and suggests a new approach to such claims: characterizing heightism, in some cases, as a kind of impermissible gender stereotyping. Part IV considers height under the ADA—including the ADA Amendments Act of 2008—and contends that height outside the “normal range,” as defined by this Article, qualifies as an “impairment”; Part IV also considers height-based claims under the ADA’s “regarded as” prong. Finally, Part V briefly examines state and local remedies for height discrimination, including state common law and antidiscrimination laws, and considers whether Congress should enact a comprehensive height discrimination law flatly prohibiting height-based employment decisions. Part V concludes that, although a comprehensive prohibition would be easiest to administer, such a prohibition would prove both gratuitous and unwise. Rather, modest amendments to the federal regulations and increased state and local enforcement should suffice.

II. THE REALITIES OF HEIGHT DISCRIMINATION

A. Heightism, Generally

Heightism is instinctive. We cannot help making subconscious height-based comparisons.10 We engage in “gaze behavior”—a primitive way of establishing social hierarchies on the basis of whether we are looking up to or down on another—whenever we encounter someone.11 To those we look down on, we ascribe less social power and negative character traits.12 We even afford short people less personal space.13 Those we look up to, however, enjoy a “halo effect,”

10 See Martel & Biller, supra note 2, at 35 (citing Seymour Fisher, Body Experience in Fantasy and Behavior (1970)).
11 See id. at 34.
12 See id. at 34–35 (citing Ralph Keyes, The Height of Your Life 92 (1980)); see also David E. Sandberg & Melissa Colsman, Assessment of Psychosocial Aspects of Short Stature, 21 Growth, Genetics & Hormones 17, 19 (2005) (“With few exceptions, both children and adults attribute significantly less favorable characteristics to short individuals compared to those of tall or average height.”) (citations omitted)); Short Guys Finish Last; Heightism, The Economist, Dec. 23, 1995, at 19 [hereinafter Short Guys Finish Last] (“Both men and women, whether short or tall, thought that short men—heights between 5’2’ and 5’5”—were less mature, less positive, less secure, less masculine; less successful, less capable, less confident, less outgoing; more inhibited, more timid, more passive; and so on.”).
13 See Martel & Biller, supra note 2, at 35 (citing Kent G. Bailey et al., Body Size as Implied Threat: Effects on Personal Space and Person Perception, 43 Perceptual &
the automatic attribution of positive personality characteristics to them because of their height. 14 This is perhaps most evident in our selection of presidents. 15 We almost always elect the taller presidential candidate. 16 In fact, we have not elected a shorter-than-average president since 1896 17 and have elected scarcely a half-dozen short presidents overall. 18 Furthermore, a candidate’s margin of victory derives, in part, from his height. 19

Heightism is also inculcated. Our language is rife with heightist idioms. 20 As children, we are constantly reminded of how much we have grown, and we are

---

Motor Skills 223, 223–30 (1976)). One study showed that we accord more than twice as much personal space to tall people than we do to short people. See id. (quoting J.J. Hartnett et al., Body Height, Position, and Sex as Determinants of Personal Space, 87 J. Psychol. 129, 134 (1974)).

14 See id. at 36 (citing Edward L. Thorndike, A Constant Error in Psychological Ratings, 4 J. Applied Psychol. 25, 25 (1920)).


17 See Timothy A. Judge & Daniel M. Cable, The Effect of Physical Height on Workplace Success and Income: Preliminary Test of a Theoretical Model, 89 J. Applied Psychol. 428, 428 (2004). President McKinley, who was first elected in 1896, was only 5’7”—and was ridiculed for being a “little boy.” See id.

18 See Persico, supra note 16, at 1021, fig.1. James Madison was the shortest at 5’4”.

See id.


20 Compare sayings that underscore the disadvantages of being short or small (short shrift; coming up short; short end of the stick; caught short; draw the short straw; short change; feel small) with those that highlight the virtues of being big or tall (look up to someone; big man on campus; head and shoulders above the rest; stand tall; be the bigger person; make it big).
encouraged to “eat our vegetables” and “drink milk” so we can grow up to be “big and strong.”

Even science has contributed to negative perceptions of short stature. In the early twentieth century, eugenists identified short stature as an inferior trait, and scientists thereafter set out to fix the “problem.” In the 1950s, scientists began treating some slow-growing children with human growth hormone (HGH) extracted from the pituitary glands of cadavers, but such treatment was administered only to those with diagnosed medical deficiencies because of limited supply. In the mid-1980s, however, the number of people receiving HGH treatment exploded with the advent of synthetic HGH, and included many who suffered from no diagnosable growth disorders. Although much of the ensuing debate over HGH focused on whether shortness (as opposed to growth deficiency) constituted a disease in itself, the underlying issue about which all seemed to agree was how bad it is to be short. Proponents of greater availability of HGH focused on the fact that being short “handicaps a [person] in the competition for schools, jobs, income, and ‘mates.’” Ultimately, the wider availability of HGH

21 See Martel & Biller, supra note 2, at 19; see also Short Guys Finish Last, supra note 12 (“As boys grow, the importance of height is drummed into them incessantly. ‘My, how tall you are!’ the relatives squeal with approval. Or, with scorn, ‘Don’t you want to grow up big and strong?’”).

22 Eugenist Charles Davenport investigated the inheritance of height. See Charles B. Davenport, Inheritance of Stature, 2 Genetics 313, 315–17 (1917). Davenport’s peers, Madison Grant and Lothrop Stoddard, promoted the inferiority of certain ethnic groups because of their “dwarfish stature.” See also Robert Michael, A Concise History of American Antisemitism 130 (2005) (noting that one scholar viewed Jews as having a “dwarfish stature”). Vernon Kellogg went so far as to criticize military conscription, on the eve of World War I, because drafts and war lessened the genetic stock of a nation; the derivative loss of tall, strong men to war diminished “the stature of the next generation.” Vernon Kellogg, Eugenics and Militarism, The Atlantic Monthly, July 1913, at 105.


25 See id.


27 See id.

28 Id. (citations omitted).
treatment made more people view short stature as a problem to be fixed.\textsuperscript{29} Perhaps that explains why more and more short—but otherwise healthy—adults have turned to “limb lengthening” to combat the stigma of short stature.\textsuperscript{30}

Yet even though our instinctive and inculcated preference for height is pervasive, we do not generally acknowledge its existence.\textsuperscript{31} There may be two reasons for this:

Either the awareness regarding discrimination is not in the consciousness of one or both individuals in a particular social situation, or verbalization of the discrimination is suppressed. The result is that the short [person] feels that something is subtly awry, but he cannot pin it down. He may believe that this discrimination is based on the social feedback that he does not look quite right, that he falls significantly short of the cultural ideal for height.\textsuperscript{32}

Strive as victims of height discrimination may, combating heightism is “like fighting a ghost.”\textsuperscript{33}

\textsuperscript{29} Joel Frader, a doctor and ethicist in the Medical Humanities and Bioethics Program at Northwestern University, characterized the FDA’s decision to approve expanded availability and use of HGH as “tragic” because it “medicalized short stature and turned it into an illness.” HALL, supra note 5, at 245.

\textsuperscript{30} Limb lengthening is an expensive and painful procedure by which a surgeon “divides a long bone into two or more sections, separates the sections slightly and braces the bone and limb with metal ‘scaffolding,’” then adjusts the pins and screws on this frame “to keep tension between the sections, enabling the bone to grow back together gradually into a complete but longer bone.” MayoClinic.com, Dwarfism: Treatments and Drugs, http://www.mayoclinic.com/health/dwarfism/DS01012/DSECTION=treatments-and-drugs (last visited Sept. 1, 2009); see also Caitlin Gibson, Growing Pains: For Caitlin Schroeder, Achieving Near-Average Height Would Require No Small Act of Courage, WASH. POST, Nov. 30, 2008, at W10 (describing the procedure and its sometimes painful consequences); Joe Kita, All to be Tall, MEN’S HEALTH, Jan/Feb 2004, at 132–35 (describing leg lengthening procedure).

Limb lengthening is particularly popular in China, where thousands of otherwise healthy Chinese people who fear height discrimination undergo the procedure. See Xun Zeng, Enforcing Equal Employment Opportunities in China, 9 U. PA. J. LAB. & EMP. L. 991, 1002 (2007); see also Joe Kita & Lee Kynaston, Men Who Risk It All to be Tall, THE INDEPENDENT, Apr. 5, 2004, at 8 (noting an increase among short men in England who turn to limb lengthening as a cosmetic procedure).

\textsuperscript{31} See MARTEL & BILLER, supra note 2, at 38 (“Discrimination against short males, although often subtle, remains a powerful factor in their lives. . . . Society positively frames an identity for the short female by labeling her as ‘cute’ or ‘dainty,’ while the short boy is just plain short.”).

\textsuperscript{32} Id.

\textsuperscript{33} Id. (quoting RALPH KEYES, THE HEIGHT OF YOUR LIFE 92 (1980)).
B. The Impact of Heightism in Employment

No matter its source, the problem is prejudice.34 Height-based prejudice permeates employment decisions—perhaps as much as race and gender.35 It begins with hiring. For example, when researchers asked a group of recruiters to make a hypothetical hiring decision between two equally qualified candidates who differed only in height, 72 percent of the recruiters chose the taller candidate.36

Height also affects wages. Data suggest that every additional inch in height is associated with a 1.8 to 2.2 percent increase in wages37—or roughly $789 per inch, per year.38 Moreover, the tallest 25 percent of the population gets a 13 percent boost in median income compared with the shortest 25 percent.39 In socially oriented jobs such as sales and management, height was shown to be predictive of earnings.40 Although some speculate that taller people earn more because of a correlation between height and intelligence,41 studies controlling for intelligence continue to find a significant relationship between height and earnings.42 Similarly controlling for gender, height continues to affect wages.43 Height’s effect does not decline over time; in fact, its importance may even increase as we age.44

Finally, height affects professional advancement. Height impacts self-esteem (how individuals regard themselves) and social esteem (how individuals are regarded by others), which in turn affect actual job performance, perceived job performance, and, ultimately, professional success.45 It is hardly a coincidence that 58 percent of Fortune 500 CEOs are six feet or taller (compared with roughly 14.5 percent of all men) and 30 percent are 6’2” or taller (compared with 3.9 percent of

34 Id. at 36 (quoting JOHN S. GILLIS, TOO TALL TOO SMALL 61 (1982)).
35 See Persico, supra note 16, at 1020–21. In fact, one economist has suggested that “[t]he gross mistake is that much of what we normally assume is sex discrimination is height discrimination. Of course, heightism affects both men and women, but because women average 4 to 5 inches shorter than men, it affects [women] more.” Dennis D. Miller, Ending Job Bias Can Be a Tall Order, WALL ST. J., Oct. 28, 1995, at 1.
36 See MARTEL & BILLER, supra note 2, at 38; see also Short Guys Finish Last, supra note 12, at 20 (noting the results of one study conducted in 1969).
37 See Persico, supra note 16, at 1021.
39 HALL, supra note 5, at 185.
40 See Judge & Cable, supra note 17, at 436–37. Although height was found to be less significant in less socially driven occupations, it still had an effect. Id.
42 See Judge & Cable, supra note 17, at 436.
43 See id.
44 See id.
45 See id.
all men). \(^{46}\) One business expert has suggested that an additional four inches in height “make[s] much more difference in terms of success in a business career than any paper qualifications you have” and that it would be better to be “5 ft. 10 and a graduate of N.Y.U.’s business school than 5 ft. 6 and a Harvard Business School graduate.” \(^{47}\) Another commentator concluded that “being short is probably as much, or more, of a handicap to corporate success as being a woman or an African American.” \(^{48}\)

In sum, heightism tangibly affects employment decisions involving short employees. Because this type of discrimination is subtle, however, many who fall prey to it may not realize, or even think to realize, that it motivated decisions against them. \(^{49}\) Nevertheless, where victims of height discrimination suspect that adverse employment actions were motivated by their height, the question remains, what remedies are available under federal law?

### III. HEIGHT DISCRIMINATION AND TITLE VII

Title VII of the Civil Rights Act of 1964 generally prohibits employers from discriminating against applicants or employees because of their race, religion, national origin, color, or sex. \(^{50}\) An employer can violate Title VII under two basic theories. The first theory, disparate impact, involves “employment practices that are facially neutral in their treatment of different groups but that in fact fall more harshly on one group than another.” \(^{51}\) This theory requires no showing of discriminatory motive. \(^{52}\) The second theory, disparate treatment, involves treating some less favorably than others because of their protected status. \(^{53}\) This theory necessarily requires a showing of discriminatory motive, be it express or implied. \(^{54}\)

---

\(^{46}\) See Gladwell, supra note 38, at 86–87. A 1980 survey showed that only 3 percent of Fortune 500 CEOs were 5’7” or shorter. See Short Guys Finish Last, supra note 12, at 20.

\(^{47}\) Georgia Harbison, A Chance to Be Taller, TIME, Jan. 8, 1990, at 70. Moreover, CEOs would rather be bald than short. See Del Jones, The Bald Truth About CEOs; Executives Say They’d Rather Have No Hair Than Be Short, USA TODAY, Mar. 14, 2008, at 1B.

\(^{48}\) Gladwell, supra note 38, at 87; see also Judge & Cable, supra note 17, at 438.

\(^{49}\) This may explain why, in those jurisdictions explicitly protecting against height discrimination, so few cases have been brought. See infra notes 287–292 and accompanying text.


\(^{52}\) See Teamsters, 431 U.S. at 335–36 n.15.

\(^{53}\) See id.

Under either theory, a plaintiff must make out a prima facie case, which an employer can rebut by (1) defeating the inference of discrimination, in disparate treatment cases; (2) undermining the evidence of causation, in disparate impact cases; or (3) marshaling an affirmative defense. This Part, however, examines only how height fits into the prima facie case as a matter of law. The post-prima facie case inquiry is incredibly fact sensitive and therefore beyond the scope of this Article, which considers merely whether height-based claims are legally cognizable.

To that end, this Part has two subsections. The first subsection analyzes claims brought under what one might call the “traditional” approaches, which have involved challenges to (1) facially neutral height restrictions under a disparate impact theory; (2) the uneven application of height restrictions as a pretext under a disparate treatment theory; and (3) height-based animus that has some demonstrable nexus to a protected trait under a disparate treatment theory. The second subsection suggests a new approach to height-based claims: challenging height-based disparate treatment as a form of impermissible gender stereotyping.

55 For a disparate treatment theory, a plaintiff must show: “(1) membership in a protected group; (2) qualification for the job in question; (3) an adverse employment action, [i.e., a failure to hire, a failure to promote, or termination]; and (4) circumstances that support an inference of discrimination.” Swierkiewicz v. Sorema, N.A., 534 U.S. 506, 510 (2002) (citing McDonnell Douglas Corp. v. Green, 411 U.S. 792, 802 (1973); Texas Dep’t of Cmty. Affairs v. Burdine, 450 U.S. 248, 253–54 n.6 (1981)). For a disparate impact theory, a plaintiff must show merely that a specific facially neutral employment practice negatively and disproportionately affects one group over another. See 42 U.S.C. § 2000e-2(k)(1)(A) (2006); see also Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 994 (1988) (determining that a plaintiff must identify a specific employment practice).

56 See McDonnell Douglas, 411 U.S. at 802 (“The burden then must shift to the employer to articulate some legitimate, nondiscriminatory reason for the employee’s rejection.”).

57 See 42 U.S.C. § 2000e-2(k)(1)(B) (2006); cf. Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 653 & n.8 (1989) (recognizing that an employer could rebut a disparate impact claim “if [the employer] could show that the percentage of selected applicants who are nonwhite is not significantly less than the percentage of qualified applicants who are nonwhite”); Int’l Bhd. of Teamsters v. United States, 431 U.S. 324, 340 (1977) (“We caution only that statistics are not irrefutable; they come in infinite variety and, like any other kind of evidence, they may be rebutted.”).

58 For example, the employer could show in a disparate treatment case that religion, sex, or national origin (but not race or color) is a “bona fide occupational qualification reasonably necessary to the normal operation of that particular business or enterprise.” 42 U.S.C. § 2000e-2(e)(1) (2006). In disparate impact cases, the employer could show that the “challenged practice is job related for the position in question . . . consistent with business necessity,” id. § 2000e-2(k)(1)(A)(i), and that “there are [no] other ways for the employer to achieve its goals that do not result in a disparate impact on a protected class.” Smith v. City of Jackson, 544 U.S. 228, 243 (2005).

59 Cf. McDonnell Douglas, 411 U.S. at 802–03 (“We need not attempt in the instant case to detail every manner which fairly could be recognized as a reasonable basis for a refusal to hire.”).
A. The “Traditional” Approaches

1. Height Restrictions Under a Disparate Impact Theory

As noted above, a disparate impact theory requires the plaintiff to point to a specific employment practice that adversely and disproportionately affects a protected group. This most commonly arises when an employer implements some sort of minimum height restriction. Such restrictions tend to adversely impact women and certain racial and ethnic groups.

With respect to gender, in the seminal Dothard v. Rawlinson, an Alabama statute required that prison guards stand at least 5’2” and weigh 120 pounds. The Supreme Court found that the minimum height and weight restrictions violated Title VII because, when combined, they would exclude 41.13 percent of the female population while excluding less than 1 percent of the male population.

Conversely, in Livingston v. Roadway Express, Inc., a 6’7” white male plaintiff was rejected for employment as a truck driver because of the company’s 6’4” maximum height limitation for the position. Livingston alleged reverse sex discrimination under Title VII because the restriction had a disparate impact on males. Statistical evidence showed that 0.9 percent of adult men were 6’4” or taller, whereas only 0.3 percent of women were 5’11” or taller; thus, the height maximum excluded three times as many men as women. Nevertheless, the Tenth Circuit rejected the reverse discrimination claim because it determined that “in impact cases, as in disparate treatment cases, a member of a favored group [like men] must show background circumstances supporting the inference that a facially neutral policy with a disparate impact is in fact a vehicle for unlawful discrimination,” which the plaintiff had failed to do.

Plaintiffs also have challenged minimum height restrictions for their disparate impact on people of certain races and national origins—namely, Asians and

---

60 See supra note 51 and accompanying text.
62 Id. at 323–24.
63 Id. at 329–31; cf. 28 C.F.R. § 50.14, pt.1 § 4(D) (2008) (“A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact . . . .”).
64 802 F.2d 1250 (10th Cir. 1986).
65 Id. at 1251.
66 Id.
67 Id.
68 Id. at 1252–53.
Hispanics. In *League of United Latin American Citizens v. City of Santa Ana*, the City of Santa Ana imposed minimum height requirements for its police and fire departments. Although Mexican Americans constituted 25.8 percent of the city’s general population, they made up only 9.2 percent of the city’s police force and 4.5 percent of its fire department. The district court found that the height requirements had served to deter Mexican American applicants, and disqualified two to three times as many Mexican American applicants as Caucasian applicants. Ultimately, the court determined “it [was] clear that by . . . the use of an arbitrary height requirement, the defendants were responsible for preventing substantial numbers of Mexican-Americans from taking the [qualification] tests in the first place.”

Likewise, in *Sondel v. Northwest Airlines*, a 4’11” woman of Sri Lankan descent applied for but was denied a position as a flight attendant and filed charges with the EEOC alleging that the airline’s height requirement discriminated against women, Asians, and Hispanics. The EEOC found that the airline’s 5’2” minimum height requirement did in fact exclude women, especially Hispanic and Asian women, from employment opportunities as flight attendants. And in *Officers for Justice v. Civil Service Commission of the City and County of San Francisco*, plaintiffs challenged the defendant commission’s use of a 5’6” preselection height minimum for patrol officers as discriminatory against Asians, Hispanics, and women. The district court agreed.

---

70  Id. at 879.
71  Id.
72  Id. at 881.
73  Id. at 882.
74  Id. at 893.
76  *Sondel II*, 1993 U.S. Dist. LEXIS 21252, at *2. In *Sondel II*, the *Sondel I* plaintiff joined forces with other rejected applicants to seek class certification, alleging discrimination on the basis of race, color, gender, and national origin. *Sondel II*, 1993 U.S. Dist. LEXIS 21252, at *1. Among the other plaintiffs was Stephanie Chung, a 5’0” woman of Korean descent. Id. at *4.
80  Id. at 380.
81  Id. In so finding, the district court did not even consider the general census data offered by the plaintiffs, which indicated that the average height of Asians and Hispanics was lower than that of African Americans or Caucasians, because the discriminatory impact “[was] clear from the height data on the applicants’ themselves. Id. But see *Arnold v. Ballard*, 390 F. Supp. 723, 727, 738 (N.D. Ohio 1975) (finding that the Akron Police Department’s minimum height requirement did not disqualify a disproportionate number of
2. Height Restrictions as Pretext Under a Disparate Treatment Theory

Some plaintiffs have successfully brought disparate treatment claims by showing that a facially neutral height restriction was a pretext for intentional discrimination where the employer applied the restriction unevenly to different groups. For example, in United States v. Lee Way Motor Freight, Inc., a trucking company had instituted a 5’7” minimum height requirement for its drivers, which it strictly enforced against minority applicants, the company selectively enforced the height restriction for white applicants who did not meet the minimum. The court found that the selective application of the height restriction to minorities violated Title VII.

Similarly, in Schick v. Bronstein, the male plaintiff applied for a position as a patrolman in the New York City Police Department but was rejected because he failed to meet the 5’7” minimum height requirement in effect when he applied. The police department also did not hire women as patrol officers when the plaintiff submitted his application. Shortly after he was rejected, the police department began hiring women as patrol officers and dropped the height requirement for all applicants. When the plaintiff reapplied for the job, however, the police department continued to apply the height restriction to him. The court determined that the department’s continued application of the height restriction to the plaintiff, when no such height requirement was applied to female applicants, “constituted a refusal of employment based on sex, in violation of Title VII.”

African American applicants and was not “used to facilitate a discriminatory attitude against” African Americans.

---

83 Id. at *13, *20.
84 Id. at *20. See also Chi. Fire Fighters Union Local No. 2 v. City of Chicago, Civ. A. Nos. 87-7295, 89-7984, 93-5438, 93-6175, 96-808, 1999 U.S. Dist. LEXIS 20310, at *181 (N.D. Ill. Dec. 30, 1999) (upholding city’s race-based affirmative action program for promotions within the fire department because a former fire commissioner had instituted pretextual job requirements—including minimum height requirements—to exclude minority applicants); McNamara v. City of Chicago, 959 F. Supp. 870, 875 (N.D. Ill. 1997) (same).
85 See United States v. Lee Way Motor Freight, Inc., 625 F.2d 918, 924 (10th Cir. 1979) (“The magnitude of the statistics established a prima facie case that during this period race was a factor in staffing the two driver categories.”).
87 Id. at 334.
88 Id. at 335. Women were hired as “policewomen,” an ostensibly different job, and were required to meet a 5’2” minimum height requirement. Id.
89 Id.
90 Id.
91 Id. at 338. See also U.S. EQUAL EMPLOYMENT OPPORTUNITY COMM’N, EEOC COMPLIANCE MANUAL § 621(b)(1) (1991) [hereinafter “EEOC Interpretive Manual”].
The same theory applies to height maximums. For example, in *Laffey v. Northwest Airlines, Inc.*, an airline did not hire women as flight attendants if they were taller than 5’9” but did hire men who were as tall as 6’0”. The court determined that the airline had violated Title VII by imposing a shorter maximum height requirement on women than it imposed on men.

3. Height-Based Animus with a Demonstrable Nexus to a Protected Trait

Many Title VII suits have challenged height-based animus as a pretext for intentional discrimination on the basis of some protected trait. Most of these cases have failed, however, because the plaintiffs could not establish a nexus between the alleged height-based animus and a trait protected by the statute; that is, they failed to establish the requisite inference of discriminatory motive. In *Ekerman v. City of Chicago*, for example, the plaintiff was a 4’10”, ninety-two-pound female detective in the Chicago Police Department who brought a Title VII claim alleging gender discrimination. She claimed that her work environment was unlawfully hostile because of a single, gender-neutral remark allegedly made by the deputy police chief, who “called out from behind her and said: ‘Boy are you short. How tall are you? How can you do this job?’” The district court granted summary judgment to the city, finding that Title VII did not generally prohibit height discrimination, that the statement was facially gender neutral, and that the plaintiff “offered no evidence to support a finding that the remark, despite being sex-neutral on its face, was actually made on account of her gender.”

Similarly, in *Cortez v. Wal-Mart Stores, Inc.*, the plaintiff had been repeatedly denied promotions to the positions of general manager of a Sam’s Club and regional personnel manager at Walmart’s home office. He alleged that Walmart did not promote him on the basis of his age and race. The district court dismissed his race discrimination claim for lack of supporting evidence and found that, despite the fact that a regional vice president referred to him as “Shortez,” the record did not support his bare allegations of endemic managerial

---

93 Id. at 775.
94 Id. at 790.
96 Id. at *2, *8.
97 Id. at *7.
98 Id. at *8. The court also dismissed the plaintiff’s retaliation claim. Id. at *12–14.
100 Id. at *6.
101 Id.
102 Id. at *17–18.
favoritism for “non-Hispanic” characteristics, such as above-average height. The district court in *Pena v. USX Corp.* reached a similar conclusion on comparable facts.

**B. A New Approach: Height as a Form of Gender Stereotyping**

Although these traditional approaches have proven moderately successful, they continue to underachieve because they fail to capture one of the most significant components of height discrimination: gender stereotyping. As this subsection shows, many cases of height-based animus, especially animus directed toward short men and tall women, result from gender stereotyping.

1. **The Theory**

Height constitutes one of many significant factors that go into overall physical attractiveness, particularly for men, and “[i]t is almost axiomatic that short males are not attractive, or at least not as attractive as their taller counterparts.” After all, we often describe the quintessential man as “tall, dark, and handsome.” Relative height preoccupies men, and short men typically struggle to form a sense of physical adequacy and competency. Women perceive short men as undesirable mates because they “do not strike [them] as true men.” Research plainly shows that “[t]he universally acknowledged cardinal rule of dating and mate selection is that the male will be significantly taller than his female partner” and that “women are actively selecting for tallness when they go looking for male partners.”

---

103 Id. at *23 n.8.
105 Id. at *11 (“In addition, the court declines to expand the ban on racial discrimination articulated in [Title VII and § 1981] to include harassment based on height for no better reason than the plaintiff’s unsupported conclusion that in his coworkers’ eyes, Mexicans were short and therefore all harassment based on height was racially motivated.”).
106 Martel & Biller, supra note 2, at 5.
107 Id. (observing that *Ms. Magazine* ran an article entitled “Short, Dark and Almost Handsome”).
108 Id.
109 Id. at 25 (citing H.G. Beigel, *Body Height in Mate Selection*, 29 J. Soc. Psychol. 257, 268 (1954)). For example, when 100 women of all heights between ages 18 and 22 were shown pictures of men whom they believed to be either short, average, or tall, all of the women found the tall men to be significantly more attractive than the short men. See id. at 26.
110 Id.
More importantly, however, “large body size has a symbolic meaning to males that is unique to their gender.”112 Put simply, we (both males and females) view large men as more manly.113 As such, short men struggle to negotiate and solidify positive male identity.114 Because we value bigness in men—in part because much of the normative male gender role involves offering security to self and others, which we assume smaller men are less capable (or even incapable) of providing—shortness in males manifests a failure to satisfy the norm.115

Conversely, because women are generally shorter than men and desire to be small,116 a tall woman may be perceived as defying her gender norm because of her uncharacteristic stature.117 Consider that endocrinologists in the 1950s prescribed hormones as growth suppressants for tall girls.118 Data collected in 1978 showed that up to one half of pediatric endocrinologists had offered estrogen to young women whose adult height they forecast to be greater than 6’1”.119 More recent data suggest this practice is out of fashion, although one in five pediatric endocrinologists had reported treating at least one girl for “tall stature” within the past five years.120

Gender stereotyping theories of liability under Title VII have become increasingly popular and effective in recent years. Beginning in Price Waterhouse v. Hopkins,121 the Supreme Court held that, “[i]n the specific context of sex stereotyping, an employer who acts on the basis of a belief that a woman cannot be

112 MARTEL & BILLER, supra note 2, at 6. The preference for male height spans every culture. See id. at 33 (citing C.S. FORD & F.A. BEACH, PATTERNS OF SEXUAL BEHAVIOR (1951)). The preference for female body type is less consistent cross-culturally. Id.
113 Id. at 6 (citing SEYMOUR FISHER, BODY CONSCIOUSNESS: YOU ARE WHAT YOU FEEL 119 (1973)).
114 Id. at 6.
115 Id. at 8. “[T]he male whose body type does not conform to the traditional image of the ideal male, that of the tall mesomorph, may face severe difficulty in accepting himself and having others accept him as truly masculine and competent in the male role.” Id. at 32 (quoting A. Gascaly & C.A. Borges, The Male Physique and Behavioral Experience Expectancies, 106 J. PSYCHOL. 97, 101 (1979)).
116 See MARTEL & BILLER, supra note 2, at 5 (citing G. Calden et al., Sex Differences in Body Concepts, 23 J. CONSULTING PSYCHOL. 378 (1959)).
117 See id. (“Femininity is often associated with petiteness and, at the very least, being large is seen as unfeminine whether it is being very tall or overweight.”); see also E. Brecher, Will Perot Be The Short-Cut Chief, or Just Cut Short?, COURIER-MAIL, June 11, 1992 (“[S]tudies have shown that bosses often reject tall women because they’re too threatening.”).
118 See ELLIOTT, supra note 26, at 242.
119 Id. (citing F.A. Conte & M.M. Grumbach, Estrogen Use in Children and Adolescents: A Survey, 62 PEDIATRICS 1091 (1978)).
120 Id. (citing Neal D. Barnard et al., The Current Use of Estrogens for Growth-Suppressant Therapy in Adolescent Girls, 15 J. PEDIATRIC & ADOLESCENT GYNECOLOGY 23, 23–26 (2002)).
121 490 U.S. 228 (1989).
aggressive, or that she must not be, has acted on the basis of gender.” The plaintiff, a senior manager for the defendant consulting firm, was passed over for partnership because, despite exhibiting desirable traits such as “strong character, independence and integrity,” she just did not act “like a woman should.” The Court concluded that “[a]n employer who objects to aggressiveness in women but whose positions require this trait places women in an intolerable and impermissible catch 22: out of a job if they behave aggressively and out of a job if they do not. Title VII lifts women out of this bind.” Subsequently, in Oncale v. Sundowner Offshore Services, Inc., a unanimous Supreme Court recognized that there could be no absolute presumption under Title VII that a person of one gender would not discriminate against another person of the same gender. As such, the Court determined that actionable sexual harassment “need not be motivated only by sexual desire to support an inference of discrimination . . . .”

Victims of discrimination on the basis of sexual orientation have successfully used this gender-stereotyping theory—despite the fact that sexual orientation is not protected under Title VII—to challenge workplace harassment under Title VII. Before Price Waterhouse and Oncale, however, same-sex harassment claims

---

122 Id. at 250.
123 Id. at 234.
124 See id. at 235 (“[S]ome of the partners reacted negatively to Hopkins’ personality because she was a woman. One partner described her as ‘macho’; another suggested that she ‘overcompensated for being a woman’; a third advised her to take ‘a course at charm school.’ Several partners criticized her use of profanity; in response, one partner suggested that those partners objected to her swearing only ‘because it’s a lady using foul language.’ Another supporter explained that Hopkins ‘ha[d] matured from a tough-talking somewhat masculine hard-nosed mgr to an authoritative, formidable, but much more appealing lady ptr candidate.’”).
125 Id. at 251. The Court further noted that “we are beyond the day when an employer could evaluate employees by assuming or insisting that they matched the stereotype associated with their group, for ‘in forbidding employers to discriminate against individuals because of their sex, Congress intended to strike at the entire spectrum of disparate treatment of men and women resulting from sex stereotypes.’” Id. (emphasis added) (quoting L.A. Dep’t of Water & Power v. Manhart, 435 U.S. 702, 707 n.13 (1978)) (some internal quotation marks omitted).
127 Id. at 78.
128 Id. at 80.
129 See Etsitty v. Utah Transit Auth., 502 F.3d 1215, 1222 (10th Cir. 2007) (“Further, this court has explicitly declined to extend Title VII protections to discrimination based on a person’s sexual orientation.” (citing Medina v. Income Support Div., 413 F.3d 1131, 1135 (10th Cir. 2005)); Vickers v. Fairfield Med. Ctr., 453 F.3d 757, 762 (6th Cir. 2006) (“[S]exual orientation is not a prohibited basis for discriminatory acts under Title VII.”); Bibby v. Phila. Coca Cola Bottling Co., 260 F.3d 257, 260–61 (3d Cir. 2001) (same); Simonton v. Runyon, 232 F.3d 33, 35 (2d Cir. 2000) (“The law is well-settled in this circuit and in all others to have reached the question that . . . Title VII does not prohibit harassment or discrimination because of sexual orientation.”).
typically failed at the motion to dismiss stage. In *Smith v. Liberty Mutual*,130 for example, the plaintiff applied for a position as a mailroom clerk with the defendant employer, but was not hired because the interviewing supervisor thought his behavior was “effeminate.”131 The Fifth Circuit rejected the plaintiff’s gender discrimination claim because he failed to allege that he was discriminated against “because he was a male.”132 Instead, the plaintiff alleged discrimination because “as a male, he was thought to have those attributes more generally characteristic of females and epitomized in the descriptive ‘effeminate’.”133 The court determined that, because Title VII did not forbid discrimination on sexual preference, it did not prohibit the conduct alleged.134 The Ninth Circuit came to a similar conclusion in *DeSantis v. Pacific Telephone & Telegraph Co.*135

In *Nichols v. Azteca Restaurant Enterprises, Inc.*,136 however, the Ninth Circuit overruled *DeSantis* in light of *Price Waterhouse* and *Oncale*.137 The *Nichols* court found that “the systematic abuse directed at [the plaintiff],”138 which “reflected a belief that [he] did not act as a man should act . . . [,] was closely linked to gender”139 and thus in violation of Title VII.140 Likewise, in *Bibby v. Philadelphia Coca Cola Bottling Co.*,141 the Third Circuit recognized that a plaintiff could prove same-sex harassment “was discrimination because of sex” using a gender stereotype theory,142 citing to similar holdings in the First, Second, and Seventh Circuits.143

---

130  569 F.2d 325 (5th Cir. 1978).
131  Id. at 326.
132  Id. at 327.
133  Id.
134  Id. at 326–27 (citing Willingham v. Macon Tel. Publ’g Co., 507 F.2d 1084 (5th Cir. 1975) (en banc)).
135  608 F.2d 327, 329–30 (9th Cir. 1979).
136  256 F.3d 864 (9th Cir. 2001).
137  Id. at 874–75.
138  Id. at 874. Specifically, “[m]ale co-workers and a supervisor repeatedly referred to [the plaintiff] in Spanish and English as ‘she’ and ‘her.’ Male co-workers mocked [him] for walking and carrying his serving tray ‘like a woman,’ and taunted him in Spanish and English as, among other things, a ‘faggot’ and a ‘fucking female whore.’” Id. at 870.
139  Id. at 874.
140  Id. at 875.
141  260 F.3d 257 (3d Cir. 2001).
142  Id. at 262–64. However, the *Bibby* plaintiff failed to prove this theory because “he did not claim that he was harassed because he failed to comply with societal stereotypes of how men ought to appear or behave.” Id. at 264.
143  Id. at 263; see also *Simonton v. Runyon*, 232 F.3d 33, 35 (2d Cir. 2000) (upholding the district court’s dismissal of a same-sex harassment claim where plaintiff’s co-workers “repeatedly assaulted him with such comments as ‘go fuck yourself, fag,’ ‘suck my dick,’ and ‘so you like it up the ass?’”); *Higgins v. New Balance Athletic Shoe, Inc.*, 194 F.3d 252, 259 (1st Cir. 1999) (denying same-sex harassment claim on appeal because, despite evidence that co-workers mocked plaintiff’s supposedly effeminate characteristics, plaintiff presented that evidence to the district court only as an example of discrimination
Turning back to height, discrimination against a short man because he is short is, in effect, discrimination against him because his short stature manifests his failure to satisfy the male gender norm, i.e., because he is less manly than he ought to be. Likewise, discrimination against a tall woman because of her tall stature can be premised on an impermissible gender stereotype. Such discrimination would arguably fall within the contours of Title VII’s prohibition against gender-based discrimination.

2. The Approach

Although pervasive, height discrimination in the workplace is often latent, manifesting predominantly either in failures to hire or promote, or in disparate wages. As a result, the “traditional” approaches tend to break down: the disparate impact claim approach underachieves because, in many cases, employers will not have an express policy of height discrimination in hiring, promotion, or wages; and, the individual disparate treatment approach typically fails because discrete instances of height-based animus have proved hard to connect to a protected trait.144

There remains a third theory, however, that combines elements of both the individual disparate treatment and disparate impact theories and which responds nicely to the frailties of both theories in the context of height discrimination—namely, a systemic disparate treatment (or “pattern or practice”) theory. A systemic disparate treatment theory still requires a showing of discriminatory intent in order to succeed,145 but it permits an inference of such intent from a pattern of adverse outcomes or “bad stats.”146 Plaintiffs can prevail under such a theory if they can “prove more than the mere occurrence of isolated or ‘accidental’ or sporadic discriminatory acts”—that is, that discrimination was the employer’s “standard operating procedure.”147 Put into context, a height discrimination plaintiff could establish impermissible gender stereotyping predicated on height by showing a pattern of adverse employment actions that disproportionately affect a protected group.148 For example, while a short man might struggle to connect his
lower wages to invidious height discrimination when considered in isolation, he might have greater success by showing that all short men working for the employer receive lower wages than their similarly qualified, taller counterparts.

The EEOC typically brings systemic disparate treatment cases, but private plaintiffs may bring the functional equivalent of such cases in the form of Rule 23 class actions. Although these actions typically are not well-suited to address uniquely individualized forms of discrimination, such as disability discrimination under the ADA, plaintiffs have successfully brought systemic disparate treatment actions challenging sexual harassment and gender stereotyping under Title VII. In *Dukes v. Wal-Mart Stores, Inc.*, for example, the class alleged that Wal-Mart discriminated against female employees by delegating unfettered decision-making authority to lower-level supervisors and managers, which resulted in lower wages and fewer or slower promotions for female workers. To support their claim, the *Dukes* plaintiffs presented evidence, both anecdotal and expert, that Wal-Mart corporate culture placed a strong emphasis on building and maintaining a uniform culture and fostered an environment that perpetuated gender stereotyping. Similar class lawsuits have challenged the delegation of unfettered decision-making authority that promotes stereotypes and, in turn, limits opportunities for women or minorities.

Although a systemic theory of liability (either disparate impact or systemic disparate treatment) seems best suited for height-based claims, plaintiffs proceeding under a systemic disparate treatment theory are more likely to succeed if they can supplement their statistical evidence with direct evidence of

---


150 See *Sutton v. United Air Lines*, 527 U.S. 471, 480 (1999) (observing that “the determination of whether an individual is [disabled] must be made on a case by case basis”); *see also* Michael A. Stein & Michael E. Waterstone, *Disability, Disparate Impact, and Class Actions*, 56 Duke L.J. 861, 864 (2006) (noting that “the class action device, which historically played a central role in group-based discrimination theory (while often going hand in hand with robust disparate impact litigation), has been virtually nonexistent under the [ADA’s] employment provisions”).

151 *See generally* EEOC *v. Dial Corp.*, 156 F. Supp. 2d 926, 946 (N.D. Ill. 2001) (listing cases in which the courts recognized “the ability of plaintiffs, including the EEOC, to proceed on a pattern-or-practice theory in litigating claims of systemic employment discrimination, including sexual harassment”).


153 *Id.*

154 *Id.* at 141, 152–53.

155 *Id.* at 151, 153–54.

Courts are generally reluctant to infer discrimination from mere statistics without anecdotal evidence of discriminatory motive.\textsuperscript{158}

\section*{IV. Height Discrimination and the Americans with Disabilities Act}

Height-based discrimination claims may also prove viable under the Americans with Disabilities Act of 1990 (“ADA”).\textsuperscript{159} The ADA generally prohibits discrimination “because of” one’s disability.\textsuperscript{160} Just like Title VII plaintiffs, an ADA plaintiff must first make out a prima facie case under either a disparate treatment theory or a disparate impact theory.\textsuperscript{161} Under either theory, however, the first and typically fatal hurdle has been establishing one’s “disability.”\textsuperscript{162}

The ADA defines disability, in part, as “a physical or mental impairment that substantially limits one or more of the major life activities of [an] individual” or

\begin{itemize}
\item \textsuperscript{157} Cf. Miller, \textit{supra} note 9, at 264–65 (“[W]here available, statistics may demonstrate the discriminatory effect of a company’s hiring, assignment and promotion policies on minorities. . . . The low percentage of little people in the labor market makes it extremely difficult to obtain any meaningful statistical evidence about little people. Since no meaningful percentage of little people are concentrated in any particular labor area, researchers must examine a diverse group of industries to describe hiring and promotion practices affecting dwarfs.”).
\item \textsuperscript{158} See Tracy A. Baron, Comment, \textit{Keeping Women Out of the Executive Suite: The Courts’ Failure to Apply Title VII Scrutiny to Upper-Level Jobs}, 143 U. Pa. L. Rev. 267, 288 (1994) (“In systemic disparate treatment cases, courts often refuse to infer discrimination solely from a pattern of exclusion demonstrated by statistics, and instead insist on anecdotal evidence of discriminatory motive.”).
\item \textsuperscript{160} See 42 U.S.C. § 12112(a) (2006) (prohibiting disability discrimination in the workplace); \textit{id.} § 12132 (prohibiting discrimination in the enjoyment of public services, programs, or activities); \textit{id.} § 12182 (prohibiting same in the enjoyment of public accommodations).
\item \textsuperscript{161} See Raytheon Co. v. Hernandez, 540 U.S. 44, 49 n.3 (2003) (observing that the courts of appeals have applied the \textit{McDonnell Douglas} framework in ADA cases); \textit{id.} at 53 (recognizing that “[b]oth disparate-treatment and disparate-impact claims are cognizable under the ADA”).
\item \textsuperscript{162} See Sullivan, \textit{supra} note 146, at 942 & n.123 (discussing a 2003 study that found only 2 percent of ADA cases were won by employee plaintiffs and that barely half made it to consideration on their merits); \textit{see generally} ADA Amendments Act of 2008, Pub. L. No. 110-325 §§ 2(a), (b), 122 Stat. 3553, 3554 (2008) [hereinafter “ADAAA”] (expressing dismay that Supreme Court decisions construed the ADA too narrowly thereby eliminating protection for many individuals whom Congress intended to protect).
“being regarded as having such an impairment.” 163 This section examines how short stature fits (or does not fit) within each of these definitions. The first part looks at height-based claims under an “actual impairment” theory. Although most courts have wholly rejected “actual impairment” claims premised on short stature, a fresh look at the regulatory scheme in light of the ADA Amendments Act of 2008 (“ADAAA”) suggests that such claims may have merit after all. The second part considers height-based claims under the “regarded as” prong. Although “regarded as” claims premised on height have largely failed, such claims may prove increasingly viable after the enactment of the ADAAA.

A. “Actual Impairment”

1. Statutory and Regulatory Framework

To be protected under the ADA’s “actual impairment” prong, a plaintiff must establish (1) that he or she suffers from a physical or mental impairment (2) that substantially limits a major life activity.164 The Supreme Court summarily observed in Sutton v. United Air Lines, Inc.,165 that “an employer is free [under the ADA] to decide that physical characteristics . . . that do not rise to the level of an impairment—such as a one’s height, build, or singing voice—are preferable to others.”166 But what actually constitutes a physical impairment? The EEOC has promulgated regulations interpreting the ADA’s statutory language, defining physical impairment as “[a]ny physiological disorder, or condition, cosmetic disfigurement, or anatomical loss affecting one or more of the following body systems: neurological, musculoskeletal, special sense organs, respiratory (including speech organs), cardiovascular, reproductive, digestive, genito-urinary, hemic and lymphatic, skin, and endocrine . . . .”167

The EEOC has further clarified that “‘impairment’ does not include physical characteristics such as eye color, hair color, left-handedness, or height, weight or muscle tone that are within ‘normal’ range and are not the result of a physiological disorder.”168 A plain reading of this language would suggest that the definition of “physical impairment” includes either (1) a normal deviation in height that is the product of a physiological disorder, or (2) an extreme deviation in height that may or may not be caused by a physiological disorder.169

---

163 42 U.S.C. §§ 12102(2)(A), (C) (2006). The ADA also defines disability in a third way, i.e., if a person has a record of impairment. Id. § 12102(2)(B). This definition of disability, however, adds little to the discussion of height under the ADA and is therefore omitted.
166 Id. at 490 (emphasis added).
169 See EEOC Interpretive Manual, supra note 91, § 902.2(c)(5) (“[N]ormal deviations in height, weight, or strength that are not the result of a physiological disorder
Many federal courts have not read the regulations this expansively, finding that short stature does not constitute an impairment and, thus, is not protected under the ADA. As this part demonstrates, however, these courts have too narrowly construed the statute and regulations, particularly in light of the ADAAA’s renewed commitment to expanding protections against disability discrimination. Extreme deviations in height may qualify as impairments and, in some cases, rise to the level of disability even in the absence of an underlying physiological disorder.

2. Extreme Short Stature Constitutes an Impairment

(a) “Just Plain Short”

As noted above, a plain reading of the EEOC’s interpretive guidance suggests that height can qualify as an impairment in one of two ways: (1) if a physiological disorder causes a normal deviation in height; or (2) if one suffers from an extreme deviation in height (i.e., outside the “normal range”), whether or not it is the result of a physiological disorder. This section focuses on the second possible reading because most cases of short stature are not the result of a physiological disorder.

Short stature has myriad causes. Some causes are environmental, such as trauma, radiation, and malnutrition. Some causes are medical, including growth hormone deficiency (e.g., hypopituitary dwarfism), congenital diseases (e.g., Turner syndrome), illness (e.g., chronic renal insufficiency), and skeletal dysplasias (e.g., achondroplasia and diastrophic dysplasia). However, medical causes account for only about 5 percent of short stature cases, which is to say that most short people suffer from no biological malfunction at all.

---

170 See supra notes 168–169 and accompanying text.
171 See Fox, supra note 24, at 1144.
172 See id. at 1144 & n.43; Mary Lee Vance & Nelly Maujas, Growth Hormone Therapy in Adults and Children, 341 NEW ENG. J. MED. 1206, 1211 (1999); BETTY M. ADELSON, DWARFISM: MEDICAL AND PSYCHOSOCIAL ASPECTS OF PROFOUND SHORT STATURE 28–31 (2005).
173 See Vance & Mauras, supra note 172, at 1211; Fox, supra note 24, at 1144.
174 See Vance & Mauras, supra note 172, at 1211–12; Fox, supra note 24, at 1144.
175 See Vance & Mauras, supra note 172, at 1213; ADELSON, supra note 172, at 17–18.
176 See Fox, supra note 24, at 1144 (citing Raymond L. Hintz, Disorders of Growth, in HARRISON’S PRINCIPLES OF INTERNAL MEDICINE 128 (Kurt J. Isselbacher et al. eds., 1994)).
177 See id. (citing Larry R. Churchill, Bias, Opportunity, and Justice in Growth Hormone Therapy, in GROWTH, STATURE, AND ADAPTATION: BEHAVIORAL, SOCIAL, AND COGNITIVE ASPECTS OF GROWTH DELAY 195, 195 (Brian Stabler & Louis E. Underwood eds., 1994); see also MARTEL & BILLER, supra note 2, at 1 (“More than half of very short individuals have no apparent endocrinological or biological abnormality.”)).
The remaining short people comprise a “heterogeneous group of otherwise apparently normal [people] who are at or below the 5th percentile for height” but who respond normally to growth hormone.178 This group includes those classified as having genetic short stature, normal-variant familial short stature (if they have short parents), constitutional delay of growth (if they experience a delay in skeletal maturation), or idiopathic short stature (in the absence of any other diagnosable cause).179 Courts have already recognized that short stature resulting from a variety of the aforementioned medical causes, notably achondroplasia and diastrophic dysplasia, qualifies as an impairment.180 Those who are just plain short, however, have had virtually no success bringing pure height-based claims under the “actual impairment” prong.

In Mehr v. Starwood Hotels & Resorts Worldwide, Inc.,181 for example, the 4’10” female plaintiff filed EEOC charges alleging, inter alia, discrimination for “being short.”182 She later recharacterized her action as an ADA claim, asserting short stature as an impairment.183 The Sixth Circuit denied her claim as meritless because it interpreted the regulations as excluding from the definition of “impairment” all “physical characteristics that are ‘not the result of a physiological disorder.’”184

178 Vance & Mauras, supra note 172, at 1212.
179 Id. Idiopathic short stature is, in effect, a diagnosis of exclusion because it merely rules out all other systemic, genetic, syndromic, organic, or psychosocial causes. See Mary M. Lee, Idiopathic Short Stature, 354 NEW ENG. J. MED. 2576, 2576 (2006).
181 72 F. App’x. 276 (6th Cir. 2003).
182 Id. at 286.
183 Id. at 287.
184 Id. (quoting Andrews v. Ohio, 104 F.3d 803, 808 (6th Cir. 1997)). Andrews involved claims that obesity and lack of cardiovascular endurance qualified as impairments. Andrews, 104 F.3d at 805–06. The Sixth Circuit rejected the claims, perfunctorily reading 29 C.F.R. § 1630.2(h) to exclude from the definition of impairment all physical characteristics not resulting from physiological disorders. See id. at 808. The court justified that “[t]o hold otherwise would . . . distort the ‘concept of an impairment [which] implies a characteristic that is not commonplace’ and would thereby ‘debase [the] high purpose [of] the statutory protections available to those truly handicapped.’” Id. at 810 (emphasis added).

The Andrews court, however, did not consider the “normal range” reading of the regulation discussed above, instead relying on its prior holding in Jasany v. United States Postal Service, 755 F.2d 1244, 1250 (6th Cir. 1985), which determined that “‘c’haracteristics such as average height or strength that render an individual incapable of performing particular jobs are not covered by the statute because they are not impairments’” under the Rehabilitation Act. Andrews, 104 F.3d at 810 (quoting Jasany,
Similarly, in *Gowins v. Greiner*, the plaintiff was a 6’5”, wheelchair-confined inmate who sued under the ADA and the Rehabilitation Act because he was not provided with a bed that suitably accommodated his height and disability. The plaintiff conceded that the inadequacy of the bed derived primarily from his height, which caused his feet to dangle over the edge of the bed. The district court rejected his ADA claim, declaring that “a person’s height is not ordinarily an ‘impairment’ covered as a disability by the Rehabilitation Act or the ADA” and finding that physical characteristics, such as height, that are not the result of a physiological disorder do not qualify as impairments.

Why have these claims failed? In part, it may be because many of them, including *Mehr* and *Gowins*, have been litigated by pro se plaintiffs. More likely, it is because the courts considering such claims have considered only one possible reading of the regulation. Despite the regulation’s plain language, and notwithstanding the EEOC’s recognition that “[a]t extremes . . . deviations [in height] may constitute impairments,” federal courts have not considered this theory in deciding height-based claims under the ADA.

(b) “Within ‘Normal Range’”

Even if federal courts had applied an “extreme deviation in height” gloss to the regulation, they would have been left without much guidance on just how extreme a deviation in height must be to qualify as an impairment—this in spite of the EEOC’s efforts to provide such guidance. To clarify what constitutes an extreme deviation in height, the EEOC Interpretive Manual offers two examples. The first example, which the EEOC suggests does not qualify under the ADA, involves a 4’10” woman denied a job as a factory worker because the employer

---

755 F.2d at 1249 (emphasis added). Nevertheless, Jasany adopted this standard out of whole cloth. See Jasany, 755 F.2d at 1249 & n.4. Thus, Mehr serves as a good example of how courts have perpetuated incomplete (and therefore improper) standards for considering height as a qualifying impairment; other courts have relied on Andrews to dismiss height-based disability claims. See infra note 188 and accompanying text.

186  Id. at *4–5, *31.
187  Id. at *31.
188  Id. at *32 (quoting Andrews v. Ohio, 104 F.3d 803, 808 (6th Cir. 1997)).
190  EEOC Interpretive Manual, supra note 91, § 902.2(c)(5).
191  See supra note 184.
192  A height of 4’10” puts a woman below the third percentile for height. See supra Table 4; see also Fox, supra note 24, at 1143 (noting that “[a] short-statured adult female may be as tall as 4 feet 11.1 inches . . .”).
thought her too small to do the job.\textsuperscript{193} Despite her below-average height, the EEOC says that “her small stature [is] not so extreme as to constitute an impairment . . . .”\textsuperscript{194} The second example, which the EEOC finds \textit{does} qualify under the ADA, involves a 4’5” man suffering from achondroplastic dwarfism.\textsuperscript{195} Unfortunately, neither of these examples actually clarifies what falls outside “normal range.”

\textit{(i) Why the EEOC’s Interpretive Examples Provide No Guidance}

The second of the EEOC’s examples is entirely inapposite to cases that do not involve medically-caused short stature because the plaintiff in that example clearly suffers from an underlying physiological disorder.\textsuperscript{196} Thus, the plaintiff in the example would qualify for protection even under the more limited reading of the regulation that many courts have applied to height-based cases. The first example, moreover, although superficially helpful, is deficient for two reasons: (1) it speaks in terms of nominal height rather than relative height (i.e., in inches, not percentiles), which is most appropriate when discussing the “normal range”; and (2) it is based substantially on \textit{American Motors Corp. v. Labor & Industry Review Commission},\textsuperscript{197} which, for several reasons discussed below, cannot inform impairment determinations under the ADA.

In \textit{American Motors}, the Supreme Court of Wisconsin considered under the state’s Fair Employment Act the discrimination claims of a 4’10”-tall woman who had been denied a job as a factory worker allegedly because of her short stature.\textsuperscript{198} Under the Wisconsin law, the plaintiff had to first establish that she was “handicapped within the meaning of the Act.”\textsuperscript{199} The statute at that time, however, did not define “handicap.”\textsuperscript{200} Rather, the Wisconsin courts had defined “handicap” by its common usage to include “such diseases . . . which make achievement unusually difficult”\textsuperscript{201} or which “limit[ ] the capacity to work.”\textsuperscript{202}

Relying on this definition of “handicap,” the \textit{American Motors} court held that “a handicap within the meaning of the Act is a physical or mental condition that

\begin{itemize}
  \item \textsuperscript{193} See EEOC Interpretive Manual, \textit{supra} note 91, § 902.2(c)(5)(i).
  \item \textsuperscript{194} \textit{Id.} (citing Am. Motors Corp. v. Wisc. Labor & Indus. Review Comm’n, 350 N.W.2d 120 (Wis. 1984)).
  \item \textsuperscript{195} \textit{Id.} § 902.2(c)(5)(1) (citing Dexler v. Tisch, 660 F. Supp. 1418, 1425 (D. Conn. 1987)).
  \item \textsuperscript{196} See \textit{supra} note 175 and accompanying text.
  \item \textsuperscript{197} 350 N.W.2d 120 (Wis. 1984).
  \item \textsuperscript{198} \textit{Id.} at 121.
  \item \textsuperscript{199} \textit{Id.} at 122.
  \item \textsuperscript{200} \textit{Id.} at 122–23. Wisconsin has since rebubbled “handicap” as “impairment” and codified the definition. See WIS. STAT. ANN. § 111.32(8)(a) (West 2002).
  \item \textsuperscript{201} Am. Motors, 350 N.W.2d at 122–23 (quoting Chi., Milwaukee, St. Paul & Pac. R.R. Co. v. Wis. Dep’t Indus. Labor & Human Relations, 215 N.W.2d 443, 446 (Wis. 1974)).
  \item \textsuperscript{202} \textit{Id.} at 123 (discussing Dairy Equip. Co. v. Dep’t Indus. & Human Relations, 290 N.W.2d 330 (Wis. 1980)).
\end{itemize}
imposes limitations on a person’s ability to achieve and capacity to work beyond the normal limitations that might render a person unable to make certain achievements or perform every possible job.”

Although the court determined that “[a]ll persons have some mental or physical deviations from the norm,” the court rightly held that “such inherent limitations or deviations . . . do not automatically constitute handicaps.” The court concluded, however, that the plaintiff’s height “[did] not constitute such a significant deviation from the norm that it [made] achievement unusually difficult” because—although the plaintiff was below the norm for height and faced “some limitations on her general ability to achieve and work, a person with her stature [remained] capable of a wide range of achievements, including many that a taller and heavier person could not do.”

The EEOC’s continued reliance on American Motors for guidance is problematic for several reasons. First, the decision deals exclusively with Wisconsin state law. Moreover, the case was decided over a decade after the enactment of the Rehabilitation Act of 1973, which internally defined “disability,” as does the ADA, as “a physical or mental impairment that substantially limits one or more major life activities.” Thus, it is significant that Wisconsin’s Fair Employment Act defined handicap and impairment differently from then-existing federal law because it suggests that the Wisconsin law was meant to apply independently of the federal law and its definitions.

Second, the American Motors court never discretely considered whether the plaintiff’s height was itself an impairment. Rather, the court approached the inquiry such that, in effect, it conflated the factual determination of whether the plaintiff was impaired with the legal determination of whether she was “handicapped” (and thus eligible for protection under the Act). The Wisconsin

203 Id.
204 Id. at 124.
205 Id.
207 See Am. Motors, 350 N.W.2d at 122 n.3 (observing that the subsequent amendments to the Wisconsin law incorporated the definition of handicap contained in the federal Rehabilitation Act, but that “[b]ecause there [was] no indication that the legislature intended a retroactive application of this statute, the provisions of the Act in effect at the time Basile filed her amended complaint apply in [the] case”); cf. King v. City of Madison, Civ. A. No. 07-295, 2008 U.S. Dist. LEXIS 25793, at *2–3 (W.D. Wis. Mar. 28, 2008) (“Unfortunately for plaintiff, a state administrative law judge’s determination under [Wisconsin’s Fair Employment Act] has no bearing on a federal court’s determination of disability under federal law.”); Jane M. Nold, Hidden Handicaps: Protection of Alcoholics, Drug Addicts, and the Mentally Ill Against Employment Discrimination Under the Rehabilitation Act of 1973 and the Wisconsin Fair Employment Act, 1983 Wis. L. Rev. 725, 732–35 (describing the differences between the Wisconsin FEA and the Rehabilitation Act).
208 See Miller, supra note 9, at 247 (observing that the plaintiff’s lawyers “stressed her ability to perform the job’s tasks, down-playing any evidence of physical impairment,” a strategy that “excluded arguments which addressed the classification of stature as a handicap”).
courts now bifurcate this analysis to parallel the ADA analysis, considering first whether a person has an impairment and then determining whether that impairment makes “achievement unusually difficult or limits the capacity to work.”

Finally, American Motors relied on the Fair Employment Act’s definition of disability—limited in its application exclusively to the employment context—which rendered impairments protected only if they “impose[d] limitations on a person’s ability to achieve and capacity to work beyond the normal limitations that might render a person unable to make certain achievements or perform every possible job.” The ADA, however, considers whether an impairment “substantially limits one or more . . . major life activities,” which encompasses far more than just working. After all, the ADA extends well beyond the employment context. Thus, the determination by the American Motors court that the plaintiff was not disabled because she could still work would not presently preclude a finding that a 4’10”-tall woman is “disabled” under the ADA if her impairment substantially limited a life activity other than working. The ADAAA has since codified a non-exhaustive list of such life activities.

In sum, although the EEOC has tried to clarify what constitutes outside the “normal range” for height, its purported guidance provides little guidance at all. In the absence of such agency insight, the question remains how to determine when a person’s short stature is sufficiently extreme to qualify as an impairment under the federal statute.

(ii) What Should Be Considered “Within Normal Range”

To determine what is within normal range for height, one might look for guidance to similar regulations governing weight. However, the weight-based regulations prove largely unhelpful for a variety of reasons, most notably that few if any pure weight-based discrimination claims are brought under the ADA. Rather, overweight ADA plaintiffs typically allege discrimination on the basis of

---

209 See, e.g., Hutchinson Tech., Inc. v. Labor & Indus. Review Comm’n, 682 N.W.2d 343, 346–37 n.4 (Wis. 2004) (quoting WIS. STAT. ANN. § 111.32(8)).
210 Am. Motors, 350 N.W.2d at 124 (emphasis added).
213 The ADA also applies to public services, see 42 U.S.C. §§ 12131–12165 (2006), and public accommodations. See id. §§ 12181–12189 (2006).
215 But pure weight requirements have been challenged under Title VII for their disparate impact on certain groups, such as women. See generally EEOC Interpretive Manual, supra note 91, §§ 621.4–621.5 (discussing how weight requirements give rise to disparate impact and disparate treatment claims under Title VII).
their obesity. To qualify as an impairment, a person’s weight must actually constitute “severe” or “morbid” obesity, i.e., it must be 100 percent over that person’s medically ideal weight. This measurement, however, is relative to a nominal value (normal height for one’s ideal weight), without reference to the frequency with which that value occurs in the population. And what constitutes normal or abnormal weight depends on not just how heavy one is, but how heavy one is compared with how tall one is. To put it simply, there theoretically could be an unlimited number of obese people if everyone weighed twice as much as he or she should, while there could never be an unlimited number of short people because some will always be taller or shorter than others. Thus, because mere weight as a characteristic is not protected under the ADA, and because obesity is not truly examined for statistical “normalcy” within the population, the definition of obesity does little to inform the determination of “normal range” for height.

The most logical remaining option for determining “normal range” would be to look to statistical principles, namely standard deviation. Standard deviation, as a “descriptive statistic,” measures “the typical or expected variation of the

---


217 EEOC Interpretive Manual, supra note 91, § 902.2(c)(5)(ii) n.15 (“The term ‘obesity’ has been defined as ‘[t]he excessive accumulation of body fat. Except for heavily muscled persons, a body weight 20% over that in standard height-weight tables is arbitrarily considered obesity.’” (quoting THE MERCK MANUAL OF DIAGNOSIS AND THERAPY 981 (Robert Berkow ed., 16th ed. 1992)).

218 See CDC.gov, Overweight and Obesity: Defining Overweight and Obesity, http://www.cdc.gov/nccdphp/dnpa/obesity/defining.htm (last visited Sept. 1, 2009) (“For adults, overweight and obesity ranges are determined by using weight and height to calculate a number called the ‘body mass index’ (BMI).”).

219 This fact is made all the more evident by the dramatic increase in obesity in the United States. For an animated map of the increasing prevalence of obesity from 1985 to 2008, see CDC.gov, Obesity and Overweight: Trends by State 1985–2008, http://www.cdc.gov/obesity/data/trends.html#State (last visited Sept. 1, 2009).

220 See Michael H. Shapiro, The Technology of Perfection: Performance Enhancement and the Control of Attributes, 65 S. CAL. L. REV. 11, 40 n.87 (1991) (“Of course, any population with a variable trait will have upper and lower regions.”).

221 See infra app. fig.1.

222 Descriptive statistics—such as averages and medians—“describe certain features of the numbers on which they are based and suppress others.” DAVID W. BARNES & JOHN M. CONLEY, STATISTICAL EVIDENCE IN LITIGATION: METHODOLOGY, PROCEDURE, AND PRACTICE 126 (1986).
numbers in a group from their average.” Standard deviation is particularly useful when examining a “normal population” of numbers, like anthropometric data (height, weight, etc.), which resemble bell-shaped curves. Looking at the “range” of a normal population, i.e., from the lowest value to the highest value, 68 percent of all values will fall within one standard deviation of the mean: 34 percent will be within one standard deviation below the mean, and 34 percent will be within one standard deviation above the mean. Roughly 96 percent of all values will fall within two standard deviations from the mean, 48 percent below and 48 percent above. Thus, only 4 percent of the population will fall beyond two standard deviations from the mean—the smallest and largest 2 percent of values relative to the mean.

One can use standard deviation to test the degree to which an assumption, called a null hypothesis, is the result of chance. An outcome is not likely to be the result of chance if it falls more than two standard deviations from the mean, because there are only four chances in 100 that the outcome is consistent with the assumption. Thus, if the assumption is that a person’s height will be average, anyone whose height falls beyond two standard deviations from the average—that is, anyone among the shortest 2 percent or the tallest 2 percent—is not of average (normal) height. Generally, statisticians reject the null hypothesis when there is less than a 5 percent probability that the outcome is the product of chance. This means that any outcome falling beyond 1.96 standard deviations above or below the mean is excluded.

Applying these principles, the shortest 2.5 percent and the tallest 2.5 percent of the population statistically fall outside “normal range” for height. Based on national anthropometric data collected by the National Center for Health Statistics, men shorter than approximately 5’4” and women shorter than roughly 4’11” would fall outside normal range.

A 2.5 percentile benchmark ostensibly comports with other regulatory and scientific benchmarks for abnormal height. For example, the FDA recommends treating idiopathic short stature in children only in cases where the child’s growth rate is unlikely to produce an adult height within normal range, which the FDA

---

223 Id. at 129.
224 Id. at 140.
225 See infra app. tbls.1–4.
226 BARNES & CONLEY, supra note 222, at 140.
227 ZIMMER, supra note 149, at 242.
228 See id.
229 Id.; see also infra app. fig.1.
230 See ZIMMER, supra note 149, at 237.
231 Id. at 243.
232 Id.
233 Id.
234 This is yet another reason why the EEOC’s interpretive example, based on the facts of American Motors, proves unhelpful. In American Motors, the plaintiff was a 4’10” female. See supra notes 193–194 and accompanying text.
projects to be sixty-three inches for men and fifty-nine inches for women. \(^{235}\) Both the American Academy of Pediatrics and American Association of Clinical Endocrinologists define short stature in a similar manner. \(^{236}\) They state that short stature is “height that falls more than two standard deviations . . . below the national mean for age and sex.” \(^{237}\) In the context of disability benefits provided under the Social Security Act, persistent height below the third percentile that is also “related to an additional specific medically determinable impairment” qualifies as a compensable disability in children. \(^{238}\) Short adults, however, do not qualify for Social Security benefits, \(^{239}\) although the qualifying standards for “disability” under the Social Security Administration tend to be more rigorous than those for protection under the ADA. \(^{240}\)

Nevertheless, the 2.5 percentile is frustratingly subjective. Although some doctors consider height below even the fifth percentile to be outside normal range, \(^{241}\) many readily admit that any measure is “totally arbitrary.” \(^{242}\) For

\(^{235}\) FDA Talk Paper, T03-56, FDA Approves Humatrope for Short Stature (July 25, 2003), available at http://www.scienceblog.com/community/older/archives/M/1/fda0850.htm; see also Lee, supra note 179, at 2578 & Table 1.

\(^{236}\) See Fox, supra note 24, at 1143.

\(^{237}\) See id.

\(^{238}\) 20 C.F.R. § 404, Subpt. P., App.1, 100.02 (2008); see also Horn v. Callahan, No. 96-5257, 1997 U.S. App. LEXIS 28599, at *4–5 (10th Cir. Sept. 26, 1997) (finding that achondroplastic dwarfism—a hereditary condition that retards bone growth—cannot be the “additional specific medically determinable impairment” to which a claimant’s growth impairment must be “related” under § 100.02); Verret v. Comm’r of Soc. Sec., Civ. A. No. 99-3647, 2001 U.S. Dist. LEXIS 2217, at *16–18 (E.D. La. Feb. 21, 2001) (finding no qualifying growth impairment, despite persistence of height below third percentile, absent evidence that established its relation to another impairment).

\(^{239}\) The Social Security Administration determines eligibility for disability benefits on the basis of regulations that describe “various physical and mental illnesses and abnormalities, most of which are categorized by the body system they affect.” Sullivan v. Zebley, S.S.R. No. 91–7c (Cum. Ed. 1991), 1991 SSR LEXIS 7, at *12 (Aug. 1, 1991) (citing 20 C.F.R. pt. 404, subpt. P, App. I (pt. A)). In addition to those categorized body systems—“musculoskeletal, special senses and speech, respiratory, cardiovascular, digestive, genito-urinary, hemic and lymphatic, skin, and endocrine”—there are four additional groups of listings not categorized by body system, which include multiple body system impairments, neurological impairments, mental disorders, and malignant neoplastic diseases. See Sullivan v. Zebley, 493 U.S. 521, 530 n.6 (1990). For children, however, the regulations add a category for “growth impairment.” Id.

\(^{240}\) Cf. Cleveland v. Policy Mgmt. Sys. Corp., 526 U.S. 795, 797–98 (1999) (observing that “[t]he Social Security Disability Insurance (SSDI) program provides benefits to a person with a disability so severe that she is ‘unable to do [her] previous work’ and ‘cannot . . . engage in any other kind of substantial gainful work which exists in the national economy[,]’ whereas the ADA considers merely whether ‘“with reasonable accommodation” she could ‘perform the essential functions’ of her job” (citing 42 U.S.C. §§ 1382c(a)(3), 12111(8)).

\(^{241}\) See HALL, supra note 5, at 303; Sandberg & Colsman, supra note 12, at 18.
example, some researchers suggest that a boy who is just one standard deviation below the average height for his age remains at risk for “psychological difficulties.” Ultimately, the problems associated with short stature may not boil down to height at all but rather a person’s “ability to cope with the stresses of being short.”

3. “Substantially Limits a Major Life Activity”

The second hurdle to establishing a viable “actual impairment” claim involves showing how one’s impairment substantially limits a major life activity. Several courts have dismissed height-based claims, before the ADAAA, because the plaintiff failed to establish this element. For example, in *Reiterman v. Costco Wholesale Management # 238*, the 4’8” female plaintiff claimed that she was disabled because her short stature made it difficult to reach her cash register and that constant reaching caused her tendinitis. The court rejected her ADA claim, perfunctorily finding that her height was not a “disability” within the meaning of the ADA “because it [did] not ‘substantially limit’ her ability to engage in the major life activity of working.” Likewise, in *Mullet v. American Cargo, Inc.*, the district court determined that the plaintiff could not proceed under an “actual impairment” theory because he could not identify any major life activity that his height kept him from performing.

The ADAAA now codifies in the statute a non-exhaustive list of qualifying life activities, which includes “caring for oneself, performing manual tasks, seeing, hearing, eating, sleeping, walking, standing, lifting, bending, speaking, breathing, learning, reading, concentrating, thinking, communicating, and working.” The ADAAA further clarifies that a qualifying impairment need only substantially limit one such activity in order to render someone “disabled.” Concededly, being just plain short in most cases does not substantially limit one’s ability to engage in any

---

242 HALL, supra note 5, at 303; see also Sharon E. Oberfield, *Growth Hormone Use in Normal, Short Children—A Plea for Reason*, 340 NEW ENG. J. MED. 557, 557–59 (1999) (suggesting that there is no good definition of “short stature”); cf. ADELSON, supra note 172, at 3 (stating that for dwarfism “any cutoff must be somewhat arbitrary”).

243 MARTEL & BILLER, supra note 2, at 99.

244 HALL, supra note 5, at 303.


246 *Id.* at *3–4.*

247 *Id.* at *8.*


249 *Id.* at *8–9.*


of the aforementioned life activities, at least under the pre-ADAAA definition of “substantially limits” (i.e., “prevents or severely restricts”). To argue otherwise would be disingenuous. Nevertheless, there may very well be some whose height falls so far below the 2.5 percentile threshold (e.g., at the 1st percentile) that their short stature does limit a qualifying life activity; but that proportion of people is small. As such, only a very limited group of individuals could likely successfully pursue an “actual impairment” claim applying the pre-ADAAA definition of “substantially limits.”

Yet all may not be lost in the post-ADAAA world. First, the statutory list of qualifying major life activities is non-exhaustive and may be interpreted in the future to include activities that short stature does substantially limit. Second, Congress has directed that the definition of disability be construed broadly and that impairments be considered in their unmitigated state—that is, before they are corrected by medication, assistive technology, accommodations, or modifications. Third, in keeping with the spirit of the Findings and Purposes sections of the ADAAA, the EEOC may in the future retool its definition of “substantially limits” to be less restrictive than the pre-ADAAA standard.

Finally, it is important to understand that presently identifying short stature as an impairment may prove necessary to overcoming those pre-ADAAA cases that casually concluded that height can never qualify for protection under the ADA. This body of law could hinder height-based claims brought, not only under the “actual impairment” prong, but also the “regarded as” prong.

B. “Regarded As” Impaired

Whatever the merits of characterizing height as a disability, there can be no doubt that we perceive extreme short stature as a handicap. It may be fair to say

---

252 See Toyota Motor Mfg., Ky., Inc. v. Williams, 534 U.S. 184, 198 (2002) (noting that it was insufficient to “merely submit evidence of a medical diagnosis of an impairment”).

253 For example, while it is hard to conceive of an otherwise healthy 5’2” man being unable to care for himself, a 4’10” man or a 4’8” woman may find it significantly more difficult to do so.


255 Id. § 4(a) (to be codified at 42 U.S.C. § 12102(4)(E)(i)).

256 See Long, supra note 251, at 219–220.

257 See supra notes 181–191 and accompanying text.

258 See infra notes 259–273 and accompanying text.

259 See Oberfield, supra note 242, at 558 (“What, then, about the use of growth hormone for short but otherwise normal children whose parents or physicians believe them to be handicapped because of their shortness, or who believe so themselves?”); Colin Krivy, Men Calming Down About Measuring Up, THE GLOBE & MAIL, Mar. 8, 1995, at A22 (“But shortness, as much as I hate to admit it, is equated with disability. . . . The assumption was always that my diminutive size made me deficient in some way and therefore I couldn’t possibly be on a par with the lucky tall ones.”).
that discrimination against short people is more the result of stereotypes than any physical limitations imposed by their short stature. As noted earlier, we often engage in “gaze behavior”—that is, we subconsciously relegate to a lesser social status those who are shorter than we are because we perceive them as physically inferior. This perception is most evident in, and perhaps has been amplified by, the debate over the use of HGH to treat normal, short children.

Before the ADAAA, however, merely viewing an extremely short person as socially inferior would not have sufficed to render him or her protected. Pre-ADAAA, a person was covered by the “regarded as” prong of the ADA if (1) a covered entity treated a physical or mental impairment as though it substantially limited a major life activity, even though the impairment did not; (2) a covered impairment substantially limited a major life activity only as a result of the attitude of others toward such impairment; or (3) a covered entity treated a person as having an impairment that substantially limited a major life activity even though a person suffered from no such impairment. The first two variations required a plaintiff to establish a qualifying impairment, while all three variations required that an impairment, be it real or perceived, either substantially limit a major life activity or be perceived by the employer to substantially limit such activity. Thus, as alluded to above, a pre-ADAAA claim premised on height suffered from two major deficiencies: (1) qualifying short stature as an actual impairment, particularly given the arbitrariness of the 2.5 percentile threshold for impairment; and (2) establishing how one’s short stature substantially limited, or was perceived by the employer to substantially limit, a major life activity.

Under the more permissive ADAAA, however, a person is protected under the “regarded as” prong if an employer merely discriminates against him or her “because of an actual or perceived physical or mental impairment whether or not the impairment limits or is perceived to limit a major life activity.” As such, the focus under the ADAAA dramatically shifts from the severity of the employer’s

---

260 Cf. ADELSON, supra note 172, at 2–3 (discussing stereotypes related to dwarfism).
261 See supra notes 11–14 and accompanying text.
262 See supra notes 23–29 and accompanying text. Parents of otherwise healthy short children have pushed the medical community to treat their children with HGH because they fear the psychosocially disabling effects of short stature. See ELLIOTT, supra note 26, at 241; see also HALL, supra note 5, at 247 (noting that “many children who are severely short have nothing obviously wrong with them medically”); Vance & Mauras, supra note 172, at 1213 (noting that “[p]arental pressure to correct the perceived ‘deficiency’ of short stature has been responsible in part for the initiation of [growth hormone] treatment” among idiopathically short children).
264 See Long, supra note 251, at 223.
265 See supra notes 181–191, 241–244 and accompanying text.
266 See supra notes 252–253 and accompanying text.
misperception about an employee’s impairment to merely whether an employee’s actual or perceived impairment motivated an adverse employment action. However, the amended “regarded as” prong does not apply to impairments that are both “minor” and “transitory” (i.e., those having “an actual or expected duration of 6 months or less”). Moreover, covered employers “need not provide a reasonable accommodation or a reasonable modification to policies, practices, or procedures to an individual” who is disabled solely under the “regarded as” prong.

The ADAAA thus resolves the two major deficiencies suffered by “regarded as” claims premised on height. First, it renders unimportant whether a person’s height is at or below the 2.5 percentile, so long as the employer regards a person’s short stature as an impairment. For example, a person at the 2.6 percentile for height—who would not otherwise suffer from a qualifying impairment—could still bring a “regarded as” claim if the employer unfairly perceived him or her as incapable because of his or her short stature. Second, it does away with the need to establish a causal nexus between one’s short stature and substantial limits on a major life activity. Finally, although the ADAAA removes the employer’s obligation to provide reasonable accommodations from the quiver of remedies available to “regarded as” plaintiffs, height discrimination typically manifests in a failure to hire, a failure to promote, or disparate wages, none of which ostensibly requires accommodations to remedy.

See supra note 251, at 224.

ADAAA, Pub. L. No. 110-325 § 4(a), 122 Stat. 3553, 3555 (2008) (to be codified at 42 U.S.C. § 12102(3)(B)). What constitutes a “minor” impairment, however, is not defined by the ADAAA. See supra note 251, at 224. Nevertheless, “minor” and “transitory” are stated in the conjunctive, suggesting an impairment that is either minor or transitory would still qualify for protection under the “regarded as” prong. Compare Executive Office of the President, Statement of Administration Policy (June 24, 2008) (“The bill [H.R. 3195] does exclude impairments that are both transitory and minor; however, those that are one or the other would be covered. . . . The Administration believes that the bill should exclude from coverage impairments that are either transitory or minor.”), with ADAAA, Pub. L. No. 110-325 § 4(a), 122 Stat. 3553, 3555 (2008) (excluding from the protections of the “regarded as” prong “impairments that are transitory and minor”) (emphasis added). Thus, even though short stature may be considered a minor impairment, it arguably qualifies under the “regarded as” prong because it is not transitory.


See supra note 234 and accompanying text.
V. ADDITIONAL PROTECTIONS AGAINST HEIGHT DISCRIMINATION

A. State and Local Laws

1. Common Law Tort

Without the explicit protections of antidiscrimination laws, many victims of height discrimination in the workplace have had to pursue common law tort claims, namely intentional infliction of emotional distress (IIED). However, a plaintiff pursuing an IIED claim because of workplace height discrimination bears a heavy burden: he or she must show that the conduct giving rise to the claim was extreme and outrageous.\footnote{See infra notes 273–277 and accompanying text.}

In Bodnovich v. ABF Freight Systems, Inc.,\footnote{No. 93-B-36, 1994 Ohio App. LEXIS 5598 (Ohio Ct. App. Dec. 13, 1994).} for example, the plaintiff was terminated from his job as a manager and sued under Ohio law, alleging that his superiors subjected him to “cruel and derisive treatment . . . because of his short stature.”\footnote{Id. at *1–2.} The court rejected his claim for intentional infliction of emotional distress because it found that he had failed to adduce evidence of the requisite extreme and outrageous conduct.\footnote{Id. at *7; see also Schlumbrecht v. Fidelity Homestead Ass’n, Civ. A. No. 03-3001, 2006 U.S. Dist. LEXIS 14487, at *26–32 (E.D. La. Mar. 30, 2006) (finding that the plaintiff had failed to establish as required by Louisiana law).}

Likewise, in Micu v. Warren,\footnote{382 N.W.2d 823 (Mich. Ct. App. 1985).} the court rejected a claim of intentional infliction of emotional distress brought by a male firefighter applicant who sued the City of Warren for rejecting his application on the basis of his short stature.\footnote{Id. at 824–25. However, the Michigan Court of Appeals did find in favor of the plaintiff with respect to his state antidiscrimination law claim. See id. at 828; see also infra note 288 (discussing Micu in the context of Michigan’s antidiscrimination law).}

Similarly, the short female plaintiff in Edwards v. New Opportunities Inc.\footnote{932 P.2d 1261 (Wash. Ct. App. 1997).} survived a motion to dismiss her IIED claim because the court found that her employer’s behavior was extreme and outrageous.\footnote{Id. at 1262–63.} She alleged that “she was referred to as Dopey because of her short stature and accent and that a nine-foot
The cutout of Snow White was positioned outside of the building with her nickname attached..."282

In sum, because common law tort claims involving height discrimination are held to a far more rigorous standard than such claims would be held to under federal antidiscrimination laws, state tort laws do not provide much protection in run-of-the-mill cases of workplace height discrimination.

2. State and Local Antidiscrimination Laws

In the absence of a federal law explicitly prohibiting height discrimination, and in light of the inadequacies of state tort law, several state and local governments have endeavored to fill the gaps in protection by promulgating statutory protections against height discrimination. One state, Michigan, affirmatively proscribes discrimination on the basis of height,283 while another, Massachusetts, has actively considered protecting against height discrimination but has not yet passed such legislation.284 The District of Columbia has taken a more general approach, condemning employment discrimination on the basis of “personal appearance,” which includes “bodily... characteristics” such as height.285 And several municipal governments in California—including those in San Francisco and Santa Cruz—have directly taken on height discrimination in the workplace.286

Nevertheless, even in jurisdictions where laws explicitly prohibit height discrimination, few pure height discrimination cases have been brought,287 and

282 Id. at *28 n.10.
287 For example, a discrimination investigator with the San Francisco Human Rights Commission could not find a single complaint alleging height discrimination since the city passed the law. See E-mail from Holy Old Man Bull (formerly known as Marcus de Maria Arana), Discrimination Investigator, San Francisco Human Rights Commission (Jan. 27, 2009, 19:03:47 EST) (on file with author). Likewise, the Assistant Director of the City of Santa Cruz Human Resources Department could not recall a single complaint or record filed under the ordinance alleging height discrimination. See E-mail from Joe McMullen,
even fewer have been successful.\textsuperscript{288} Part of the problem may be, as mentioned above, that height discrimination tends to be implicit, rarely manifesting in overtly discriminatory conduct. Thus, victims of height discrimination either may not realize they are being discriminated against or may not be able to amass sufficient evidence of discriminatory motive. Another part of the problem may be that some misapprehend what these laws actually protect against. For example, given that Michigan’s law was passed almost contemporaneously with the Supreme Court’s decision in \textit{Dothard v. Rawlinson},\textsuperscript{289}—which held that an Alabama statute that imposed minimum height and weight requirements on prison guards violated Title VII because it disqualified over 40 percent of female applicants but less than 1 percent of male applicants\textsuperscript{290}—some might mistake that law’s solicitude of height as just another (i.e., less onerous) way to allege a gender-based disparate impact claim.\textsuperscript{291} And yet another part of the problem might be that these jurisdictions have instituted protections against height discrimination merely to put themselves in the “vanguard of anti-discrimination,” without any genuine or practical concern for height-based discrimination in the workplace.\textsuperscript{292}

\textsuperscript{288} There appears to be only one case of record in which a plaintiff alleging a claim of height discrimination prevailed under District of Columbia law. See Coleman v. District of Columbia, 700 A.2d 232, 233 (D.C. 1997). In Michigan, which has prohibited height discrimination since 1976, most of the mere handful of height-based cases litigated have failed because of pleading deficiencies or the absence of evidence establishing discriminatory motive. See, e.g., Terry v. DaimlerChrysler Corp., No. 263339, 2005 Mich. App. LEXIS 3202, at *4–5 (Mich. Ct. App. Dec. 20, 2005) (finding no direct evidence of discrimination where the plaintiff testified that she thought she was too short for the job to which she was assigned while admitting that management personnel told her they thought she could do the job); Berry v. Way Bakeries, Inc., No. 248841, 2005 Mich. App. LEXIS 459, at *11–13 (Mich. Ct. App. Feb. 22, 2005) (finding that plaintiff was properly terminated, in spite of her short stature, because plaintiff’s height affected her ability to perform certain tasks and gave rise to safety concerns). However, plaintiffs have successfully brought more “traditional” claims premised on height. See, e.g., Roskamp v. Cigna Sec., Inc., Civ. A. No. 89-60122, 1989 U.S. Dist. LEXIS 17850, at *1–3 (W.D. Mich. Sept. 28, 1989) (denying employer’s motion for summary judgment where pseudo-achondroplastic dwarf alleged that he was not hired because of his height); Micu v. City of Warren, 382 N.W.2d 823, 827–28 (Mich. Ct. App. 1985) (finding fire department could not discriminate on the basis of height in hiring firefighters unless it established height as a bona fide occupational qualification).


\textsuperscript{290} \textit{Dothard}, 433 U.S. at 329–31.

\textsuperscript{291} Cf. Kristen, \textit{supra} note 7, at 101 (“Michigan’s Ombudsman commented that one reason height . . . [was] added to the law was that ‘certain height . . . characteristics tend to be linked to certain ethnic groups or to women.’”).

\textsuperscript{292} Cf. E-mail from Joe McMullen, Assistant Dir., City of Santa Cruz Human Res. Dep’t. (Jan. 13, 2009, 20:31:32 EST) (on file with author) ("The City Manager recalls that
B. Why a Federal Prohibition on Height Discrimination Is Unnecessary

As discussed above, gender- and disability-based theories under existing federal law provide protection against many cases of height discrimination—but neither theory is perfect. The gender-based theory would substantially protect short men and tall women, leaving short women, tall men, and all people of average height with no protection. A disability-based theory would only protect people whose height falls outside the “normal range” and those whose short stature an employer perceives as an impairment, but would not protect the vast majority of people.

Nevertheless, these two theories do capture the two most significant prejudices underlying height discrimination. The question remains, what should be done, if anything, about height-based employment decisions involving everyone else? A flat federal prohibition on height-based employment decisions first comes to mind.

1. Such a Prohibition Would Be Overinclusive

Although height is an immutable trait—and despite the fact that antidiscrimination laws are premised, at least in part, on the notion that arbitrary decisions based on immutable traits are unfair and immoral—293—not all height-based employment decisions are motivated by prejudice. Thus, a flat federal prohibition on height-based employment decisions, while being the easiest to administer, would be overinclusive. After all, an employer is generally free to prefer certain traits to others so long as that preference does not rise to the level of impermissible prejudice.294

Still, some federal antidiscrimination laws, such as Title VII, overinclude by prohibiting trait-based employment decisions that may not actually be motivated by prejudice.295 That is, Title VII prohibits the consideration of covered traits in all their variations and renders everyone a member of the protected class: its protections extend not only to minorities who are most likely to be victims of the City Council wanted to be in the vanguard of anti-discrimination, but that there was not an event or anything that occurred that they were responding to [in passing the law].”); Kubilis, supra note 284, at 225 (quoting the Michigan’s Ombudsman as saying that “[t]here wasn’t even much debate about [passing the Elliot-Larsen Civil Rights Act]” and that the law passed with ease).

293 See ZIMMER, supra note 149, at 31.
294 See id.; see also Sutton v. United Airlines, Inc., 527 U.S. 471, 490 (1999) (“By its terms, the ADA allows employers to prefer some physical attributes over others and to establish physical criteria.”).
295 Cf. City of Boerne v. Flores, 521 U.S. 507, 530 (1997) (“The appropriateness of remedial measures must be considered in light of the evil presented. Strong measures appropriate to address one harm may be an unwarranted response to another, lesser one.” (citing South Carolina v. Katzenbach, 383 U.S. 301, 308, 334 (1966))).
prejudice, but to everyone. Thus, Congress has implicitly determined that the price of overinclusion is justified because the distinction between invidious prejudice and benign preference as to Title VII’s covered traits may be too subtle to be trusted.

Conversely, some federal antidiscrimination laws, like the Age Discrimination in Employment Act of 1967 (ADEA)\(^\text{296}\) and the ADA, are carefully circumscribed to prohibit consideration of specified traits only in limited cases, protecting only certain subgroups of people sharing a common trait. For example, although everyone has an “age,” the ADEA only prohibits an employer from considering the age of those who are at least 40 years old because, according to Congress, such older employees are most likely to be victims of prejudice.\(^\text{297}\) As such, younger employees are not protected even though they share the common trait of “age” with protected employees. Similarly, although millions of people may be classified as “impaired,” protections under the ADA extend only to employees whose impairments rise to the level of a disability; those who merely have impairments that do not qualify as disabilities are not protected, even though they share the common trait of “impairment” with those covered by the statute.\(^\text{298}\) Thus, Congress has implicitly determined that the price of overinclusion as to these traits is not justified because the prejudice-preference distinction can be clearly drawn with respect to these traits.

2. Overinclusion Would Not Be Justified

So how does one determine whether the price of overinclusion is justified? A useful shortcut may be the Equal Protection Clause itself. Classifications based on certain traits receive heightened scrutiny under the clause.\(^\text{299}\) These include all of those traits covered by Title VII: race, religion, national origin, color, and gender.\(^\text{300}\) Under heightened scrutiny, classifications based on these traits are presumptively invalid.\(^\text{301}\) However, classifications based on traits such as age and


\(\text{297}\) See id. § 621(a)(1) (“The Congress hereby finds and declares that . . . in the face of rising productivity and affluence, older workers find themselves disadvantaged in their efforts to retain employment, and especially to regain employment when displaced from jobs.”).

\(\text{298}\) See supra Part IV.A.


\(\text{300}\) See generally id. at 541–42, 671(describing strict scrutiny).

\(\text{301}\) See id. at 541–42 (observing that the proponent of the law, i.e., the government, bears the burden of proving the validity of the law under intermediate and strict scrutiny review); see also City of Cleburne v. Cleburne Living Ctr., 473 U.S. 432, 440 (1985) (“The general rule gives way, however, when a statute classifies by race, alienage, or national origin. These factors are so seldom relevant to the achievement of any legitimate state interest that laws grounded in such considerations are deemed to reflect prejudice and antipathy—a view that those in the burdened class are not as worthy or deserving as others.
disability receive mere rational basis review and are presumptively valid; they need only be rational to survive. Put simply, the level of overinclusiveness tolerated of a federal antidiscrimination law ostensibly coincides with the review given to the traits covered by that law under the Equal Protection Clause.

This brings us full circle. Whether the costs of a flat prohibition on height-based employment decisions would be justified, despite its overinclusiveness, depends on the level of scrutiny a court would apply to height under the Equal Protection Clause. Based on precedential and pragmatic considerations, however, height need not receive heightened review, and, thus, any law prohibiting height discrimination could not broadly prohibit all height-based considerations in employment.

When deciding whether to apply heightened scrutiny to a given classification, the Supreme Court has generally favored (1) immutable characteristics, (2) groups traditionally unable to protect themselves through the political process, and (3) groups with a history of being discriminated against. Although height is surely immutable, as an abstract trait it does not constitute a “discrete and insular minorit[y],” and thus generally fails to satisfy the second and third factors. From a practical standpoint, moreover, the Supreme Court has not seen fit within the past 30 years to broaden the categories of classifications receiving heightened scrutiny. The Court’s decision in *City of Cleburne, Texas v. Cleburne Living Center, Inc.*, succinctly captures the essence of the Court’s reluctance. When faced with the plaintiffs’ request to treat mental retardation as a quasi-suspect classification entitled to heightened scrutiny, the Court declined, observing that

For these reasons, and because such discrimination is unlikely to be soon rectified by legislative means, these laws are subjected to strict scrutiny, and will be sustained only if they are suitably tailored to serve a compelling state interest.”).


303 See *Chemerinsky, supra* note 299, at 540 (noting that, under rational basis review, the opponent of the law has the burden of proving its invalidity).

304 See id. at 672–73.

305 United States v. Carolene Prods. Co., 304 U.S. 144, 152 n.4 (1938) (“[W]hether prejudice against discrete and insular minorities may be a special condition, which tends seriously to curtail the operation of those political processes ordinarily to be relied upon to protect minorities, and which may call for a correspondingly more searching judicial inquiry.”).

306 However, it may very well be that short people lack political power commensurate with their numbers among the general population. See *supra* notes 15–19 and accompanying text (observing the paucity of short presidents); cf. *Chemerinsky, supra* note 299, at 672 (“Women, for example, are more than half the population, but traditionally they have been severely underrepresented in political offices.”).

307 See *Chemerinsky, supra* note 299, at 672.

[If the large and amorphous class of the mentally retarded were deemed quasi-suspect . . . it would be difficult to find a principled way to distinguish a variety of other groups who perhaps have immutable disabilities setting them off from others, who cannot themselves mandate the desired legislative responses, and who can claim some degree of prejudice from at least part of the public at large. One need mention in this respect only the aging, the disabled, the mentally ill, and the infirm. We are reluctant to set out on that course, and we decline to do so.

Doubtless, there have been and there will continue to be instances of discrimination against the retarded that are in fact invidious, and that are properly subject to judicial correction under constitutional norms. But the appropriate method of reaching such instances is not to create a new quasi-suspect classification and subject all governmental action based on that classification to more searching evaluation. Rather, we should look to the likelihood that governmental action premised on a particular classification is valid as a general matter, not merely to the specifics of the case before us.309

Put simply, height-based classifications would and should only receive rational basis review.310 To be sure, there may be circumstances where the government invidiously classifies on the basis of height (as it has with disability); but, such classifications are likely to fail even under the least rigorous rational basis review.311 Because height-based classifications would receive rational basis review under the Equal Protection Clause, any federal law prohibiting height-based employment decisions would have to be narrowly drawn to address only those employment decisions motivated by a height-based prejudice. Such a narrowly drawn prohibition would necessarily cover short people of all genders, and tall women. However, these groups are already protected, albeit somewhat crudely,

309 Id. at 445–46 (emphasis added). The Court went on to note that “mental retardation is a characteristic that the government may legitimately take into account in a wide range of decisions, and . . . we will not presume that any given legislative action, even one that disadvantages retarded individuals, is rooted in considerations that the Constitution will not tolerate.” Id. at 446. Ultimately, the Court invalidated the classifications as lacking a rational relation to a legitimate government purpose. See id. at 450.


311 See, e.g., City of Cleburne, 473 U.S. 432, 447–50; Romer v. Evans, 517 U.S. 620, 632 (1996) (finding unconstitutional Colorado’s Amendment 2—which effectively prohibited the passing of antidiscrimination laws protecting homosexuals—because “its sheer breadth is so discontinuous with the reasons offered for it that the amendment seems inexplicable by anything but animus toward the class that it affects; it lacks a rational relationship to legitimate state interests”).
under existing federal laws: short men under the ADA and Title VII; short women under the ADA; and tall women under Title VII. As such, any new federal law prohibiting height-based employment decisions would be redundant. Moreover, passing such a sweeping law would be practically and politically problematic, given increasing aversion to the expansion of antidiscrimination laws.

C. A Proposal

Although a federal bar on height-based employment decisions would be gratuitous, modest changes to the regulations and interpretive guidance covering Title VII and the ADA would help to clarify the scope and contours of those laws’ protections with respect to height. A change in the regulations, moreover, is preferable to a change in the statutes themselves because the federal rulemaking process is less political and more flexible than the legislative process.

For example, 29 C.F.R. pt. 1630, App. § 1630.2(h) states that “‘impairment’ does not include physical characteristics such as . . . height, weight or muscle tone that are within ‘normal’ range and are not the result of a physiological disorder.” As noted above, however, most courts have read this to mean that height cannot constitute an impairment on its own if not caused by a physiological disorder. Although the EEOC has not adopted such a narrow view in its Interpretive Manual, the regulation itself could be rephrased to make the disjunctive nature of the exclusion more evident. Thus, 29 C.F.R. pt. 1630, App. § 1630.2(h) could be amended to read that “‘impairment’ does not include physical characteristics such as . . . height . . . that are either within ‘normal’ range or are not the result of a physiological disorder.” Such a simple change to the regulation would counteract the bulk of pre-ADAAA cases that wrongly read the regulations to prohibit claims premised on height, and would permit height-based claims to proceed beyond motions to dismiss, where so many discrimination plaintiffs lose now that the Supreme Court has made federal pleading requirements more rigorous.
In addition, state and local governments should consider extending protections against height-based prejudice in the workplace, even if such protections have not been widely invoked by plaintiffs in those jurisdictions that have enacted such protections. \(^{318}\) Jurisdictions that decide to pass such laws should consider explicitly but conservatively expanding protections beyond those offered by the federal laws. For example, a state law typifying an extreme deviation in height as an impairment might do well to broaden the definition to height at or below the third percentile, ensuring that victims at the margin of the clinical definition of profound short stature are not left unprotected.

Of course, jurisdictions could also adopt sweeping protections against height-based employment decisions, such as those enacted in Michigan and the District of Columbia, if they determine that such protections are worth their commensurate costs. After all, the availability of state and local prohibitions may actually decrease litigation by increasing the frequency of mediation and negotiation. \(^{319}\) Moreover, having access to multiple court systems would help to more evenly distribute the burden of resolving height-based discrimination claims (if such claims become fashionable). Ultimately, a victim of height discrimination should have at his or her disposal a full panoply of remedies, both federal and other, when he or she decides to challenge heightism that has adversely affected him or her at work.

VI. CONCLUSION

Heightism does exist, and its effects in the employment context are undeniable, even if its causes are less obvious. Many victims of height discrimination have tried in earnest to challenge the adverse treatment they faced because of their stature, and many have failed. Nevertheless, discrimination claims premised on height remain viable under both Title VII and the ADA.

---

318 See supra notes 287–292 and accompanying text.

319 Cf. Kubilis, supra note 284, at 232–33 (describing the salutary effects caused by the Michigan, Santa Cruz, and San Francisco laws).
Discrimination against short men and tall women because of their height may constitute discrimination on the basis of a gender stereotype; employment decisions motivated by such stereotypes are unlawful. Discrimination against the profoundly short (i.e., those at or below the 2.5 percentile) because of their stature is likewise impermissible. Recent changes to the ADA, coupled with a fresh look at the regulatory scheme, suggests pre-ADAAA cases were wrong in summarily concluding that height-based discrimination is not actionable under the ADA.

In the end, federal law should prohibit height-based employment decisions motivated by prejudice, but a comprehensive law would be unfeasible. The fact is that existing federal law already prohibits the vast majority of cases involving such prejudice, and the costs of expanding federal law to ban all height-based employment decisions outweigh the benefits of such expansion. Nevertheless, modest amendments to the federal regulations—as well as the passing of protections against height discrimination by state and local governments—would go a long way in preventing this form of invidious prejudice.

"Short People are just the same
As you and I
(A Fool Such As I)
All men are brothers
Until the day they die
(It’s A Wonderful World)."320

---

320 Newman, supra note 1.
## Figure 1 – Standard Deviation

![Graph showing standard deviation distribution](image)

### Table 1 – Mean Standing Height in Inches for All Males Age 20 and Older: United States, 1999–2002

<table>
<thead>
<tr>
<th>Age (in years)</th>
<th>Mean</th>
<th>5th</th>
<th>10th</th>
<th>15th</th>
<th>25th</th>
<th>50th</th>
<th>75th</th>
<th>85th</th>
<th>90th</th>
<th>95th</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 and older</td>
<td>69.3</td>
<td>64.1</td>
<td>65.4</td>
<td>66.2</td>
<td>67.3</td>
<td>69.3</td>
<td>71.3</td>
<td>72.3</td>
<td>73.1</td>
<td>74.2</td>
</tr>
<tr>
<td>20-29</td>
<td>69.6</td>
<td>64.1</td>
<td>65.7</td>
<td>66.5</td>
<td>67.5</td>
<td>69.6</td>
<td>71.4</td>
<td>72.6</td>
<td>73.5</td>
<td>74.6</td>
</tr>
<tr>
<td>30-39</td>
<td>69.5</td>
<td>64.2</td>
<td>65.7</td>
<td>66.3</td>
<td>67.4</td>
<td>69.3</td>
<td>71.6</td>
<td>72.6</td>
<td>73.3</td>
<td>74.4</td>
</tr>
<tr>
<td>40-49</td>
<td>69.7</td>
<td>65.1</td>
<td>66.1</td>
<td>66.7</td>
<td>67.7</td>
<td>69.6</td>
<td>71.7</td>
<td>72.6</td>
<td>73.2</td>
<td>74.4</td>
</tr>
<tr>
<td>50-59</td>
<td>69.2</td>
<td>64.0</td>
<td>65.4</td>
<td>66.2</td>
<td>67.2</td>
<td>69.3</td>
<td>71.4</td>
<td>72.3</td>
<td>73.0</td>
<td>73.9</td>
</tr>
<tr>
<td>60-69</td>
<td>68.8</td>
<td>64.4</td>
<td>65.1</td>
<td>65.8</td>
<td>67.0</td>
<td>68.9</td>
<td>70.7</td>
<td>71.7</td>
<td>72.3</td>
<td>73.1</td>
</tr>
<tr>
<td>70-79</td>
<td>68.0</td>
<td>63.0</td>
<td>64.3</td>
<td>64.9</td>
<td>65.9</td>
<td>68.1</td>
<td>70.1</td>
<td>71.0</td>
<td>71.5</td>
<td>72.5</td>
</tr>
<tr>
<td>80 and older</td>
<td>67.2</td>
<td>62.8</td>
<td>63.7</td>
<td>64.5</td>
<td>65.4</td>
<td>67.1</td>
<td>69.0</td>
<td>70.0</td>
<td>70.6</td>
<td>71.7</td>
</tr>
</tbody>
</table>

---

Table 2 – Mean Standing Height in Inches for All Females Age 20 and older: United States, 1999–2002

<table>
<thead>
<tr>
<th>Age (in years)</th>
<th>Mean</th>
<th>5th</th>
<th>10th</th>
<th>15th</th>
<th>25th</th>
<th>50th</th>
<th>75th</th>
<th>85th</th>
<th>90th</th>
<th>95th</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 and older</td>
<td>63.8</td>
<td>59.4</td>
<td>60.3</td>
<td>61.0</td>
<td>62.0</td>
<td>63.8</td>
<td>65.6</td>
<td>66.5</td>
<td>67.3</td>
<td>68.2</td>
</tr>
<tr>
<td>20-29</td>
<td>64.1</td>
<td>59.9</td>
<td>60.5</td>
<td>61.4</td>
<td>62.3</td>
<td>64.1</td>
<td>65.8</td>
<td>67.0</td>
<td>67.5</td>
<td>68.4</td>
</tr>
<tr>
<td>30-39</td>
<td>64.2</td>
<td>59.9</td>
<td>60.8</td>
<td>61.5</td>
<td>62.4</td>
<td>64.2</td>
<td>65.9</td>
<td>66.9</td>
<td>67.5</td>
<td>68.6</td>
</tr>
<tr>
<td>40-49</td>
<td>64.3</td>
<td>59.9</td>
<td>60.8</td>
<td>61.5</td>
<td>62.5</td>
<td>64.3</td>
<td>66.1</td>
<td>67.0</td>
<td>67.6</td>
<td>68.7</td>
</tr>
<tr>
<td>50-59</td>
<td>63.9</td>
<td>59.7</td>
<td>60.6</td>
<td>61.3</td>
<td>62.1</td>
<td>63.9</td>
<td>65.5</td>
<td>66.4</td>
<td>67.2</td>
<td>68.1</td>
</tr>
<tr>
<td>60-69</td>
<td>63.2</td>
<td>59.0</td>
<td>59.9</td>
<td>60.7</td>
<td>61.6</td>
<td>63.2</td>
<td>64.9</td>
<td>65.8</td>
<td>66.5</td>
<td>67.1</td>
</tr>
<tr>
<td>70-79</td>
<td>62.6</td>
<td>58.4</td>
<td>59.1</td>
<td>59.9</td>
<td>61.0</td>
<td>62.6</td>
<td>64.1</td>
<td>65.3</td>
<td>65.8</td>
<td>66.6</td>
</tr>
<tr>
<td>80 and older</td>
<td>61.3</td>
<td>57.4</td>
<td>58.1</td>
<td>58.9</td>
<td>59.8</td>
<td>61.4</td>
<td>62.8</td>
<td>63.7</td>
<td>64.4</td>
<td>65.6</td>
</tr>
</tbody>
</table>

Table 3 – Stature-for-Age Percentiles for Boys Age 20 (Derived from CDC Growth Charts)

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Height (in inches)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd</td>
<td>64.3</td>
</tr>
<tr>
<td>5th</td>
<td>65</td>
</tr>
<tr>
<td>10th</td>
<td>66</td>
</tr>
<tr>
<td>25th</td>
<td>67.7</td>
</tr>
<tr>
<td>50th</td>
<td>69.6</td>
</tr>
<tr>
<td>75th</td>
<td>71.5</td>
</tr>
<tr>
<td>90th</td>
<td>73.2</td>
</tr>
<tr>
<td>95th</td>
<td>74.2</td>
</tr>
<tr>
<td>97th</td>
<td>74.9</td>
</tr>
</tbody>
</table>

Table 4 – Stature-for-Age Percentiles for Girls Age 20 (Derived from CDC Growth Charts)

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Height (in inches)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd</td>
<td>59.5</td>
</tr>
<tr>
<td>5th</td>
<td>60.1</td>
</tr>
<tr>
<td>10th</td>
<td>61.0</td>
</tr>
<tr>
<td>25th</td>
<td>62.6</td>
</tr>
<tr>
<td>50th</td>
<td>64.3</td>
</tr>
<tr>
<td>75th</td>
<td>66.0</td>
</tr>
<tr>
<td>90th</td>
<td>67.6</td>
</tr>
<tr>
<td>95th</td>
<td>68.5</td>
</tr>
<tr>
<td>97th</td>
<td>69.1</td>
</tr>
</tbody>
</table>

---

322 Id.
324 Id.
DIVIDED LOYALTIES: HOW THE *METLIFE V. GLENN* STANDARD DISCOUNTS ERISA FIDUCIARIES’ CONFLICTS OF INTEREST

Beverly Cohen*

TABLE OF CONTENTS

I. INTRODUCTION ................................................................................................................................. 956

II. THE HISTORY OF THE STANDARD OF REVIEW APPLIED TO ERISA PLAN BENEFIT DENIALS ........................................................................................................................................ 957

   A. The Fiduciary Standards Under ERISA .......................................................................................... 957
   B. The Firestone Standard of Review ................................................................................................. 959

III. COURTS’ CONFUSED APPLICATION OF THE FIRESTONE STANDARD TO CONFLICTS OF INTEREST .......................................................................................................................... 960

   A. Courts’ Disagreement on the Existence of a Conflict ............................................................... 960
   B. Courts’ Disagreement on How an Identified Conflict Alters the Standard of Review .......... 966

IV. *METROPOLITAN LIFE INSURANCE CO. V. GLENN*: THE SUPREME COURT’S RECENT PRONOUNCEMENT OF A STANDARD OF REVIEW FOR ERISA PLAN BENEFIT DENIALS BY A CONFLICTED FIDUCIARY ......................................................................................................................... 971

   A. The District Court’s Decision ....................................................................................................... 972
   B. The Sixth Circuit Court of Appeals’ Reversal .............................................................................. 972
   C. The Supreme Court’s Affirmation ................................................................................................. 973

V. HOW WEIGHING THE CONFLICT AS A FACTOR AFFECTS THE REVIEW: AN ANALYSIS OF THE POST-FIRESTONE CASES .................................................................................................................. 976

   A. The Conflict Had Little Impact in Post-Firestone Cases on Determining Whether Fiduciaries Had Abused Their Discretion ............................................................................................................. 977
   B. Post-Firestone Cases Finding an Abuse of Discretion Relied Upon Substantive Inconsistencies Between the Fiduciary’s Decision and the Record ................................................................................. 978
   C. The Fiduciary’s Conflict of Interest Does Not Act as a Tiebreaker in Close Cases .................. 982
   D. Weighing the Conflict as a Factor Is Indistinguishable from De Novo Review ......................... 986

VI. CONCLUSION .................................................................................................................................. 991

* © 2009 Beverly Cohen, Professor of Law, Albany Law School. Douglass College of Rutgers University (B.A., 1968); Alfred State College (A.A.S., 1979); Albany Law School (J.D., 1987). The author wishes to thank Theresa Colbert and Evette Tejada for technical assistance, and Mohammad Ali Naquvi, Albany Law School Class of 2010, and the research staff at the Albany Law School Schaffer Law Library for research assistance.
I. INTRODUCTION

Since Firestone Tire & Rubber Co. v. Bruch\(^1\) was decided two decades ago, federal courts have been giving employer health and disability benefit plans\(^2\)—regulated under the Employee Retirement Income Security Act of 1974 (“ERISA”)\(^3\)—the enormous advantage of deferential review whenever plan members and beneficiaries have challenged benefit denials by plan fiduciaries.\(^4\) As a result, ERISA plan members and beneficiaries have faced a substantial, often insurmountable hurdle. To reverse a denial of benefits, a claimant needed to prove that the fiduciary’s decision was arbitrary and capricious, or unreasonable.\(^5\)

Making for an even tougher challenge for plan members and beneficiaries, often the fiduciary that had denied the claim had operated under a conflict of interest.\(^6\) A claim denied was a claim the fiduciary employer or insurance company providing coverage did not have to pay, so the fiduciary was arguably incentivized to deny claims due to financial self-interest.\(^7\) Although courts have considered whether fiduciaries have operated under a conflict of interest as a factor in determining whether fiduciaries have abused their discretion, often the presence of a conflict has not influenced the outcome.\(^8\)

In the recently decided Metropolitan Life Insurance Co. v. Glenn,\(^9\) the Supreme Court had an opportunity to revisit the deferential standard of review applied to claim denials by conflicted ERISA fiduciaries, and it opted to reaffirm the Firestone standard.\(^10\) As a result, substantial obstacles remain for ERISA plan members and beneficiaries challenging benefit denials. For ERISA health and disability plans in particular, this is a harsh result because an unfairly denied claim may leave a member without benefits for an expensive and urgently needed medical procedure, or without financial support after suffering a catastrophic disability.

This Article examines the historical development of the standard of review applied to ERISA plan benefit denials by conflicted fiduciaries\(^11\) and takes a critical look at the Supreme Court’s recent reaffirmation of the Firestone standard in Metropolitan Life Insurance Co. v. Glenn.\(^12\) Ultimately, this Article concludes that the standard operates to discount a fiduciary’s conflict of interest as a

\(^{1}\) 489 U.S. 101 (1989).
\(^{2}\) Although the principles articulated in this Article apply to all types of ERISA benefit plans, this Article focuses on ERISA health and disability plans in particular.
\(^{4}\) See infra Part II.B.
\(^{5}\) See infra note 43 and accompanying text.
\(^{6}\) See infra Part III.A.
\(^{7}\) See infra Part III.A.
\(^{8}\) See infra Part V.A.
\(^{9}\) 128 S. Ct. 2343 (2008).
\(^{10}\) See infra Part IV.
\(^{11}\) See infra Part IV.
\(^{12}\) See infra Part II.
determining factor in the review. As a result, conflicted decision making in the context of ERISA health and disability plans will continue to be a key feature of these plans, raising a question of whether the undivided loyalty promised by ERISA to plan members and beneficiaries truly can be achieved.

II. THE HISTORY OF THE STANDARD OF REVIEW APPLIED TO ERISA PLAN BENEFIT DENIALS

A. The Fiduciary Standards Under ERISA

Congress passed ERISA in response to congressional investigations that revealed labor unions’ mismanagement and looting of union-sponsored pension plans. To insulate employee benefit plans from corruption, ERISA mandated that all such plans be held in trust for the exclusive benefit of the plan members and their beneficiaries.

All persons who exercise discretionary authority over ERISA plan assets are denominated ERISA fiduciaries. ERISA imposes two principal duties upon plan fiduciaries. First, an ERISA fiduciary must be prudent in administering plan assets. The duty of prudence requires an ERISA fiduciary to administer the plan "with the care, skill, prudence, and diligence under the circumstances then

---

13 See infra Part V.
14 See infra Parts V, VI.
15 See ERISA § 404(a)(1)(A), 29 U.S.C. § 1104(a)(1)(A) (2006) (stating that an ERISA fiduciary “shall discharge his duties with respect to a plan solely in the interest of the participants and beneficiaries”).
17 ERISA provides that “all assets of an employee benefit plan shall be held in trust.” ERISA § 403(a), 29 U.S.C. § 1103(a) (2006). This provision, however, does not apply to insurance policies of plan assets held by an insurer. See id. § 1103(b)(1)–(2) (“The requirements of subsection (a) [holding assets of plans in trust] shall not apply—(1) to any assets of a plan which consist of insurance contracts or policies issued by an insurance company qualified to do business in a State; (2) to any assets of such an insurance company or any assets of a plan which are held by such an insurance company.”); see also Langbein, supra note 16, at 1326 n.62 (noting that plans funded with insurance policies are excused from the trust requirement).
18 See ERISA § 3(21)(A), 29 U.S.C. § 1002(21)(A) (2006) (“Except as otherwise provided in subparagraph (B), a person is a fiduciary with respect to a plan to the extent (i) he exercises any discretionary authority or discretionary control respecting management of such plan or exercises any authority or control respecting management or disposition of its assets, (ii) he renders investment advice for a fee or other compensation, direct or indirect, with respect to any moneys or other property of such plan, or has any authority or responsibility to do so, or (iii) he has any discretionary authority or discretionary responsibility in the administration of such plan.”).
prevailing that a prudent man acting in a like capacity and familiar with such
matters would use in the conduct of an enterprise of a like character.”20 Second, is
the duty of loyalty, whereby the fiduciary “shall discharge his duties with respect
to a plan solely in the interest of the participants and beneficiaries.”21 This duty of
loyalty requires a fiduciary to administer plan funds “for the exclusive purpose of .
. . providing benefits to participants and their beneficiaries” and “defraying
reasonable expenses of administering the plan.”22 The Supreme Court has declared
that “a trustee bears an unwavering duty of complete loyalty to the beneficiary to
the trust, to the exclusion of the interests of all other parties,” which must be
enforced with “uncompromising rigidity.”23

Despite this duty of unwavering loyalty, an ERISA fiduciary may have
financial interests adverse to the beneficiaries.24 Employers, for example, can be
ERISA fiduciaries, even though they may take certain actions that disadvantage
plan members.25 Employers may injure employee beneficiaries when they fire
them or lay them off, or when they decide to reduce or restructure the benefits that
are available under an employee benefit plan.26 However, when an ERISA
fiduciary makes fiduciary decisions—those involving discretionary judgment over
plan assets,27 such as determining the benefits due to a particular plan member28—
ERISA then requires the fiduciary to disregard its own interests and to make the
fiduciary decision for the sole benefit of the beneficiary and the plan as a whole.29

If an ERISA plan member or beneficiary is denied benefits to which he feels
entitled under the plan, ERISA accords him the right to bring a civil action “to
recover benefits due to him under the terms of his plan, to enforce his rights under

20 Id.
21 Id. § 1104(a)(1).
22 Id. § 1104(a)(1)(A); see, e.g., Bedrick v. Travelers Ins. Co., 93 F.3d 149, 154 (4th
Cir. 1996) (stating that “[a] fiduciary with a conflict of interest must act as if he is ‘free’ of
such a conflict”).
23 Music v. W. Conference of Teamsters Pension Trust Fund, 712 F.2d 413, 417 (9th
use “an officer, employee, agent or other representative” as a fiduciary); see also Pegram v.
Herdrich, 530 U.S. 211, 225 (2000) (commenting that ERISA fiduciaries may have
interests adverse to beneficiaries).
25 Pegram, 530 U.S. at 225.
26 Id.
fiduciary with respect to a plan to the extent (i) he exercises any discretionary authority or
discretionary control respecting management of such plan or exercises any authority or
control respecting management or disposition of its assets”).
28 See Rud v. Liberty Life Assurance Co. of Boston, 438 F.3d 772, 774 (7th Cir.
2006) (noting that an insurer is a fiduciary when it makes determinations of eligibility to
receive benefits under the plan).
29 Pegram, 530 U.S. at 225 (stating that “ERISA does require, however, that the
fiduciary with two hats wear only one at a time, and wear the fiduciary hat when making
fiduciary decisions”).
the terms of the plan, or to clarify his rights to future benefits under the terms of
the plan.” ERISA does not, however, specify the standard of review that courts
should apply in reviewing the plan member’s claim that he was wrongfully denied
a plan benefit by a fiduciary.

B. The Firestone Standard of Review

In the 1989 landmark case *Firestone Tire & Rubber Co. v. Bruch*, the
Supreme Court supplied ERISA’s missing standard of review as applied to
members’ challenges of benefit denials. In *Firestone*, six former Firestone
employees challenged the company’s denial of severance benefits under its
termination pay plan. The central issue raised on appeal was whether Firestone’s
interpretation of its plan was to be reviewed de novo or deferentially.

Applying trust principles, the Court determined that because the plan did not
give Firestone, as administrator, discretionary authority to construe ambiguous
terms in the plan, Firestone’s interpretation of the plan would be subject to de
novo review. In what is arguably the most quoted passage in ERISA cases, the
Court declared the standard of review as follows: “[W]e hold that a denial of
benefits challenged under § 1132(a)(1)(B) is to be reviewed under a de novo
standard unless the benefit plan gives the administrator or fiduciary discretionary
authority to determine eligibility for benefits or to construe the terms of the
plan.” Under this standard, whenever a trustee is “given power to construe
disputed or doubtful terms [in the plan], . . . the trustee’s interpretation will not be
disturbed if reasonable.”

Also, the Court commented on how the existence of a conflict of interest by a
fiduciary would affect the standard of review. Where a plan does not confer
discretion on the fiduciary—regardless of whether the fiduciary is acting under a

---

“ERISA does not set out the appropriate standard of review for actions under §
1132(a)(1)(B) challenging benefit eligibility determinations”).
32 *Id.*
33 *Id.* at 105.
34 *Id.* Deferential review requires courts to leave fiduciaries’ discretionary decisions
undisturbed if they are reasonable. *Id.* at 110. In contrast, “deference is unwarranted” in de
novo review, allowing courts to replace the fiduciaries’ decisions with others the courts
prefer. *Id.* at 107.
35 *Id.* at 111 (“In determining the appropriate standard of review for actions under §
1132(a)(1)(B), we are guided by principles of trust law.”).
36 *Id.* (finding “no evidence that under Firestone’s termination pay plan the
administrator has the power to construe uncertain terms or that eligibility determinations
are to be given deference”).
37 *Id.* (stating that “a deferential standard of review [is] appropriate when a trustee
exercises discretionary powers”).
38 *Id.* at 115.
39 *Id.* at 111.
possible or actual conflict of interest—review of the fiduciary’s decision will be de novo.\textsuperscript{40} However, in dicta that became the subject of much controversy, the Court further declared that when the plan confers discretionary authority on a fiduciary who acts under a conflict, the review will be deferential, but “that conflict must be weighed as a ‘factor’ in determining whether there is an abuse of discretion.”\textsuperscript{41}

Ultimately, pronouncement of the \textit{Firestone} standard of review resulted in the amendment of virtually all ERISA plans to confer discretion on fiduciaries to construe or interpret the terms of their plans.\textsuperscript{42} By so doing, employers secured for themselves under the \textit{Firestone} standard the substantial benefit of receiving deferential rather than de novo review when plan members challenged benefit denials. This deferential standard gave employers greater ability to determine claims in a way that better controlled the costs of their employee benefit plans, as they could construe plan terms to promote cost efficiency and would be subject to reversal by a court only if their interpretations were arbitrary and capricious.\textsuperscript{43} Thus, although recognizing that ERISA was promulgated to protect employee benefit plans for the benefit of employee members, the Court provided a means whereby employers could insulate their benefit denials from more searching de novo review.\textsuperscript{44}

III. COURTS’ CONFUSED APPLICATION OF THE FIRESTONE STANDARD TO CONFLICTS OF INTEREST

Application of the \textit{Firestone} standard did not produce uniform results. When post-\textit{Firestone} courts applied it, they could not agree on the extent to which fiduciaries presented a conflict situation,\textsuperscript{45} or how weighing the conflict as a factor should alter the review.\textsuperscript{46}

\textbf{A. Courts’ Disagreement on the Existence of a Conflict}

Where an ERISA plan purchased health or disability coverage from an insurance company, typically the insurance company determined whether claims would be paid or denied under the terms of its policy. Under this scenario, courts

\begin{itemize}
  \item \textsuperscript{40} \textit{Id.} at 115 ("[T]he de novo standard of review applies regardless of whether . . . the administrator or fiduciary is operating under a possible or actual conflict of interest.").
  \item \textsuperscript{41} \textit{Id.} (quoting \textit{RESTATEMENT (SECOND) OF TRUSTS} § 187 cmt. d (1959)).
  \item \textsuperscript{42} \textit{See} Langbein, \textit{supra} note 16, at 1324 (observing that “[p]lan drafters routinely seize upon Bruch’s invitation to instruct the courts to defer to plan decisionmaking”).
  \item \textsuperscript{43} A decision is arbitrary and capricious if it lacks substantial evidence, based upon the record as a whole. \textit{See, e.g.,} Caldwell v. Life Ins. Co. of N. Am., 287 F.3d 1276, 1282 (10th Cir. 2002).
  \item \textsuperscript{44} \textit{See} Langbein, \textit{supra} note 16, at 1323 (characterizing the \textit{Firestone} court’s determination that “the employer or other plan sponsor has the authority to defeat the de novo standard” as a “disastrous misstep”).
  \item \textsuperscript{45} \textit{See infra} Part III.A.
  \item \textsuperscript{46} \textit{See infra} Part III.B.
\end{itemize}
generally agreed that the insurer was a fiduciary of the plan acting under at least a potential conflict of interest. The potential conflict arose because if an insurance company denied a claim, it would not have to pay the claim out of its own funds collected from premiums. The courts disagreed, however, on whether the mere possibility that an insurance company might deny claims due to financial self-interest was a conflict sufficient enough to alter either the standard of review or the court’s ultimate ruling. While some courts found that an insurer’s financial incentive to deny claims was sufficient to demonstrate actual conflict, others required a more concrete showing.

47 See, e.g., Maciejczak v. Procter & Gamble Co., 246 F. App’x 130, 132 (3d Cir. 2007) (“Conflicts of interest are far more likely to arise when an insurance company . . . both funds and administers the plan.”); Abatie v. Alta Health & Life Ins. Co., 458 F.3d 955, 965 (9th Cir. 2006) (observing that “an insurer that acts as both the plan administrator and the funding source for benefits operates under what may be termed a structural conflict of interest”); Brown v. Blue Cross & Blue Shield of Ala., Inc., 898 F.2d 1556, 1562 (11th Cir. 1990) (asserting that a ‘strong conflict of interest [exists] when the fiduciary making a discretionary decision is also the insurance company responsible for paying the claims”).

48 See, e.g., Carolina Care Plan Inc. v. McKenzie, 467 F.3d 383, 387 (4th Cir. 2006) (stating that “when an entity both administers and insures a plan, its profits (or losses) depend in part on how the actual cost of providing coverage diverges from the projections on which it based premiums. Over time, a predilection to deny coverage pays well, even for inexpensive and infrequent treatments. . . . [B]ut in reality, health insurance companies benefit financially when they deny claims”); Lemaire v. Hartford Life & Accident Ins. Co., 69 F. App’x 88, 92 (3d Cir. 2003) (“[A] conflict of interest is presumed where an insurance company both determines eligibility for benefits and pays out those benefits from its own funds”); Pinto v. Reliance Standard Life Ins. Co., 214 F.3d 377, 388 (3d Cir. 2000) (stating that “insurance carriers have an active incentive to deny close claims in order to keep costs down and keep themselves competitive so that companies will choose to use them as their insurers”); Pitman v. Blue Cross & Blue Shield of Okla., 24 F.3d 118, 122 (10th Cir. 1994) (stating that an insurer’s “fiduciary role lies in perpetual conflict with its profit-making role as a business”); see also Langbein, supra note 16, at 1331 (noting that even when an employer’s insurance policy is experience rated, wherein the employer agrees to reimburse the insurer for benefit payments, the insurer may still have an incentive to deny claims “because the market for insurance services is intensely competitive” and “[t]he more effectively an insurer contains costs under an experience-rated policy, the better that insurer’s chance of retaining the account and getting others”).

49 See, e.g., Lemaire, 69 F. App’x at 92 (stating that “a conflict of interest is presumed where an insurance company both determines eligibility for benefits and pays out those benefits from its own funds, because there exists ‘an active incentive to deny close claims in order to keep costs down and keep themselves competitive so that companies will choose to use them as their insurers’” (citation omitted)); Brown, 898 F.2d at 1562 (finding a “strong conflict of interest . . . when the fiduciary making a discretionary decision is also the insurance company responsible for paying the claims” (citation omitted)).

50 See, e.g., Schatz v. Mutual of Omaha Ins. Co., 220 F.3d 944, 947 (8th Cir. 2000) (stating that a “less deferential standard of review is triggered where the claimant presents ‘material, probative evidence demonstrating that . . . a palpable conflict of interest existed’” that caused the plan fiduciary to breach its fiduciary duties (quoting Woo v. Deluxe Corp., 144 F.3d 1157, 1160 (8th Cir. 1998))); Jones v. Aetna U.S. Healthcare, 136 F. Supp. 2d
For example, the Seventh Circuit Court of Appeals balked at finding a conflict “whenever an insurer is being asked to dip into its own pocket to pay a claim for benefits.”\(^1\) Observing that the parties to every contract have adverse interests, the court declared that “[t]he ubiquity of such a situation makes us hesitate to describe it as a conflict of interest.”\(^2\) Therefore, the court refused to alter its deferential standard of review unless the claimant presented substantive evidence showing some basis to assume a conflict.\(^3\) The court favored requiring the claimant to “demonstrate the existence of a real and not merely notional conflict of interest” because “[i]t strikes the right balance, consistent with *Firestone*, between freedom of contract and a realistic sense of its limits.”\(^4\)

Other courts, also conceding an insurer fiduciary’s “apparent” conflict, required even more to recognize an “actual” conflict. These courts required the claimant to make a showing that the insurer’s financial self-interest had infected its determination of the claim.\(^5\) Then the courts examined the insurance company’s justification for its denial, finding an actual conflict only where the reasons

---
\(^1\) Rud v. Liberty Life Assurance Co. of Boston, 438 F.3d 772, 775 (7th Cir. 2006)
\(^2\) *Id.*
\(^3\) *Id.* at 777 (stating that it might, for example, find that an employer interpreting its own plan was conflicted where the claimant presented evidence that the plan was underfunded, that the claim was large, and that honoring it would require the employer to transfer additional assets to the plan).
\(^4\) *Id.*

\(^5\) See, e.g., Smith v. First Unum Life Ins. Co., 157 F. App’x 332, 333 (2d Cir. 2005) (holding that the defendant insurer’s dual status as a plan administrator and plan insurer does not demonstrate a conflict of interest per se); Bendixen v. Standard Ins. Co., 185 F.3d 939, 943 (9th Cir. 1999) (“In order to establish a serious conflict, the beneficiary has the burden to come forward with ‘material, probative evidence, beyond the mere fact of the apparent conflict, tending to show that the fiduciary’s self-interest caused a breach of the administrator’s fiduciary obligations to the beneficiary.’” (quoting Atwood v. Newmont Gold Co., Inc., 45 F.3d 1317, 1323 (9th Cir. 1995))); Mers v. Marriott Int’l Group Accidental Death & Dismemberment Plan, 144 F.3d 1014, 1020 (7th Cir. 1998) (presuming that “a fiduciary is acting neutrally unless a claimant shows by providing specific evidence of actual bias that there is a significant conflict”); Whitney v. Empire Blue Cross & Blue Shield, 106 F.3d 475, 477 (2d Cir. 1997) (stating that “a reasonable interpretation of the Plan will stand unless the participants can show not only that a potential conflict of interest exists, . . . but that the ‘conflict affected the reasonableness of the Committee’s decision’” (quoting Sullivan v. LTV Aerospace & Defense Co., 82 F.3d 1251, 1259 (2d Cir. 1996))); Mikrut v. Unum Life Ins. Co. of Am., No. 3:03cv1714 (SRU), 2006 U.S. Dist. LEXIS 92265, at *21–22 (D. Conn. Dec. 20, 2006) (noting that “Unum’s dual role as both plan administrator and plan insurer is alone insufficient as a matter of law to trigger stricter review,” and that “[t]he plaintiff must also show that Unum’s decision was in fact influenced by the conflict” to alter the standard of review).
proffered by the insurer for the denial were inconsistent with the evidence in the record.\textsuperscript{56}

Several courts evaluated the potential for an insurer’s conflict by examining whether the insurer was determining claims under its own policy of insurance that it had sold to an ERISA health benefits plan,\textsuperscript{57} or whether it was doing so as a third-party administrator of a self-insured plan,\textsuperscript{58} retained by the employer to process the claims on behalf of the plan.\textsuperscript{59} Where a claims denial was made by a third-party administrator of a self-insured plan, the administrator admittedly had no direct economic motive to deny the claim.\textsuperscript{60} Ostensibly, the third-party administrator would receive its administration fee from the plan regardless of how the claim was decided, and payment of the claim was coming from the employer,

\textsuperscript{56} See, e.g., Lemaire v. Hartford Life & Accident Ins. Co., 69 F. App’x 88, 93 (3d Cir. 2003) (finding that “Hartford’s conflict of interest drove its decision-making,” but that the decision was arbitrary and capricious because Hartford’s determination “lacked both logic and support in the record”); see also additional cases discussed infra Part V.B.

Some courts noted that the amount of the claim might imply that self-interest had tainted the insurer’s denial. See, e.g., Carolina Care Plan Inc. v. McKenzie, 467 F.3d 383, 387 (4th Cir. 2006) (stating that “a frequent and expensive claim might well demand comparatively more scrutiny on the ‘sliding scale’ than an inexpensive and infrequent claim”); Calvert v. Firstar Fin., Inc., 409 F.3d 286, 292 (6th Cir. 2005) (considering the factor that payment of the permanent disability claim would be expensive for the insurer); Pinto v. Reliance Standard Life Ins. Co., 214 F.3d 377, 386 (3d Cir. 2000) (stating that where more money was at stake, “the potential conflict might invite closer scrutiny”); Adamson v. Wadley Health Sys., Civ. No. 07-1081, 2008 U.S. Dist. LEXIS 89335, at *26 (W.D. Ark. Oct. 20, 2008) (declaring that the fact that the denial of benefits would save the plan $200,000 demonstrated a nexus between the conflict of interest and breach of fiduciary duty). \textit{But see Mers}, 144 F.3d at 1020–21 (opining that the impact of granting or denying a $200,000 benefits claim is “minuscule compared to AIG’s bottom line” and that the amount may be “too slight to compromise the impartiality of the trustees” (citing Van Boxel v. Journal Co. Employees’ Pension Trust, 836 F.2d 1048, 1061 (7th Cir. 1991))).

\textsuperscript{57} ERISA plans that purchase insurance to cover the beneficiaries are referred to as “insured plans.” See FMC Corp. v. Holliday, 498 U.S. 52, 66 (1990) (Stevens, J., dissenting) (contrasting “self-insured plans” and “insured plans”).

\textsuperscript{58} A self-insured employee benefits plan is one in which the employer itself pays the claims, and either pays a third-party administrator to interpret the plan and make plan benefits determinations, or creates an internal benefits committee vested with the discretion to interpret the plan’s terms and administer benefits. \textit{See Pinto}, 214 F.3d at 383 (explaining the concept of self-funded plans).

\textsuperscript{59} See id. (stating that self-funded employee benefit plans could hire an independent third party (often insurance companies) to interpret the plan and make plan benefit determinations).

\textsuperscript{60} See Metro. Life Ins. Co. v. Glenn, 128 S. Ct. 2343, 2349 (2008) (observing that, with regard to an insurance company acting as a third party administrator for an ERISA plan, “paying an individual claim does not come to the same extent from the company’s own pocket”).
not from the insurer’s own coffers. For this reason, some courts were reluctant to recognize a conflict when the benefit denial was by a third-party administrator.

Other courts, however, recognized that even a third-party administrator might have motives to deny claims. For example, an administrator might want to demonstrate to the plan that it processed claims in a cost-effective manner, so that it would be more likely to be rehired as administrator for another contract term. An insurance company third-party administrator that is also selling health or disability insurance policies to ERISA health benefits plans and private individuals might have an incentive to determine the third-party claims in a manner that would be consistent with the denials it makes for its other business. Therefore, third-party claims processing does not necessarily indicate an administrator is free from conflict.

Finally, courts have considered whether a conflict of interest is likely to exist where the employer itself functions as the claims administrator. Some courts have recognized that although conflicts of interest may be less likely to arise when an employer is funding and administering an ERISA health or disability plan.

---

61 See id. (observing that “the insurance company typically charges a fee that attempts to account for the cost of claims payouts”).

62 See, e.g., Pinto, 214 F.3d at 383 (stating that arrangements where the employer funds a plan and pays an independent third party to interpret the plan and make plan benefit determinations “do not, in themselves, typically constitute the kind of conflict of interest mentioned in Firestone”); Ladd v. ITT Corp., 148 F.3d 753, 754 (7th Cir. 1998) (speculating that because MetLife functioned only as an administrator of the ERISA plan, and not as its insurer, “it [might] be tempting to infer that [it does not have] a conflict of interest in administering the plan”).

63 See Ladd, 148 F.3d at 754 (speculating that MetLife, hired by the employer as the claims administrator for the plan, might be “inclined to resolve close cases against the claimant” when the employer has to “dig into its own pocket if claims exceeded contributions”).

64 In a similar context, health care commentators have opined that insurance companies acting as intermediaries or carriers for the Medicare program might have an incentive to determine Medicare claims in a manner that is consistent with their determination of similar claims in their private business. See, e.g., RAND E. ROSENBLATT ET AL., LAW AND THE AMERICAN HEALTH CARE SYSTEM 393–94 (1997) (stating that when health insurers act as both private insurers and Medicare contractors, they may have “a financial interest in denying Medicare coverage in order to restrict coverage in their private policies”).

65 See, e.g., Leffew v. Ford Motor Co., 258 F. App’x 772, 778 (6th Cir. 2007) (stating that “[w]hen . . . the employer operates the plan as both the administrator determining which claims are covered and the sponsor responsible for paying those claims, there is a readily apparent conflict of interest”).

66 See, e.g., Maciejczak v. Procter & Gamble Co., 246 F. App’x 130, 132 (3d Cir. 2007) (observing that “[e]mployer-funded plans present a decreased 'risk of a conflict of interest . . . because the employer has incentives to avoid the loss of morale and higher wage demands that could result from denials of benefits’” (quoting Smathers v. Multi-Tool, Inc., 298 F.3d 191, 197 (3d Cir. 2002))); Rud v. Liberty Life Assurance Co. of Boston, 438 F.3d 772, 776 (7th Cir. 2006) (pointing out that “an employer cannot reap a long-run
conflicts could arise based upon how a plan is structured. For example, an employer who pays claims from a fixed fund might have little incentive to deny payment for a particular claim, as the denial would not affect the amount of funding the employer is contributing to the plan. On the other hand, where an employer structures the plan so that claims are paid out of the company’s general operating revenues, the employer might have a stronger financial incentive to deny claims. Further, internal procedures under an ERISA plan may signal that claims benefit from reducing welfare benefits”); *Pinto*, 214 F.3d at 389 (stating that employer-funded plans are less likely to be conflicted because the employer has “incentives to avoid the loss of morale and higher wage demands that could result from denials of benefits” (quoting *Nazay v. Miller*, 949 F.2d 1323, 1335 (3d Cir. 1991))).

67 *See, e.g.*, *Maciejczak*, 246 F. App’x at 132 (observing that where the employer makes fixed contributions to the plan, there is likely to be little conflict, as “the employer ‘incurs no direct expense as a result of the allowance of benefits, nor does it benefit directly from the denial or discontinuation of benefits’” (quoting *Pinto*, 214 F.3d at 388)); *Rud*, 438 F.3d at 777 (stating that where “benefits are paid from a trust fund that has ample assets that cannot be transferred back to the employer,” the employer’s stake in a particular benefits determination is minimal); *Pinto*, 214 F.3d at 386 (stating that the employer’s conflict is “insufficiently compelling” where its “contributions to the [employee benefit] plan were fixed such that it ‘incurs no direct expense as a result of the allowance of benefits, nor does it benefit directly from the denial or discontinuation of benefits’” (quoting *Abnathya v. Hoffmann-La Roche, Inc.*, 2 F.3d 40, 45 n.5 (3d Cir. 1993))).

68 *See, e.g.*, *Maciejczak*, 246 F. App’x at 132 (declaring that “heightened scrutiny may be appropriate when the ‘plan is “unfunded,” that is, when it pays benefits out of operating funds rather than from a separate ERISA trust fund’” (quoting *Vitale v. Latrobe Area Hosp.*, 420 F.3d 278, 2882 (3d Cir. 2005))); *Post v. Hartford Ins. Co.*, 501 F.3d 154, 163 (3d Cir. 2007) (stating that a plan’s financial structure raises concern when the plan is funded on a “case-by-case basis,” that is, where “the administrator pays the claims out of the [employer’s] operating budget, rather than from segregated monies that the employer sets aside according to an actuarial formula,” because “it means that each dollar paid out is a dollar out of the administrator’s pocket” so that “the administrator has a financial incentive to deny claims”); *Rud*, 438 F.3d at 777 (observing that where a plan “is underfunded, the benefits claim is large, [and] the employer will have to transfer assets to the plan if the claim is honored,” there is an increased risk of “sharp dealing”); see also *Houston v. Unum Life Ins. Co. of Am.*, 246 F. App’x 293, 300 (6th Cir. 2007) (stating that “[c]ourts should be particularly vigilant in situations where . . . the plan sponsor bears all or most of the risk of paying claims” (quoting *Univ. Hosps. of Cleveland v. Emerson Elec. Co.*, 202 F.3d 839, 846 n.4 (6th Cir. 2000))); *Williams v. BellSouth Telecommms., Inc.*, 373 F.3d 1132, 1135–37 (11th Cir. 2004) (finding a self-insured plan to be conflicted despite its retention of a third-party administrator to process claims, where the employer retained the ultimate authority to determine when payments should be made out of its own assets); *Peruzzi v. Summa Med. Plan*, 137 F.3d 431, 433 (6th Cir. 1998) (recognizing “the conflict of interest inherent in self-funded plans”); *Ladd*, 148 F.3d at 754 (speculating that ITT might have a conflict of interest because it “would have to dig into its own pocket if claims exceeded contributions”); *Godfrey v. BellSouth Telecommms., Inc.*, 89 F.3d 755, 758 (11th Cir. 1996) (finding that the employer was conflicted because granting disability benefits to Godfrey would require it to hire a replacement worker or lose the services of Godfrey’s position).
are being determined in a conflicted manner. For example, where an employer gives bonuses to claims processors based upon the dollar value of claim denials— as opposed to rewarding claims processors for accuracy—a court might be more likely to find that the claims determinations are influenced by the employer’s financial self-interest.\textsuperscript{69}

\textbf{B. Courts’ Disagreement on How an Identified Conflict Alters the Standard of Review}

While courts generally have recognized that, per \textit{Firestone}, identification of a conflict required the conflict to be “weighed as a factor” in determining whether the claim had been unreasonably denied, they did not agree at all as to how the weighing was to be carried out.\textsuperscript{70} In fact, courts developed a number of approaches to altering the manner of review once a conflict was identified.

Some courts refused to alter the standard of review at all based upon the mere potential for conflict.\textsuperscript{71} However, upon finding that a conflict actually had affected

\textsuperscript{69} See Potter v. Liberty Life Assurance Co. of Boston, 132 F. App’x 253, 259 (11th Cir. 2005) (stating that the insurer argued that its claims reviewers were not compensated or given bonuses for denying claims).

\textsuperscript{70} See, e.g., Fought v. Unum Life Ins. Co., 379 F.3d 997, 1003–04, 1008 (10th Cir. 2004) (stating that while all circuit courts agree that a conflict of interest triggers a less deferential standard of review, “[t]he courts . . . differ over how this lesser degree of deference alters their review” and that it is “difficult to distinguish [heightened deferential review] from pure arbitrary and capricious deference”); Chambers v. Family Health Plan Corp., 100 F.3d 818, 825 (10th Cir. 1996) (stating that the courts “differ over how this lesser degree of deference alters their review process”); see also Abatie v. Alta Health & Life Ins. Co., 458 F.3d 955, 969 (9th Cir. 2006) (stating that “an inherent conflict of interest . . . ought to have some effect on judicial review [and t]he question is, what effect,” and observing that the review is “indefinite,” requiring “something akin to a credibility determination” of the administrator’s motives); Williams, 373 F.3d at 1138 (describing “heightened arbitrary and capricious review” as somewhere between the \textit{de novo} and “mere” arbitrary and capricious standards,” but questioning “where is that ‘somewhere’?”); Pinto, 214 F.3d at 392 (stating that “[o]nce the conflict becomes a ‘factor’ . . . it is not clear how the process required by the typical arbitrary and capricious review changes”).

\textsuperscript{71} See, e.g., Werdehausen v. Benicorp Ins. Co., 487 F.3d 660, 664 (8th Cir. 2007) (stating that to obtain a less deferential review, a claimant must “show a ‘palpable’ financial conflict of interest that ‘has a connection to the substantive decision reached’ and raises ‘serious doubts as to whether the result reached was the product of an arbitrary decision or the plan administrator’s whim’” (quoting Sahulka v. Lucent Techs., Inc., 206 F.3d 763, 768 (8th Cir. 2000))); Wright v. R. R. Donnelley & Sons Co. Group Benefits Plan, 402 F.3d 67, 74 (1st Cir. 2005) (stating that a “‘conjectural conflict will not strip the fiduciary’s determination of the deference that otherwise would be due’”(quoting Leahy v. Raytheon Co., 315 F.3d 11, 16 (1st Cir. 2002))); Smith v. Fortis Benefits Ins. Co., 77 F. App’x 532, 533 (1st Cir. 2003) (stating that the arbitrary and capricious standard is not altered by a potential conflict of interest); Bendixen v. Standard Ins. Co., 185 F.3d 939, 943 (9th Cir. 1999) (stating that to alter the traditional abuse of discretion review, the
a fiduciary’s decision, some courts went to the other extreme, giving no deference and applying de novo review to neutralize the taint of the conflict.  

Probably the most prevalent approach to weighing the conflict as a factor was to assess the seriousness of the conflict and—while still retaining the deferential standard—to accord more or less deference to the fiduciary’s decision depending on how obvious and deleterious the conflict appeared to be. Some of these courts expressly referred to a “sliding scale” to be used for determining the level of deference. Others did not use the term “sliding scale,” but spoke more generally
in terms of lessening deference in proportion to the seriousness of the conflict.\textsuperscript{75} For the most part, courts determined the seriousness of the conflict by examining the manner of the administrator’s denial and its stated justification. Depending on the extent to which the record did or did not support the denial, the courts assigned a rough approximation of the seriousness of the conflict and diminished the deference depending upon how injurious it adjudged the conflict to be.\textsuperscript{76} In cases where the fiduciary’s denial appeared to be rationally supported by substantial evidence in the record, the deference accorded was at the high end of the scale.\textsuperscript{77} For example, in cases where the administrator’s decision was not supported by the record, and there were a number of unexplained irregularities by the fiduciary or inconsistencies between the record and the fiduciary’s conclusions, the courts diminished the deference to the lower end of the scale and subjected the denial to increased scrutiny.\textsuperscript{78}

\textsuperscript{75} See, e.g., DeGrado v. Jefferson Pilot Fin. Ins. Co., 451 F.3d 1161, 1168 (10th Cir. 2006) (stating that “[w]e . . . are obligated to ‘take a hard look at the evidence [to determine whether the fiduciary’s decision was] untainted by the conflict of interest’” (quoting Fought, 379 F.3d at 1003)); Ladd v. ITT Corp., 148 F.3d 753, 754 (7th Cir. 1998) (declaring that “judicial inquiry is more searching” where the administrator has a conflict of interest); Vega v. Nat’l Life Ins. Servs., Inc., 145 F.3d 673, 678 (5th Cir. 1998) (stating that identification of a conflict requires the court to “carefully consider” the administrator’s decision and to perform a “close examination”); Bedrick v. Travelers Ins. Co., 93 F.3d 149, 152 (4th Cir. 1996) (stating that “deference will be lessened to the degree necessary to neutralize any untoward influence resulting from the conflict” (quoting Bailey v. Blue Cross & Blue Shield of Virginia, 67 F.3d 53, 56 (4th Cir. 1995))).

\textsuperscript{76} See, e.g., Kosiba v. Merck & Co., 384 F.3d 58, 61 (3d Cir. 2004) (applying a “moderately heightened” arbitrary and capricious standard of review where there was evidence of procedural bias in the employer’s intervention in the appeals process); Lasser, 344 F.3d at 385 (3d Cir. 2003) (applying a “moderate degree of deference” in view of an “inherent structural conflict” but no evidence that it had tainted the claims decision).

\textsuperscript{77} See, e.g., Wade, 493 F.3d at 541 (stating that “[a] potential conflict. . . where an Administrator serves the dual role of both administrator and insurer, results in only a ‘modicum less deference’ than would otherwise be afforded”); Havens, 186 F. App’x at 212 (applying only a “low to moderate degree of heightened scrutiny, near the arbitrary and capricious end of the sliding scale”); Lasser, 344 F.3d at 385 (finding no evidence of conflict other than the inherent structural conflict, and therefore holding that the correct standard of review was “at the mild end of the heightened arbitrary and capricious scale” affording a “moderate degree of deference”); Vega, 214 F.3d at 394 (declaring that the sheer number of irregularities by the fiduciary warranted review via a “further heightened” standard); Pinto, 214 F.3d at 394 (declaring
Other courts adopted a “burden shifting” approach. Upon identification of a potential conflict, these courts shifted the burden to the fiduciary to show that its claims decision had not been tainted by self-interest and that the decision operated exclusively for the best interests of the plan members and their beneficiaries.

Still other courts developed hybrid approaches. For example, the Tenth Circuit adopted a sliding-scale approach in addition to shifting the burden to the administrator to prove that its decision was reasonable. The Ninth and Third Circuits, upon identifying a conflict, applied factors to determine the level of skepticism of the review.

This “burden shifting” approach is also known as the “presumptively void” approach, because a decision rendered by a conflicted fiduciary is presumed to be arbitrary unless the fiduciary can demonstrate that its decision was reasonable. See, e.g., Chambers v. Family Health Plan Corp., 100 F.3d 818, 826 (10th Cir. 1996) (stating that the administrator’s decision is presumptively void “unless the administrator can demonstrate that either (1) under de novo review, the result reached was nevertheless ‘right’ or (2) the decision was not made to serve the administrator’s conflicting interest”).

Finding that additional reduction in deference is appropriate once a conflict is identified, and declaring that “the plan administrator bears the burden of proving the reasonableness of its decision”).

See Buchanan v. Standard Ins. Co., No. 05-16651, 2007 U.S. App. LEXIS 24526, at *6 (9th Cir. Oct. 15, 2007) (stating that “[a] court may weigh a conflict more heavily if, for example, the administrator provides inconsistent reasons for denial, fails adequately to investigate a claim or ask the plaintiff for necessary evidence . . . or has repeatedly denied benefits to deserving participants by interpreting plan terms incorrectly or by making decisions against the weight of evidence in the record” (quoting Abatie v. Alta Health & Life Ins. Co., 458 F.3d 955, 968–69 (9th Cir. 2006))); Post v. Hartford Ins. Co., 501 F.3d 154, 163 (3d Cir. 2007) (identifying four factors to consider in determining the level of deference: “(1) the sophistication of the parties, (2) the information accessible to the

79 See, e.g., Fought, 379 F.3d at 1006 (adopting the burden-shifting approach).

80 See, e.g., Potter v. Liberty Life Assurance Co. of Boston, 132 F. App’x 253, 259 (11th Cir. 2005) (stating that even if a conflicted insurer’s denial is reasonable, it must show that it was not tainted by self-interest); Allison v. Unum Life Ins. Co. of Am., 381 F.3d 1015, 1022 (10th Cir. 2004) (stating that “when the plan administrator operates under . . . an inherent conflict of interest, . . . the plan administrator bears the burden of proving the reasonableness of its decision” (quoting Fought, 379 F.3d at 1006)); Newell v. Prudential Ins. Co. of Am., 904 F.2d 644, 651 (11th Cir. 1990) (declaring that a fiduciary’s decision is arbitrary “if it advances the conflicting interest of the fiduciary at the expense of the affected beneficiary . . . unless the fiduciary justifies the interpretation on the ground of its benefit to the class of all participants and beneficiaries” (quoting Brown v. Blue Cross & Blue Shield, Inc., 898 F.2d 1556, 1566–67 (11th Cir. 1990))); Brown, 898 F.2d at 1566 (stating that “when a plan beneficiary demonstrates a substantial conflict of interest on the part of the fiduciary . . . the burden shifts to the fiduciary to prove that its interpretation of plan provisions committed to its discretion was not tainted by self-interest”).
The most complex hybrid approach was undeniably the one developed by the Eleventh Circuit Court of Appeals. As the first step of its analysis, the Eleventh Circuit evaluated the claims administrator’s interpretation of the plan to determine whether it was “wrong.” This step required the court to perform a de novo review to determine whether it agreed or disagreed with the fiduciary’s denial of benefits. If the court determined that the fiduciary’s decision was “wrong,” it next decided whether the claimant had proposed a “reasonable” interpretation of the plan. If the fiduciary’s decision was wrong and the claimant’s proffered interpretation was reasonable, then the court determined whether the fiduciary’s interpretation was also reasonable. If the fiduciary’s denial was wrong but reasonable, then it was entitled to deference, and the plan prevailed. If the fiduciary was conflicted, however, its interpretation was not necessarily entitled to deference. In this case, the court proceeded to “gauge the self interest of the claims administrator [fiduciary],” and applied a “heightened arbitrary and beneficiary, (3) the financial arrangement between the employer and administrator, and (4) the financial status of the administrator”); Abatie, 458 F.3d at 967 (noting that comment d to section 187 of the Restatement of Trusts lists key factors in determining whether a trustee has abused its discretion, including “the motives of the trustee in exercising or refraining from exercising [a power granted to the trustee]; [and] the existence or nonexistence of an interest in the trustee conflicting with that of the beneficiaries”); Ceccanecchio v. Cont’l Cas. Co., 50 F. App’x 66, 70 (3d Cir. 2002) (applying a heightened level of review in conflict cases, sliding the scale to determine the measure of scrutiny, and considering the following factors: sophistication of the parties, information accessible to the parties, the financial arrangement between the insurer and the plan, and the financial status of the fiduciary); Siebert v. Standard Ins. Co. Group Long-Term Disability Policy, 220 F. Supp. 2d 1128, 1135 (C.D. Cal. 2002) (identifying the following factors, inter alia, as demonstrating an administrator’s actual conflict of interest: inconsistent reasons for denial of benefits, changes in position, reliance upon an improper definition of a plan term, no supporting evidence for a determination of a material fact, failure to provide reasons for denying benefits); see also Brown, 898 F.2d at 1564–65 (listing six factors from comment d to section 187 of the Restatement (Second) of Trusts and factors listed by a leading treatise to consider in determining whether a trustee abused its discretion, including whether the trustee had an interest conflicting with that of the beneficiaries).

83 See HCA Health Servs. of Ga., Inc. v. Employers Health Ins. Co., 240 F.3d 982, 993 (11th Cir. 2001) (declaring that “[r]egardless of whether arbitrary and capricious or heightened arbitrary and capricious review applies, the court evaluates the claims administrator’s interpretation of the plan to determine whether it is ‘wrong’”).

84 Id. (“[I]t is fundamental that the fiduciary’s interpretation first must be ‘wrong’ from the perspective of de novo review before a reviewing court is concerned with the self-interest of the fiduciary.” (quoting Brown, 898 F.2d at 1566 n.12)). The court defined “wrong” as meaning that “the court disagrees with the claims administrator’s plan interpretation.” Id. at 994 n.23.
capricious” review.\textsuperscript{90} The heightened standard shifted the burden to the fiduciary to show that “its wrong but reasonable interpretation of the plan benefits the class of participants and beneficiaries.”\textsuperscript{91} Finally, even if the fiduciary satisfied this burden,\textsuperscript{92} the claimant still was entitled to benefits if he could show that the fiduciary’s determination was arbitrary and capricious for other reasons.\textsuperscript{93}

IV. \textit{METROPOLITAN LIFE INSURANCE CO. V. GLENN: THE SUPREME COURT’S RECENT PRONOUNCEMENT OF A STANDARD OF REVIEW FOR ERISA PLAN BENEFIT DENIALS BY A CONFLICTED FIDUCIARY}

Due to the numerous unresolved differences among the circuit courts regarding when a conflict existed and how it impacted the standard of review in ERISA benefit denial cases, the Supreme Court granted certiorari in \textit{Metropolitan Life Insurance Co. v. Glenn},\textsuperscript{94} decided on June 19, 2008. Metropolitan Life Insurance Co. (hereinafter “MetLife”) was the administrator and insurer for Sears, Roebuck & Co.’s long-term disability plan, which was governed by ERISA.\textsuperscript{95} The plan gave MetLife discretion to determine employees’ claims, which MetLife paid out of its own funds.\textsuperscript{96} When claimant Wanda Glenn was diagnosed with a heart disorder, MetLife paid her an initial twenty-four months of disability benefits.\textsuperscript{97} It also encouraged her to apply for Social Security permanent disability benefits, which she was awarded based on the agency’s decision that she could no longer perform the tasks required by any job for which she was qualified.\textsuperscript{98} MetLife retained most of these benefits as an offset to its own plan benefits.\textsuperscript{99}

After Glenn’s initial twenty-four months of disability benefits from MetLife were exhausted, MetLife required Glenn to meet a stricter standard to receive extended plan benefits. The stricter standard required that Glenn be incapable of performing the duties not only of her former position, but of any occupation for

\textsuperscript{90} Id.
\textsuperscript{91} Id. at 994–95.
\textsuperscript{92} This burden could be satisfied by a showing that the insurer’s interpretation of the plan was “calculated to maximize benefits to participants in a cost-efficient manner.” Id. at 1001 (quoting Lee v. Blue Cross/Blue Shield, 10 F.3d 1547, 1552 (11th Cir. 1994)).
\textsuperscript{93} Id. at 995. \textit{But see} Doyle v. Liberty Life Assurance Co. of Boston, 511 F.3d 1336, 1344–46 (11th Cir. 2007) (opining that the Eleventh Circuit’s heightened standard of review was “flawed” because it was (1) “inconsistent with the Supreme Court’s announcement in Bruch of two standards under which an administrator’s decision should be reviewed,” (2) was contrary to traditional burden of proof requirements, and (3) placed a “remarkably difficult burden” on administrators to prove that their decisions were not tainted by conflict).
\textsuperscript{94} 128 S. Ct. 2343 (2008).
\textsuperscript{95} Id. at 2346.
\textsuperscript{96} Id.
\textsuperscript{97} Id.
\textsuperscript{98} Id. at 2346–47.
\textsuperscript{99} Id. at 2347.
which she was “reasonably qualified.” Upon determining that Glenn was capable of performing full-time sedentary work, MetLife denied the extended benefits. After exhausting administrative appeals, Glenn brought suit in federal court.

A. The District Court’s Decision

Applying the Firestone standard of deferential review and weighing MetLife’s conflict of interest as a factor, the district court denied relief to Glenn. The court found that there was no inconsistency between MetLife’s denial of extended benefits and the Social Security Administration’s award of permanent benefits because MetLife had access to later medical records that had not been available to Social Security. The court viewed the opinion of Glenn’s personal physician, Dr. Patel, as suspect because he had initially stated Glenn could do sedentary work, and then later recanted, stating she was unable to return to any employment. The court pointed to other evidence that supported MetLife’s decision. An independent consultant in internal medicine, Dr. Moyer, had concluded that Glenn was able to perform sedentary work. A vocational coordinator had found there were alternative sedentary occupations available in Glenn’s geographic area that she could perform. The report of another cardiologist, Dr. Pujara, indicated Glenn was capable of sedentary work, at least on a trial basis. Supported by this substantial evidence, the district court found MetLife’s denial was rational and not an abuse of discretion.

B. The Sixth Circuit Court of Appeals’ Reversal

The Sixth Circuit Court of Appeals reversed, declaring at the outset that it appeared the district court had not given MetLife’s conflict of interest any weight, and concluding that “this factor did not receive appropriate consideration by the

100 Id.
101 Id.
102 Id.
103 See Glenn v. Metro. Life Ins. Co., 2005 U.S. Dist. LEXIS 11007, *12 (S.D. Ohio June 8, 2005) (“In the instant case, defendants decide whether a claimant is eligible for benefits, and pays those benefits. A conflict of interest therefore exists which must be considered as a factor when applying the abuse of discretion standard.”).
104 Id. at *1.
105 Id. at *16.
106 Id. at *19–21.
107 Id. at *20.
108 Id.
109 Id. at *21.
110 Id. at *21–22.
district court." Specifically, the court of appeals disagreed with many of the district court’s factual findings. It faulted MetLife for failing to explain why it had disagreed with the Social Security Administration’s determination that Glenn was permanently disabled. The court found MetLife to be arbitrary in relying upon a check-off form filled out by Glenn’s physician, Dr. Patel, that said she was capable of sedentary work, but discrediting his numerous and more detailed written reports that clearly stated she was incapable of working. The court observed that none of the professional opinions MetLife relied upon was the result of a physical examination of Glenn. When MetLife requested Dr. Pujara to review the file, it failed to furnish him with letters from Dr. Patel. Finally, the court criticized MetLife for discounting stress as a factor in Glenn’s condition, when Glenn’s physician and others pointed out that it was an important aspect of her diagnosis. The court concluded that MetLife’s denial “was not the product of a principled reasoning process.”

The Supreme Court granted certiorari due to the conflicts among the circuit courts as to whether fiduciaries such as MetLife were conflicted, and how the conflict would affect the review.

C. The Supreme Court’s Affirmation

With regard to the threshold question of whether a plan administrator that both evaluates claims and pays them out of its own funds operates under a conflict of interest, the Court ruled in the affirmative. When an employer holds these dual roles, the Court declared that a conflict was “clear,” as “every dollar provided in benefits is a dollar spent by . . . the employer; and every dollar saved . . . is a dollar in [the employer’s] pocket.” Although conceding that the answer is “less clear” where the plan administrator is not the employer but an insurance company, the court concluded that a conflict nonetheless exists. The Court observed that the employer’s own conflict may bias its selection of an insurer, as the employer

112 Id. at 666.
113 Id. at 667.
114 Id. at 670.
115 Id. at 671.
116 Id.
117 Id. at 672–73.
118 Id. at 674.
119 See Metro. Life Ins. Co. v. Glenn, 128 S. Ct. 2343, 2347 (2008) (stating that MetLife had asked the Court to “determine whether a plan administrator that both evaluates and pays claims operates under a conflict of interest in making discretionary benefit determinations” and that the Solicitor General “suggested that we also consider ‘how’ any such conflict should ‘be taken into account on judicial review of a discretionary benefit determination’” (citation omitted)).
120 Id. at 2348–50.
121 Id. at 2348.
122 Id. at 2349.
may be more interested in low rates than in accurate claims processing.\textsuperscript{123} Moreover, the Court declared that ERISA’s fiduciary standards impose “higher-than-marketplace” quality standards on insurers.\textsuperscript{124} Even if there are differences in the conflicts under which employers and insurer administrators operate, the Court suggested that these differences can be taken into account when reviewing their fiduciary decisions.\textsuperscript{125}

Turning next to the question of how the conflict should be weighed as a factor, the Court declared that the Firestone standard did not imply that de novo review should occur whenever there was a conflict.\textsuperscript{126} In view of the fact that virtually all ERISA plan benefit claims are decided by fiduciaries that are conflicted to some extent (including employers, third-party administrators, and insurance companies providing the coverage), applying de novo review in any conflict situation would “bring about near universal review by judges de novo . . . of the lion’s share of ERISA plan claims denial.”\textsuperscript{127} Next, the Court rejected “special burden-of-proof rules” because they “create further complexity, adding time and expense to a process that may already be too costly for many of those who seek redress.”\textsuperscript{128} Ultimately, the Court adopted a conflict standard exactly as it was worded in Firestone, i.e., that the conflict should be weighed as a factor in determining whether there was an abuse of discretion.\textsuperscript{129}

In attempting to provide some direction to lower courts about how to apply the standard, the Court stated that the conflict could act as a tiebreaker when other factors were closely balanced.\textsuperscript{130} The conflict could be weighed more heavily when there was an increased likelihood that it had affected a particular benefit denial, or where there was a history of biased claims determinations.\textsuperscript{131} On the other hand, the conflict would be less influential where the administrator had taken steps to reduce bias, such as walling off claims processing personnel from finance departments, penalizing inaccurate claims determinations, or implementing incentives to promote accuracy.\textsuperscript{132}

\begin{itemize}
\item \textsuperscript{123} Id. at 2349–50.
\item \textsuperscript{124} Id. at 2350.
\item \textsuperscript{125} Id.
\item \textsuperscript{126} Id.
\item \textsuperscript{127} Id. Many believe that converting to de novo review for all ERISA claims will increase caseloads and be more time consuming. See, e.g., Langbein, supra note 16, at 1334 (noting the suspicion that resisting deferential review would cause caseloads to increase and would be more time consuming than presuming the correctness of the fiduciary’s decision).
\item \textsuperscript{128} Metro. Life Ins. Co., 128 S. Ct. at 2351.
\item \textsuperscript{129} Id. (“Firestone means what the word ‘factor’ implies, namely, that when judges review the lawfulness of benefit denials, they will often take account of several different considerations of which a conflict of interest is one.”).
\item \textsuperscript{130} Id.
\item \textsuperscript{131} Id.
\item \textsuperscript{132} Id.
\end{itemize}
Well aware that the chief defect with its reaffirmed *Firestone* standard was its lack of precision, the Court conceded that “our elucidation of *Firestone’s* standard does not consist of a detailed set of instructions.” Nevertheless, the Court declared that its standard was appropriate in view of “the intractability of any formula to furnish definiteness of content for all the impalpable factors involved in judicial review.”

Applying its standard to Glenn’s claim, the Court affirmed the Sixth Circuit’s ruling that MetLife had abused its discretion in denying benefits to Glenn. It held that the Sixth Circuit’s analysis properly “illustrates the combination-of-factors method of review,” whereby the court of appeals gave some weight to the conflict, but did not find the conflict to be determinative. Instead, the court of appeals had relied upon other factors, including MetLife’s ignoring the Social Security Administration’s finding, MetLife’s selective reliance upon only those reports that supported a denial of benefits, and MetLife’s failure to provide all of the available evidence to its independent consultants. As the court of appeals had appropriately weighed MetLife’s conflict along with these other “serious concerns,” the Court concluded that there was “nothing improper in the way in which the court [of appeals] conducted its review.”

In a concurring opinion, Chief Justice Roberts criticized the majority for endorsing an approach that weighs a conflict even when there has been no showing that the conflict affected the denial of benefits. Observing that a conflict of interest “is a common feature of ERISA plans,” Chief Justice Roberts opined that the majority’s invitation to weigh the conflict “invites the substitution of judicial discretion for the discretion of the plan administrator,” thereby effectively converting the review to de novo. Denigrating the majority’s approach as “kitchen-sink,” he concluded that “[t]he court leaves the law more uncertain, more unpredictable than it found it.”

In a separate opinion that concurred in part and concurred in the judgment, Justice Kennedy sought to remand the case to allow MetLife to defend its denial under the new standard. He urged that MetLife should be afforded the opportunity to show whether it “employed structural safeguards to avoid conflicts of interest,” to give MetLife “a chance to defend its decision under the standards the Court articulates today.”

---

133 Id. at 2352.
134 Id. (quoting Universal Camera Corp. v. NLRB, 71 S. Ct. 457, 577 (1951)).
135 Id.
136 Id. at 2351.
137 Id. at 2351–52.
138 Id. at 2352.
139 Id.
140 Id. at 2352–53 (Roberts, C.J., concurring).
141 Id. at 2353.
142 Id.
143 Id.
144 Id. at 2354.
145 Id. In a separate opinion that concurred in part and concurred in the judgment, Justice Kennedy sought to remand the case to allow MetLife to defend its denial under the new standard. Id. at 2356 (Kennedy, J., concurring in part and concurring in the judgment). He urged that MetLife should be afforded the opportunity to show whether it “employed structural safeguards to avoid conflicts of interest,” to give MetLife “a chance to defend its decision under the standards the Court articulates today.” Id.
Justice Scalia dissented, with Justice Kennedy joining. While the dissent agreed with the majority that MetLife had a conflict of interest, Justice Scalia stated that he had a “fundamental disagreement” with the majority’s determination of how the conflict should bear upon the judicial review.\footnote{Id. at 2357 (Scalia, J., dissenting).} Because the majority did not provide explicit instructions on how the review should be undertaken, the dissent characterized its standard as “unpredictable,” “painfully opaque,” and “nothing but de novo review in sheep’s clothing.”\footnote{Id. at 2358.}

Rather, the dissent argued that the conflict should never be weighed as a factor, but was relevant only for determining whether the fiduciary had acted with an improper motive, a separate basis under the Restatement of Trusts for finding an abuse of discretion.\footnote{Id. at 2360–61 (applying \textsc{Restatement (Second) of Trusts} § 187 cmt. i (1959)).} Rejecting the possibility that a conflict can ever be relevant to determining the “substantive reasonableness of [a fiduciary’s] decision,”\footnote{Id. at 2358.} the dissent urged a remand to determine the reasonableness of MetLife’s denial without regard to the existence of a conflict of interest.\footnote{Id. at 2361.}

V. HOW WEIGHING THE CONFLICT AS A FACTOR AFFECTS THE REVIEW: AN ANALYSIS OF THE POST-FIRESTONE CASES

Both Chief Justice Roberts’s concurring opinion and Justice Scalia’s dissenting opinion in \textit{MetLife} criticized the conflict standard adopted by the majority for its lack of precision. Justice Scalia described it as a “(so-called) test,”\footnote{Id. at 2357.} whereby the factors deemed relevant to the decision are “all . . . chucked into a brown paper bag and shaken up to determine the answer.”\footnote{Id. at 2358.} Chief Justice Roberts characterized the standard as “so imprecise” as to exacerbate the problem of how to treat the conflict.\footnote{Id. at 2360.}

With regard to the thousands of benefit denials that occur each year in ERISA health and disability benefit plans, it is extremely important to understand how the \textit{MetLife} conflict standard will be applied to benefit plan denials. Virtually all ERISA health and disability plans have delegated discretionary authority to their administrators, and these administrators are most likely to be insurance companies that are either administering their own policies of insurance (as in the \textit{MetLife} case) or processing claims as third-party administrators on behalf of the employer sponsors of the plans. As the \textit{MetLife} Court found, virtually all of these insurers, as well as the employers who hired them, perform their fiduciary duties under a
conflict of interest. Therefore, the MetLife conflict standard will be universally applied to review ERISA health and disability benefit denials.

Fortunately, since the MetLife Court simply adopted the historical conflict standard precisely as it was articulated in Firestone, there are hundreds of post-Firestone cases that have already applied the MetLife standard. An examination of these cases demonstrates how the MetLife conflict approach will likely impact the outcomes of cases wherein plan beneficiaries challenge benefit denials made by fiduciaries who are acting under a conflict of interest.

A. The Conflict Had Little Impact in Post-Firestone Cases on Determining Whether Fiduciaries Had Abused Their Discretion

Upon reviewing cases to determine the impact of the conflict on the courts’ findings of whether or not the fiduciary had abused its discretion, the conclusion is overwhelming that the conflict had little to no impact upon the courts’ ultimate findings.

This is most obvious in those cases where, despite an initial observation that the fiduciary had operated under a conflict of interest, the courts failed to refer to the conflict again as they proceeded through their analyses. Even in the numerous post-Firestone cases where each court explicitly recognized a conflict, noted that the conflict heightened the scrutiny of its review, and then actually found an abuse of discretion by the fiduciary, each court failed to link the conflict in any way to its ultimate finding that the fiduciary’s decision was unreasonable. Thus, there is no indication in these decisions of how, or whether, the conflict influenced the result.

In fact, demonstrating what appears to be an extreme reluctance to anchor a finding of abuse of discretion to the fiduciary’s conflict of interest, a number of courts, after first recognizing the fiduciary’s conflict, then explicitly disavowed that their decision had been influenced by the conflict at all. Instead, these courts found the fiduciary’s conflict irrelevant in concluding that the fiduciary had abused

---

156 See id. at 2348–50; see also Langbein, supra note 16, at 1325 (stating that “most ERISA plan benefit denials are the work of decisionmakers operating under serious conflicts of interest.”).


158 See, e.g., Leffew v. Ford Motor Co., 258 F. App’x 772, 779–81 (6th Cir. 2007); Force v. Ameritech Corp., 250 F. App’x 662, 667–70 (6th Cir. 2007); Brooking v. Hartford Life & Accident Ins. Co., 167 F. App’x 544, 546–50 (6th Cir. 2006); Darland v. Fortis Benefits Ins. Co., 317 F.3d 516, 528–33 (6th Cir. 2003); see also Houston v. Unum Life Ins. Co. of Am., 246 F. App’x 293, 299–300 (6th Cir. 2007) (implying the insurer was conflicted because it bears the risk of paying claims, but failing to discuss the conflict during the review).

its discretion, using instead the traditional arbitrary and capricious standard.\textsuperscript{160} Certainly, these cases demonstrate most emphatically that the conflict did not influence the outcome.

\textbf{B. Post-Firestone Cases Finding an Abuse of Discretion Relied on Substantive Inconsistencies Between the Fiduciary’s Decision and the Record}

Rather than relying upon the fiduciary’s conflict of interest, post-\textit{Firestone} courts that found an abuse of discretion by the fiduciary premised their holdings upon substantive inconsistencies between the fiduciary’s conclusions and the evidence in the record.\textsuperscript{161} In many of these cases, the fiduciary had failed to adequately explain the reasons for its denial,\textsuperscript{162} or had stated conclusions that were plainly inconsistent with undisputed medical evidence in the file.\textsuperscript{163} Often, the fiduciary had relied upon reports of consultants who had merely examined the claimant’s file, performed no physical exam,\textsuperscript{164} and rejected contrary reports by physicians who had treated claimant for years,\textsuperscript{165} a practice the courts criticized as selective evaluation of the medical proof.\textsuperscript{166} Courts also considered procedural


\textsuperscript{161} See, e.g., Gilchrest v. Unum Life Ins. Co. of Am., 255 F. App’x 38, 42 (6th Cir. 2007) (stating that the “record reveals that Unum ignored or disregarded undisputed evidence that Gilchrest was not performing the occupation defined by [the] job description”).

\textsuperscript{162} See, e.g., Bennett, 514 F.3d at 554 (registering its “serious concern that the final denial letter fails to explain the reasons for its decision”). \textit{Contrast} Creech v. UNUM Life Ins. Co. of North America, 162 F. App’x 445, 457 (6th Cir. 2006) (finding Unum’s conclusions regarding Creech’s pain and limitations to be reasonable despite the presence of conflicting medical opinions in the file because Unum pointed to specific flaws in the opinions and “made sufficiently detailed, cogent criticisms of the contrary opinions and findings”).

\textsuperscript{163} See, e.g., Bennett, 514 F.3d at 555 (concluding that the insurer did not adequately explain why the reviewers reached decisions contrary to the medical evidence presented).

\textsuperscript{164} See, e.g., \textit{id.} (questioning why the insurer implied that Bennett was not credible, when “no one who actually examined Bennett reached that conclusion”).

\textsuperscript{165} See, e.g., Mikrut v. Unum Life Ins. Co. of Am., No. 3:03cv1714 (SRU), 2006 U.S. Dist. LEXIS 92265, at *23–24 (D. Conn. Dec. 20, 2006) (observing that Unum “rejected the evaluations of Mikrut’s treating physicians in favor of the opinions of consultants who reviewed Mikrut’s file but never met or examined her”).

Courts have also questioned the propriety of a fiduciary’s placing sole reliance upon reviews performed by its own employees or consultants it retained. See, e.g., Elliott v. Metropolitan Life Ins. Co., 473 F.3d 613, 620 (6th Cir. 2006) (noting that “physicians repeatedly retained by benefits plans may have an incentive to make a finding of ‘not disabled’ in order to save their employers money and to preserve their own consulting arrangements”).

\textsuperscript{166} See, e.g., Helms v. Gen. Dynamics Corp., 222 F. App’x 821, 832–33 (11th Cir. 2007).
irregularities by the fiduciary, such as an unexplained reversal of position, as a factor in finding an abuse of discretion. While declaring that there was no per se requirement that the fiduciary’s decision needed to be consistent with findings by the Social Security Administration, a number of courts faulted fiduciaries for not bothering to explain why their decisions were incompatible with the agency’s findings. In particular, where an insurer first had urged a claimant to apply for Social Security benefits and had offset its own payments with the Social Security award, the insurer’s later position contrary to the Social Security Administration’s was suspect. As a result, the post-Firestone cases give a strong indication that it

---

167 See, e.g., Abatie v. Alta Health & Life Ins. Co., 458 F.3d 955, 974 (9th Cir. 2006) (noting procedural irregularities by Alta when it “tack[ed] on a new reason for denying benefits in a final decision”); Jones v. Aetna U.S. Healthcare, 136 F. Supp. 2d 1122, 1135 (C.D. Cal. 2001) (finding that Aetna’s failure to abide by ERISA’s disclosure requirements was an additional factor demonstrating that Aetna’s decision was tainted by conflict and in breach of its fiduciary duties).

168 See, e.g., Houston v. Unum Life Ins. Co. of Am., 246 F. App’x 293, 301 (6th Cir. 2007) (finding “Unum’s abrupt, inexplicable reversal of its own conclusion that Houston was suffering from post-breast cancer edema in her arm is per se unreasoned and unprincipled”); Brown v. Blue Cross & Blue Shield of Ala., Inc., 898 F.2d 1556, 1569 (11th Cir. 1990) (concluding that Blue Cross’ change in position highlighted its conflict of interest and called into question its motives in denying benefits).

169 See, e.g., Bennett v. Kemper Nat’l Servs. Inc., 514 F.3d 547, 547 (6th Cir. 2008) (declaring that an administrator’s decision cannot be considered arbitrary solely because the Social Security Administration reached a different decision); Calvert v. Firstar Fin., Inc., 409 F.3d 286, 295 (6th Cir. 2005) (stating that the Social Security Administration’s disability determination, standing alone, does not render an insurer’s contrary determination arbitrary and capricious).

170 See, e.g., Bennett, 514 F.3d at 554 (concluding that the insurer’s silence as to the Social Security Administration’s disability determination supported a finding that the insurer’s denial of benefits was arbitrary); Force v. Ameritech, 250 F. App’x 662, 670 (6th Cir. 2007) (noting Ameritech’s failure to consider the Social Security Administration’s finding that Force was disabled); Calvert, 409 F.3d at 296 (faulting the reviewing physician for failing to mention the Social Security Administration’s disability determination at all); Darland v. Fortis Benefits Ins. Co., 317 F.3d 516, 530 (6th Cir. 2003) (finding it inconsistent for Fortis to ignore the Social Security Administration’s determination that Darland was disabled); see also Leffew v. Ford Motor Co., 258 F. App’x 772, 779 (6th Cir. 2007) (stating that “the SSA determination, though certainly not binding, is far from meaningless”).

171 See, e.g., Mikrut v. Unum Life Ins. Co. of Am., 2006 U.S. Dist. LEXIS 92265, at *27 (D. Conn. Dec. 20, 2006) (faulting Unum’s decision for using the SSA award of disability benefits to demand a return of funds but failing to factor the award into its analysis of Mikrut’s claim); Calvert, 409 F.3d at 294–95 (stating that “a decision by a plan administrator to seek and embrace an SSA determination for its own benefit, and then ignore or discount it later, ‘casts additional doubt on the adequacy of their evaluation of . . . [a] claim’” (citation omitted)); Ladd v. ITT Corp., 148 F.3d 753, 756 (7th Cir. 1998) (finding the denial of disability benefits arbitrary where “the defendants encouraged and supported her effort to demonstrate total disability to the Social Security Administration, going so far as to provide her with legal representation. To further lighten that cost, it then
was the factual inconsistencies between the fiduciary’s decision and the substantive proof in the record that served as the basis for the arbitrary and capricious finding.\textsuperscript{172} Not only was the fiduciary’s conflict of interest not the driving force behind these decisions, but the conflict had little apparent impact on the results.\textsuperscript{173}

Even in those post-	extit{Firestone} cases where the court found that the fiduciary’s conflict of interest actually had tainted its decision, the conflict itself was not the basis for the courts’ finding of abuse of discretion. Instead, it is clear that in this scenario, too, the inconsistencies between the fiduciary’s decision and the proof in the record drove the decision, not the conflict of interest.\textsuperscript{174}

\textsuperscript{172} See, e.g., Leffew, 258 F. App’x at 781 (reversing the denial of disability benefits on the basis that the court was “unable to discern a reasoned basis for the denial or substantial evidence to support it”); McGraw v. Prudential Ins. Co. of Am., 137 F.3d 1253, 1259–60 (10th Cir. 1998) (noting that Dr. Shook, Prudential’s medical director, had acknowledged that before making the decision, he did not review Ms. McGraw’s medical records, did not talk to her neurologist, did not examine Ms. McGraw, and did not read any medical literature on her condition).

\textsuperscript{173} See, e.g., Elliott v. Metro. Life Ins. Co., 473 F.3d 613, 621 (6th Cir. 2006) (stating that even if the conflict had been shown to have impacted MetLife’s decision, it merely “would lend greater weight to the conclusion that MetLife’s determination was arbitrary and capricious”).

\textsuperscript{174} See, e.g., Evans v. Unum Provident Corp., 434 F.3d 866, 879–80 (6th Cir. 2006) (finding that the insurer’s decision to terminate the claimant’s benefits was arbitrary and capricious, and that its conflict of interest “unduly influenced its evaluation of plaintiff’s claim,” on the grounds that the insurer never sought independent medical review of the claimant, ignored reliable medical evidence proffered by the claimant, and based its decision on erroneous findings of fact); Lemaire v. Hartford Life & Accident Ins. Co., 69 F. App’x 88, 92 (3d Cir. 2003) (finding that “Hartford’s conflict of interest drove its decision-making,” but that the decision was arbitrary and capricious because Hartford’s determination “lacked both logic and support in the record”); McGraw, 137 F.3d at 1259–60, 1263 (finding that although Prudential’s medical director “understood every ‘medical’ decision would impact Prudential’s profitability,” Prudential denial of coverage was arbitrary because it did not review the claimant’s medical records, did not talk to her neurologist, did not examine the claimant, did not read any medical literature, and adopted an interpretation of the plan that alters its scope); Jones v. Aetna U.S. Healthcare, 136 F. Supp. 2d 1122, 1134–35 (C.D. Cal. 2001) (finding that Aetna’s conflict of interest caused a breach of its fiduciary duties, and that it did not act reasonably in denying benefits to
For example, in *Bedrick v. Travelers Insurance Co.*, Travelers, which provided coverage to members of an ERISA health benefit plan, limited benefits for physical and occupational therapy and denied claims for durable medical equipment to a fourteen-month-old beneficiary with cerebral palsy and spastic quadriplegia on the ground that the child’s “potential for progress is mild.” Ample evidence in the file supported the court’s finding that the claimants had not received a “full and fair review.” The physician reviewer at Travelers’ Home Office had noted that “the field office ‘needs another opinion and support of Home Office.’” He had conducted his review without supplementing or updating the file. He had characterized his job as providing “support for the legal department” and “support work for medical directors in the field who have problem cases.” The physician had admitted that he was not familiar with textbooks or treatises on cerebral palsy, and that his opinion had been based on a single medical journal article. The requirement imposed by Travelers that the patient make “significant progress” in order to receive continued benefits was not in the plan. Travelers’ utilization review agent who had made the initial determination to limit benefits had little experience with cerebral palsy patients. When asked to justify her decision to limit benefits contrary to the recommendations of the claimant’s treating physicians, she stated: “I have no idea. I have not examined the patient whether it is appropriate or inappropriate. But that isn’t a decision I was asked to make.”

Based upon the above evidence, the court concluded that the utilization review agent “at least ‘unconsciously’ had put the financial interest of Travelers above her fiduciary duty to [the child]” and that Travelers had breached its fiduciary duty of undivided loyalty to the claimant. Thus, it was the numerous and compelling inconsistencies between the fiduciary’s decision and the proof in the record, not the fiduciary’s proven conflict of interest, that caused the court to find an abuse of discretion.

It is even more obvious that conflicts did not affect the results in the post-*Firestone* cases where the courts affirmed fiduciaries’ denials of benefits despite

---

175  93 F.3d 149 (4th Cir. 1996).
176  Id. at 151.
177  Id. at 153.
178  Id. (citation omitted).
179  Id.
180  Id.
181  Id. at 154.
182  Id. at 153.
183  Id.
184  Id. at 154.
185  Id.
the conflicts. Where a fiduciary’s denial of benefits was substantially supported by the record, the fact that the fiduciary had operated under a conflict of interest was insufficient to alter the courts’ conclusions that the fiduciary had not abused its discretion. Therefore, these cases also show that the existence of a conflict of interest had no effect on the outcome of the judicial review.

C. The Fiduciary’s Conflict of Interest Does Not Act as a Tiebreaker in Close Cases

Mindful that it would be foolish to adopt a conflict standard with no real impact on how the cases would be decided, the Supreme Court in MetLife v. Glenn pressed counsel at oral argument to explain precisely how weighing the conflict as a factor could change the outcome. The Court fully recognized that the standard would be universally applied to discretionary denials of ERISA plan benefits, as “the potentiality of conflict is inherent in all of these commonplace dual arrangements in ERISA welfare benefit plans.” When counsel for MetLife argued that the conflict “must be weighed, but it doesn’t change the abuse of

---

186 See, e.g., Wade v. Hewlett-Packard Dev. Co. LP Short Term Disability Plan, 493 F.3d 533, 541 (5th Cir. 2007) (stating that “[e]ven taking into account any alleged conflict of interest of the Administrator, we affirm the district court’s grant of summary judgment in favor of the defendant-appellee” because “[t]here is substantial evidence in the record to support the Administrator’s decision to deny benefits”).

187 See, e.g., id. at 536–37 (finding that, despite the administrator’s conflict, an opinion of a consulting neurophysiologist that claimant’s condition did not constitute a disability, an opinion by a psychiatrist employed by a disability care management service company, and a review of all the documentation by another psychiatrist constituted substantial evidence in the record to support the administrator’s decision to deny benefits); Creech v. Unum Life Ins. Co. of N. Am., 162 F. App’x 445, 457 (6th Cir. 2006) (finding that “even when we consider the conflict of interest inherent in Unum’s dual role, we conclude that its assessment of Creech’s residual functional capacity was not arbitrary and capricious”); Allison v. Unum Life Ins. Co. of Am., 381 F.3d 1015, 1024 (10th Cir. 2004) (holding that, despite a conflict, Unum’s requests for additional documentation were reasonable, and that the claimant’s repeated failure to respond to the requests resulted in the denial of her claim); Williams v. BellSouth Telecomms., Inc., 373 F.3d 1132, 1139 (11th Cir. 2004) (holding that, despite a conflict, the plan’s denial of benefits was supported by “more than sufficient medical evidence to contradict [claimant’s] claim,” including the reports of several doctors, and even the claimant’s own doctor, who did not say that claimant was incapable of working, and claimant’s remarks to an independent medical examiner that she was caring for two young children and a granddaughter, cooking all meals, performing housework, tending to finances, and attending religious services, upon which the examiner concluded that claimant’s stress was “not overwhelming her capacity for coping”).


189 Id. at 3; see also Langbein, supra note 16, at 1321 (stating that “conflicted plan decisionmaking is a structural feature of ERISA plan administration”).
Chief Justice Roberts asked frankly: “How does that work? I don’t understand.”190 Apparently not satisfied with counsel’s responses, Chief Justice Roberts asked again, “How is the review different in each of those cases as a practical matter?”191 and still again, “How does the review differ as a functional matter?”192 Justice Scalia followed with the key question, “What was a reasoned decision for someone who doesn’t have a conflict becomes an unreasoned decision simply by reason of the fact that he has a conflict?”193

The post-Firestone cases themselves provide a response to Justice Scalia’s question—a conflict never serves to convert a reasonable fiduciary decision into an arbitrary and capricious one.

Virtually all of the post-Firestone courts endorsed an approach of decreasing deference to the fiduciary’s decision in proportion to the seriousness of the conflict. Thus, in cases where it appeared that the fiduciary’s decision was clearly supported by the evidence in the record, the court assigned little weight to the conflict and accorded nearly full deference to the fiduciary.194 In these cases, since the conflict had little weight, it was understandably incapable of causing a reversal of the fiduciary’s determination, which typically was affirmed on the ground that it was supported by substantial evidence.195 Conversely, where it appeared that the fiduciary’s decision was inconsistent with the evidence in the record, the courts

191 Id.
192 Id.
193 Id. at 10. Justice Scalia also stated that “since this fiduciary has a conflict, what was a reasonable decision, whoop, the added weight, it becomes an unreasonable decision.” Id. at 13. Referring to the “weighing the conflict as a factor” standard, he stated “it’s beautiful art, but I don’t understand how it turns into law.” Id. at 46. Other justices articulated the same confusion. Justice Souter asked, “Being on the fence, may I take into consideration the fact that there was a conflict of interest?” Id. at 14. Justice Ginsberg repeated the concern: “I am in equipoise about this case, not the terms of the plan but the decision that was made to deny her benefits. So which way do I call it? That’s the question that Justice Souter posed, and you seem not to want to face up to it and answer it.” Id. at 19.
194 See, e.g., Lasser v. Reliance Standard Life Ins. Co., 344 F.3d 381, 385 (3d Cir. 2003) (finding no evidence that the fiduciary’s decision had been affected by conflict, and therefore reviewing the decision “at the mild end of the heightened arbitrary and capricious scale” affording a “moderate degree of deference”).
195 See, e.g., Bendixen v. Standard Ins. Co., 185 F.3d 939, 944 (9th Cir. 1999) (holding that because evidence supported the insurer’s denial of benefits, “we find that Standard’s rationale for denying benefits was not born out of the apparent conflict of interest”); Godfrey v. BellSouth Telecomms., Inc., 89 F.3d 755, 758–59 (11th Cir. 1996) (concluding that the fiduciary’s determination was arbitrary and capricious because it rejected clear medical evidence without examining the claimant itself or seeking treatment notes from her doctors, and ignored the effects of the medication that the claimant was required to take due to her condition).
gave more weight to the conflict and decreased deference. But in these cases the seriousness of the conflict was not the determining factor, as there was already sufficient basis to reverse the fiduciary’s decision on the ground that it was inconsistent with the record. Thus, the discretionary assignment of weight to the conflict actually guaranteed its inability to influence the outcome of judicial review in most cases.

It is only the cases that fall in the middle of these two extremes that potentially could be affected by the conflict, i.e., those cases where there is both substantial evidence in favor of awarding benefits to the claimant and substantial evidence in favor of denying benefits. Arguably, the fiduciary’s decision in such cases would always be reasonable because either an award of benefits or denial of benefits would be supported by substantial evidence. Here, Justice Scalia’s inquiry was key: can the conflict ever convert an otherwise reasonable decision into an arbitrary and capricious one?

It is instructive that while several post-Firestone courts declared that weighing the conflict as a factor could convert a reasonable fiduciary decision into an unreasonable one, it does not appear any such cases actually occurred.

196 See, e.g., Post v. Hartford Ins. Co., 501 F.3d 154, 168 (3d Cir. 2007) (ruling that a number of irregularities by the fiduciary warranted review under a “further heightened” standard); Pinto v. Reliance Standard Life Ins. Co., 214 F.3d 377, 394 (3d Cir. 2000) (finding that due to “procedural anomalies,” the court’s review would be at the “far end of the arbitrary and capricious ‘range,’ and we examine the facts before the administrator with a high degree of skepticism”).

197 See cases cited supra Part V.B and accompanying notes.

198 The “two extremes” referred to are those fiduciary decisions that are clearly supported by the record, and those fiduciary decisions that are clearly inconsistent with the record.

199 The term “substantial evidence” has been defined as “more than a scintilla but less than a preponderance,” Schatz v. Mutual of Omaha Ins. Co., 220 F.3d 944, 949 (8th Cir. 2000), and evidence that “a reasonable mind might accept as adequate to support” the decision, Siebert v. Standard Ins. Co. Group Long-Term Disability Policy, 220 F. Supp. 2d 1128, 1141 (C.D. Cal. 2002).

200 See Transcript of Oral Argument at 10, Metro. Life Ins. Co. v. Glenn, 128 S. Ct. 2343 (2008) (No. 06-923) (asking whether “a reasoned decision for someone who doesn’t have a conflict becomes an unreasoned decision simply by reason of the fact that he has a conflict?”).

201 See, e.g., Whitney v. Empire Blue Cross & Blue Shield, 106 F.3d 475, 477 (2d Cir. 1997) (stating that “a reasonable interpretation of the Plan will stand unless the participants can show not only that a potential conflict of interest exists, but that the conflict affected the reasonableness of the Committee’s decision”) (citation omitted); Godfrey v. BellSouth Telecommuns., Inc., 89 F.3d 755, 758 (11th Cir. 1996) (stating that a conflict could render a fiduciary’s “apparently reasonable interpretation” arbitrary and capricious).

202 See, e.g., Creech v. Unum Life Ins. Co., 162 F. App’x 445, 455–58 (6th Cir. 2006) (finding that although Unum was a conflicted administrator of an ERISA disability plan, Unum’s assessment of Creech’s residual functional capacity was not arbitrary and capricious where the medical opinions in the record were themselves inconsistent).
Even the post-\textit{MetLife} cases tend to bear out that in close cases, where the conflict conceivably could be useful to tip the balance, the courts have resisted using the conflict as a tiebreaker.\footnote{\textit{See}, e.g., Dunn v. GE Group Life Assurance Co., 289 F. App’x 778, 782 (5th Cir. 2008) (finding that even where “two equally rational interpretations of the Plan language are involved in this dispute,” the insurer’s “minimal” conflict of interest was insufficient to disturb its denial of benefits); Doyle v. Liberty Life Assurance Co., 542 F.3d 1352, 1363 (11th Cir. 2008) (“Because the evidence is close, we cannot say, even accounting for the conflict, that Liberty Life abused its discretion in denying Doyle benefits.”); \textit{see also} Feigenbaum v. Merrill Lynch & Co., 308 F. App’x 585, 587 (3d Cir. 2009) (stating that “there is not a sufficiently close balance for the conflict of interest to act as a tiebreaker in favor of finding that [MetLife] abused its discretion”); Smith v. Health Servs. of Coshocton, 314 F. App’x 848, 861–62 (6th Cir. 2009) (ruling in the insurer’s favor despite its conflict of interest because “the other factors are not closely balanced in this case”); Wakkinen v. UNUM Life Ins. Co., 531 F.3d 575, 582 (8th Cir. 2008) (concluding that “there is not a sufficiently close balance for the conflict of interest to act as a tiebreaker in favor of finding that UNUM abused its discretion”).}

Justice Scalia well recognized the ineffectuality of the \textit{MetLife} conflict standard to change the outcome in any particular case. At oral argument, Justice Scalia was openly skeptical that consideration of the conflict could render a reasonable decision unreasonable. He said, “Now, let’s take a perfectly reasonable decision. It is not arbitrary. . . . [D]oes giving weight to [a conflict] mean that what was reasonable becomes unreasonable? . . . If it doesn’t mean that, then it means nothing.”\footnote{Transcript of Oral Argument at 28, \textit{Metro. Life Ins. Co. v. Glenn}, 128 S. Ct. 2343 (2008) (No. 06-923).}

In his dissent, Justice Scalia declared outright:

\begin{quote}
Common sense confirms that a trustee’s conflict of interest is irrelevant to determining the substantive reasonableness of his decision. A reasonable decision is reasonable whether or not the person who makes it has a conflict. It makes no sense to say that a lurking conflict of interest . . . can make a reasonable decision unreasonable, or a well-thought-out, informed decision uninformed or arbitrary.\footnote{\textit{Id.}}
\end{quote}

Where the fiduciary’s decision comports with the proof in the record, Justice Scalia pointed out that it is unlikely that the fiduciary was influenced by the inherent conflict. He stated, “[I]f one is to draw any inference about a fiduciary from the fact that he made an informed, reasonable, though apparently self-serving discretionary decision, it should be that he \textit{suppressed} his selfish interest . . . in compliance with his duties of good faith and loyalty.”\footnote{\textit{Id.}} Accordingly, Justice...
Scalia concluded that the conflict should not be considered at all in determining whether MetLife abused its discretion.207

D. Weighing the Conflict as a Factor Is Indistinguishable from De Novo Review

Justice Scalia criticized the MetLife conflict standard for inviting de novo review of a fiduciary’s discretionary decision:

[T]he Court seems to advance a gestalt reasonableness standard (a “combination-of-factors method of review,” the opinion calls it, . . .) by which a reviewing court, mindful of being deferential, should nonetheless consider all the circumstances, weigh them as it thinks best, then divine whether a fiduciary’s discretionary decision should be overturned.208

Gradating reasonableness, and making it a “factor” in the improper-motive determination would have the precise effect of eliminating the discretion that the settlor has intentionally conferred upon the trustee with a conflict, for such a trustee would be foreclosed from making an otherwise reasonable decision.209

As a result, Justice Scalia characterized the majority’s conflict standard as “nothing but de novo review in sheep’s clothing.”210 Pre-MetLife courts, too, had noted the close similarity between the heightened arbitrary and capricious standard and de novo review.211

An examination of the MetLife case itself shows how close to the mark Justice Scalia was in observing that weighing the conflict as a factor is indistinguishable

---

207 Id. at 2361 (recommending a remand to the Court of Appeals to allow it to determine whether MetLife’s determination was reasonable “without regard to the existence of a conflict of interest”). Justice Robert concurred. See id. at 2355 (Roberts, C.J., concurring) (stating that “[t]hese facts [of inconsistencies between MetLife’s denial and proof in the record] simply prove that MetLife abused its discretion in failing to consider relevant, expert evidence on the question of Glenn’s disability status. There is no basis for supposing that the conflict of interest lent any greater significance to these factors, and no logical reason to give the factors an extra dollop of weight because of the structural conflict”); see also Fought v. Unum Life Ins. Co., 379 F.3d 997, 1007–08 (10th Cir. 2004) (declaring that “[t]he guiding inquiry must be whether the plan administrator’s decision was objectively reasonable given the administrative record”).


209 Id. at 2360.

210 Id. at 2358. Chief Justice Roberts agreed on this point. See id. at 2353 (Roberts, C.J., concurring) (commenting that weighing the conflict as a factor “invites the substitution of judicial discretion for the discretion of the plan administrator”).

211 See, e.g., Williams v. BellSouth Telecomms., Inc., 373 F.3d 1132, 1137 (11th Cir. 2004) (noting that “the distinctions between the heightened arbitrary and capricious, arbitrary and capricious, and de novo standards of review have become difficult to discern over time”).
from de novo review. In *MetLife*, there was substantial evidence tending to show that MetLife’s denial of benefits was arbitrary. Both the Sixth Circuit Court of Appeals and the Supreme Court cited a number of factors showing arbitrary decision making by MetLife. MetLife initially encouraged Glenn to apply for Social Security benefits, reaped financial benefits when the award was granted, and then adopted a contrary view of Glenn’s disability that allowed MetLife to avoid paying continued benefits.\(^{212}\) MetLife ignored numerous statements from Glenn’s physician, Dr. Patel, that Glenn could not work.\(^{213}\) MetLife dismissed stress as a factor in Glenn’s disability, despite Dr. Patel’s repeated statements that stress was a critical factor.\(^{214}\) MetLife instead relied upon a report by its own consultant, Dr. Pujara, even though Dr. Pujara’s opinion was “at best, ambiguous,”\(^{215}\) and was based upon a review of Glenn’s medical file rather than a medical examination.\(^{216}\)

However, the district court, in granting judgment to MetLife, had noted a number of factors supporting MetLife’s denial of benefits.\(^{217}\) After receiving short-term disability benefits from MetLife, Glenn experienced an improved condition as a result of treatment.\(^{218}\) In 2000, Dr. Patel described her as “clinically well” and “stable.”\(^{219}\) The percentage measurement of blood being pumped out of her heart improved from 25 percent in 2000 to her pre-disability level of 40 to 50 percent in 2001.\(^{220}\) In 2002, Dr. Patel completed a physical capacity evaluation in which he

\(^{212}\) Glenn v. MetLife, 461 F.3d 660, 667 (6th Cir. 2006) (finding that MetLife “failed to address Social Security’s contrary determination of Glenn’s status”).

\(^{213}\) Specifically, in April 2000, Dr. Patel wrote that Glenn “cannot return to any kind of job that would require any significant physical or psychological stress.” *Id.* at 672. In June 2002, Dr. Patel prepared a progress report that stated, “I do not believe she will handle any kind of stress well at her work and she would be better off being on disability.” *Id.* at 670. In July 2002, Dr. Patel wrote:

At the present time, I do not believe Wanda should be forced to return to any kind of even sedentary work particularly because it is the psychologic stress of work that really exacerbates her cardiovascular condition and symptomology. The patient basically should be considered completely disabled from her dilated cardiomyopathy as well as history of ventricular tachycardia.

*Id.* at 664. In February 2003, Dr. Patel submitted a letter to MetLife stating that “there was never a time where I felt that this patient would be able to return to full-time employment.” *Id.* at 665. Dr. Patel reiterated his position that “she should be considered completely disabled.” *Id.*

\(^{214}\) *Id.* at 672.

\(^{215}\) *Id.* at 665.

\(^{216}\) *Id.* at 671.


\(^{218}\) *Id.* at *4.

\(^{219}\) *Id.*

\(^{220}\) *Id.*
indicated that in an eight-hour workday, Glenn could sit eight hours, stand four hours, and walk two hours. Dr. Patel checked “yes.” Three months later, Dr. Patel completed another physical capacity evaluation in which he again indicated that Glenn could sit eight hours, stand four hours, and walk two to four hours. He also indicated that Glenn could lift and carry up to ten pounds occasionally, and engage in other activities. In 2002, Dr. Moyer, an independent physician consultant who reviewed Glenn’s file, concurred with Dr. Patel that Glenn could perform sedentary work. A vocational coordinator concluded there were alternative sedentary occupations available in Glenn’s geographic area that Glenn would be able to perform. In 2003, Dr. Pujara, an independent cardiology consultant, reported that Glenn had “achieved a relatively stable cardiac status,” and that she was “capable of sedentary activity.” Thus, there not only appears to be substantial evidence supporting MetLife’s denial of benefits, but the district court, arguably an objective and reasonable fact finder, found the facts sufficient to sustain MetLife’s decision.

Therefore, it is somewhat surprising that the Sixth Circuit reversed the district court’s ruling when applying an arbitrary and capricious standard to MetLife’s decision. The arbitrary and capricious standard “is the least demanding form of judicial review.” All that is required to affirm a fiduciary’s decision under the standard is that it is “possible to offer a reasoned explanation, based on the evidence, for a particular outcome.” The fiduciary’s decision will not pass

---

221 Id. at *5.
222 Id.
223 Id.
224 Id.
225 Id.
226 Id. at *6.
227 Id. at *8–9.
228 Id. at *21–22 (“The Court finds that the above constitutes substantial evidence supporting MetLife’s determination that plaintiff was no longer totally disabled as the Plan defines that term. Even considering MetLife’s conflict of interest, the Court holds that MetLife’s decision to terminate plaintiff’s LTD benefits was rational and did not constitute an abuse of discretion.”).
229 Henderson v. Ameritech Corp., No. 00-1686, 2001 WL 1823813, at *6 (E.D. Mich. Dec. 20, 2001). The Unum scandal, in which various states discovered systemic irregularities in Unum’s evaluation of ERISA disability claims, demonstrates the extent to which the deferential standard of review can facilitate extremely questionable or even outright improper denials of claims. See Langbein, supra note 16, at 1317–21 (discussing the scandal).
230 Henderson, 2001 WL 1823813, at *6; see also Creech v. Unum Life Ins. Co., 162 F. App’x 445, 447 (6th Cir. 2006) (stating that a denial of benefits will not be considered arbitrary “so long as it is possible to offer a reasoned explanation, based on the evidence,” for the denial); Fought v. Unum Life Ins. Co., 379 F.3d 997, 1008 (10th Cir. 2004) (stating that an arbitrary and capricious decision is “objectively reasonable given the administrative record”); Schatz v. Mutual of Omaha Ins. Co., 220 F.3d 944, 949 (8th Cir. 2000) (stating
muster only if it is “whimsical, random, or unreasoned.”

Indeed, the standard is so easily satisfied that Judge Posner has described it as any determination that is not “off the wall.”

Moreover, the reversal was not due to the district court’s use of an improper standard of review. Both the district court and the Sixth Circuit recognized that MetLife had a conflict of interest, and both courts applied the arbitrary and capricious standard, weighing the conflict as a factor in determining whether MetLife had abused its discretion. Rather, the basis for the Sixth Circuit’s reversal was its finding that the district court had failed to assign sufficient weight to the conflict:

[T]he court’s analysis of the plan administrator’s basis for terminating benefits does not include any discussion of the role that MetLife’s conflict of interest may have played in its decision nor appear

that the decision to deny benefits must be supported by substantial evidence, “meaning more than a scintilla but less than a preponderance”).

 Judges, supra note 16, at 1324.

Some courts, however, have stated that deferential review is more demanding. See, e.g., Gilchrest v. Unum Life Ins. Co., 255 F. App’x 38, 42 (6th Cir. 2007) (stating that to pass muster under deferential review, the Unum’s decision “must be consistent with the quantity and quality of the evidence in the record”); Evans v. Unum Provident Corp., 434 F.3d 866, 876 (6th Cir. 2006) (commenting that the arbitrary and capricious standard “is not, however, without some teeth”); Creech v. UNUM Life Ins. Co. of North America, 162 F. App’x 445, 448 (6th Cir. 2006) (declaring that “we will not rubber-stamp the administrator’s decision”); Moon v. Unum Provident Corp., 405 F.3d 373, 379 (6th Cir. 2005) (declaring that “deferential review is not no review, and deference need not be abject”); Vega v. Nat’l Life Ins. Servs., Inc., 188 F.3d 287, 297 (5th Cir. 1999) (“The essential requirement of reasonableness has substantial bite . . . .”); see also Fought, 379 F.3d at 1015 (finding UNUM’s inability to offer “more than a scintilla” of evidence in support of its denial of benefits to be unreasonable under deferential review); Adamson v. Wadley Health Sys., Civ. No. 07-1081, 2008 WL 4649084, at *5 (W.D. Ark. Oct. 20, 2008) (stating that although courts are “hesitant to interfere” with the administrative decision,” courts evaluate “[b]oth the quantity and the quality of the evidence” under the deferential standard).

Glenn v. MetLife, 461 F.3d 660, 666 (6th Cir. 2006) (“MetLife is authorized both to decide whether an employee is eligible for benefits and to pay those benefits. This dual function creates an apparent conflict of interest.”); Glenn v. Metro. Life Ins. Co., No. 2:2003cv0572, 2005 U.S. Dist. LEXIS 11007, at *12 (S.D. Ohio June 8, 2005) (finding that MetLife operated under a conflict of interest because it both “decide[d] whether a claimant is eligible for benefits, and pays those benefits”).

Glenn v. MetLife, 461 F.3d at 666 (“[W]e are entitled to take into account the existence of a conflict of interest . . . .”); Glenn v. Metro. Life Ins. Co., 2005 U.S. Dist. LEXIS 11007, at *12 (stating “that conflict must be weighed as a ‘factor’ in determining whether there is an abuse of discretion’”) (citation omitted).
to give that conflict any weight. It appears to us, as a result, that this factor did not receive appropriate consideration by the district court.\textsuperscript{235}

But inexplicably, the Sixth Circuit’s reversal similarly “does not include any discussion of the role that MetLife’s conflict of interest may have played in its decision.”\textsuperscript{236} Rather than pointing to any concrete evidence showing that the conflict had tainted MetLife’s decision, the Sixth Circuit simply reconsidered the factual evidence and adopted an alternate view. Specifically, the Sixth Circuit stated that its reversal was based upon the following substantive grounds:

1. MetLife’s denial conflicted with the finding by the Social Security Administration;
2. MetLife offered no explanation for crediting a brief form filled out by Dr. Patel while overlooking his more detailed reports;
3. The occupational skills analyst and independent medical consultant were not provided with all of the information from Dr. Patel; and
4. MetLife improperly ignored the role that stress played in aggravating Glenn’s condition.\textsuperscript{237}

Thus, it was the Sixth Circuit’s disagreement with the district court’s evaluation of the facts that drove the reversal, not MetLife’s conflict of interest. Other than simply declaring that “MetLife acted under a conflict of interest,”\textsuperscript{238} the circuit court neither specified the correct weight to be assigned to the conflict nor explained how weighing the conflict had altered the outcome.

Judging from the Sixth Circuit’s opinion in \textit{MetLife}, it appears that weighing the conflict as a factor simply allows the reviewing court to engage in a de novo review of the facts and to give no deference to the fiduciary’s decision. Even when the fiduciary’s decision is supported by substantial evidence, as it arguably was in \textit{MetLife}, the court can reverse it merely by declaring that the conflict was entitled to great, or greater, weight. As Justice Scalia observed, “[T]he notion that there are degrees of deference is absurd. . . . [T]he court either defers, or it does not. ‘Some deference,’ or ‘less than total deference,’ is no deference at all.”\textsuperscript{239}

Justice Scalia fully understood that weighing the conflict as a factor in arbitrary and capricious review is “nothing but \textit{de novo} review in sheep’s clothing.”\textsuperscript{240}

\textsuperscript{235} Glenn v. MetLife, 461 F.3d at 666.
\textsuperscript{236} \textit{Id.}
\textsuperscript{237} \textit{Id.} at 674 (stating “these factors reflect a decision by MetLife that can only be described as arbitrary and capricious”).
\textsuperscript{238} \textit{Id.}
\textsuperscript{240} \textit{Id.} at 2358 (Scalia, J., dissenting); see also Rud v. Liberty Life Assurance Co., 438 F.3d 772, 774 (7th Cir. 2006) (noting cases that “say not that the conflict alters the standard of review, but that it affects how the standard is applied; but that comes to the
VI. Conclusion

Professor Langbein, a noted ERISA scholar, has argued convincingly in favor of applying de novo review to ERISA benefit denials where the fiduciary acted under a conflict of interest. While the Supreme Court in *MetLife* rejected the de novo standard in favor of retaining *Firestone*’s arbitrary and capricious standard coupled with weighing the conflict as a factor, application of the *MetLife* standard yields a result that is so close to de novo review as to be virtually indistinguishable. By allowing the reviewing court to assign a discretionary weight to the conflict, it is uncertain how much, if any, deference will be given to the fiduciary’s decision. This discretionary weighing of the conflict thereby allows the reviewing court to overturn any fiduciary decision on the stated basis that the conflict warrants greater weight in the review.

Although the *MetLife* standard certainly moves us closer to the de novo approach endorsed by Professor Langbein, it is still questionable whether it achieves the “undivided loyalty” that ERISA promises to plan beneficiaries. Even where a fiduciary’s denial of benefits is reasonable and supported by substantial evidence, there is no way to determine with certainty that self-interest was not one of the bases for the denial. Moreover, it is questionable whether health and disability beneficiaries are receiving their promised undivided loyalty. Since most health and disability claims decisions are rendered by insurers acting as either the insurance issuers or the third-party administrators of ERISA plans, these insurers are subject to the fiduciary duty of loyalty whenever they process ERISA claims. This fiduciary duty of loyalty, however, is not required for the insurers’ non-ERISA business.

---

241 See Langbein, supra note 16, at 1342 (opining that “impartial judicial review in [ERISA conflicts cases] is an essential safeguard against self-serving conduct”).

242 See, e.g., Glenn v. MetLife, 461 F.3d 660, 666 (reversing the district court’s decision on the basis that the fiduciary’s conflict of interest “did not receive appropriate consideration by the district court”); see also discussion supra at Part IV.D.

243 See Bedrick v. Travelers Ins. Co., 93 F.3d 149, 154 (4th Cir. 1996) (stating that ERISA’s command that fiduciaries act “free” of conflict “is an absolute. There is no balancing of interests”).

244 See, e.g., Fought v. Unum Life Ins. Co., 379 F.3d 997, 1004 (10th Cir. 2004) (stating that a conflicted fiduciary may unconsciously favor its interests over the interests of the plan beneficiaries); Brown v. Blue Cross & Blue Shield, Inc., 898 F.2d 1556, 1565 (11th Cir. 1990) (also stating that a conflicted fiduciary may unconsciously favor its interests over the interests of the plan beneficiaries).

difference in duty would arguably require insurers to implement ERISA claims processing machinery and procedures, separate from their non-ERISA business, to deliver the additional duty of loyalty to ERISA claimants. However, if insurance companies process both ERISA and non-ERISA claims in precisely the same way—employing the same internal guidelines and claims processing software—we should question by what means an ERISA beneficiary receives any duty of loyalty from insurance fiduciaries over and above the standard contract duties that insurers deliver to their non-ERISA policyholders.

As there is no way to guarantee that ERISA claims are processed free of the taint of a fiduciary’s conflict of interest, the only way to ensure delivery of the undivided loyalty promised by ERISA is to prohibit conflicted entities from acting as fiduciaries altogether. In the health and disability plan context, this would prohibit claims processing by employers and insurance companies, and instead would require claims to be determined by randomly assigned independent third parties.

But we are currently struggling with a national economic crisis, with the costs of medical treatment and health insurance skyrocketing beyond the ability of many families and employers to afford. Therefore, it is extremely unlikely that Congress will implement such a colossal and costly change to ERISA claims review. As a result, for the present at least, ERISA’s promise to plan members and beneficiaries of undivided loyalty when benefits claims are determined by conflicted fiduciaries is a promise that largely cannot be kept.

---

246 See Pinto v. Reliance Standard Life Ins. Co., 214 F.3d 377, 389 (3d Cir. 2000) (observing that “direct evidence of a conflict is rarely likely to appear in any plan administrator’s decision”); Langbein, supra note 16, at 1321 (stating that “[t]he danger pervades the ERISA-plan world that a self-interested plan decisionmaker will take advantage of its license under Bruch to line its own pockets by denying meritorious claims,” and that “in the hands of subtler operators such misbehavior is much harder to detect”).

247 Some states have prohibited the use of discretionary clauses in insurance policies, thereby effectively foreclosing deferential review for insured ERISA plans. See Langbein, supra note 16, at 1340 (noting the movement among state insurance commissioners to forbid discretionary clauses in insurance policies in the wake of the Unum scandal, wherein Hawaii, Maine, and Minnesota passed statutes forbidding the clauses, and Insurance Commissioners in California, Illinois, Indiana, Montana, Nevada, New Jersey, Oregon, Texas, and Utah declared opposition to their use).

248 See Wright v. R. R. Donnelley & Sons Co. Group Benefits Plan, 402 F.3d 67, 75 (1st Cir. 2005) (noting that the district court was “troubled by what it deemed the ‘false premise’ that an ERISA plan administrator that also has a financial stake in the benefit decisions can act as a disinterested trustee”); Music v. W. Conference of Teamsters Pension Trust Fund, 712 F.2d 413, 418 (9th Cir. 1983) (observing that “it would be anomalous to conclude that [ERISA] imposed fiduciary obligations on trustees but that federal courts were powerless to enforce them”) (citation omitted).
WEB SITE PROPRIETORSHIP AND ONLINE HARASSMENT

Nancy S. Kim*

Although harassment and bullying have always existed, when such behavior is conducted online, the consequences can be uniquely devastating. The anonymity of harassers, the ease of widespread digital dissemination, and the inability to contain and/or eliminate online information can aggravate the nature of harassment on the Internet. Furthermore, section 230 of the Communications Decency Act provides Web site sponsors with immunity for content posted by others and no incentive to remove offending content.

Given the unique nature of online harassment, ex post punitive measures are inadequate to redress grievances. In this Article, I propose the imposition of proprietorship liability upon Web site sponsors who fail to adopt “reasonable measures” to prevent foreseeable harm, such as online harassment. I also introduce several proposals to deter online harassment that would qualify as reasonable measures. These proposals incorporate contractual and architectural restraints, limits on anonymity, and restrictions on posting certain types of digital images.

TABLE OF CONTENTS

I. INTRODUCTION ................................................................. 995
II. WHAT IS CYBERHARASSMENT? ...................................... 999
   A. Verbal Cyberharassment ............................................... 1001
      1. Online Threats ....................................................... 1001
      2. Online Insults ....................................................... 1002

* © 2009 Nancy S. Kim, Associate Professor, California Western School of Law; Visiting Associate Professor, Rady School of Management, UC San Diego. This paper was presented at TPRC’s 36th Research Conference on Communication, Information, and Internet Policy, held at George Mason University School of Law on September 26–28, 2008; the Conference of Asian Pacific American Law Faculty, held at Denver University School of Law on April 24–26, 2008; and a faculty workshop at the Loyola University of Chicago School of Law. I greatly benefited from the comments that were made by audience participants at these presentations. Ideas in this paper were also presented at the Law & Information Society Symposium: Intermediaries in the Information Society, Center on Law and Information Policy, Fordham Law School, held in New York, New York, on March 27, 2009; and the 22nd Annual Media and the Law Seminar, held in Kansas City, Missouri, on April 17, 2009. I would like to express my gratitude to Tom Barton, Ed Dauer, James Grimmelmann, and Cynthia Ho for their very helpful comments and suggestions on prior drafts of this paper. I would also like to express my appreciation to Janet Bowermaster, Larry Benner, Seth Burns, and Fred Yen for their helpful comments in discussions regarding this paper, to my research assistants, Melissa Henkel and Brian Suba, and to the conscientious editors at the Utah Law Review for their diligent efforts.
“Now, once you had communal fires and cooking and a higher calorie diet, the social world of our ancestors changed, too. Once individuals were drawn to a specific attractive location that had a fire, they spent a lot of time around it together. This was clearly a very different system from wandering around chimpanzee-style, sleeping wherever you wanted, always able to leave a group if there was any kind of social conflict. We had to be able to look each other in the eye. We couldn’t react with impulsivity. Once you are sitting around the fire, you need to suppress reactive emotions that would otherwise lead to social chaos. Around that fire, we became tamer.”

I. INTRODUCTION

Several female law students were the subject of malicious and obscene comments on AutoAdmit, a Web site catering to law school students. The cruel nature of the posts, the hostile, unrepentant mob mentality of the anonymous posters, and the adamant refusal of the Web site operators to remove the offensive posts shocked the legal community and attracted media attention. After repeated, unsuccessful requests to the Web site operator to remove the posts, two of the women sued, claiming that the posts caused them emotional distress and diminished their professional opportunities. As part of the litigation, the identities of several of the posters were revealed. One of the named defendants was later dropped from the case; he then sued the plaintiffs, alleging emotional distress and damage to his professional reputation.

The AutoAdmit case illustrates the damage online harassment wreaks upon both the targets of the harassment and the harassers. Although the harm to targets of online smear campaigns is obvious, what is less evident is that the harassers also risk damage to their own reputations. Online communication is often phrased as

---

1 Claudia Dreifus, A Conversation with Richard Wrangham, N.Y. TIMES, Apr. 21, 2009, at D2 (noting that Richard Wrangham is a professor of biological anthropology at Harvard University and author of CATCHING FIRE: HOW COOKING MADE US HUMAN (2009)).
3 Margolick, supra note 2.
4 Id.
5 Id.
6 Id.
7 One of the defendants in the AutoAdmit case claimed that the lawsuit has ruined his life. Id. Another defendant claims he has been unable to find a job as a result of the negative publicity surrounding the Web site and the litigation. Id.
expressive speech,8 but the nature of online discourse is often shaped by the Web site itself. For example, the founder of AutoAdmit marketed the Web site as an alternative to message boards that filtered out inflammatory posts.9 This Article adopts a new approach to the problem of online harassment or “cyberharassment”10 by treating it primarily as a failure of business norms,11 rather than as a matter of unfettered speech.

Framing online harassment as a private-sector problem resolves or reduces many of the free speech concerns raised by First Amendment advocates.12 It empowers and encourages Web site sponsors to shape developing norms as part of good business practices. Setting expectations for user conduct empowers Web site sponsors to better control their Web site image (i.e., their “brand”). To require users to conform to the law and prevailing offline social norms reinforces positive community values, and reduces the likelihood of conflicts between and amongst users regarding expectations of conduct on a particular Web site.

8 See, e.g., Eugene Volokh, Freedom of Speech and Information Privacy: The Troubling Implications of a Right to Stop People From Speaking About You, 52 STAN. L. REV. 1049, 1051 (2000) (“While privacy protection secured by contract [turns out to be] constitutionally sound, broader information privacy rules are not easily defensible under existing free speech law.”).

9 Margolick, supra note 2.

10 I use the term online harassment or “cyberharassment” to characterize the use of the Internet as a medium for disseminating harmful material about another individual. The definition is deliberately loose to accommodate different types of conduct.

11 This Article argues that the social problem of online harassment is one that should be addressed by Web site sponsors. By “Web site sponsors,” I refer to the companies and individuals who control or have the ability to control activity on the Web site. I am not referring to Internet service providers or other Web hosting companies (ISPs) that technically enable those businesses, unless those companies also sponsor or control the activity on the site (such as AOL, which provides Internet access but also sponsors its own message board). See, e.g., People Connection Blog, http://www.peopleconnectionblog.com (last visited Sept. 1, 2009) (showing that AOL sponsors its own message board).

The volume of traffic that a single ISP transports is typically many times greater than that hosted by any single Web site sponsor. See Mark A. Lemley, Rationalizing Internet Safe Harbors 12 (Stanford Public Law Working Paper, Paper No. 979836, 2007), available at http://ssrn.com/abstract=979836 (noting that it is “simply impossible for a search engine—to say nothing of an ISP or bandwidth conduit—to cull through the literally billions of links and messages they process each day and identify all those messages and Web pages that may create liability under any law”). Accordingly, different standards should apply to ISPs and Web site sponsors, and I confine my discussion in this Article to Web site sponsors.

12 This is not to suggest that government regulation is not an appropriate way to address the problem of cyberharassment, only that First Amendment issues are more relevant where the government is directly regulating conduct rather than where a private entity or industry creates and adopts its own standards. For a discussion of an Internet-based civil rights strategy, see generally Danielle Keats Citron, Cyber Civil Rights, 89 B.U. L. REV. 61 (2009) (arguing that cyberharassment harms ought to be understood and addressed as civil rights violations).
Currently, Web site sponsors exercise power over their Web sites in an inconsistent and self-serving manner. They exercise property-like control over certain aspects of the site, yet claim powerlessness when it comes to removing harassing content. More important, despite their ability to control content, Web site sponsors are immune from liability as publishers under section 230 of the Communications Decency Act of 1996 (“section 230”). Courts have generally interpreted this provision to grant broad immunity to Web site sponsors. Section 230 thus places responsibility for content directly—and exclusively—upon those who create it, and absolutely relieves from liability the Web sites that profit (or hope to profit) from it. Yet the immunity granted to them under section 230 as publishers should not mean that Web site sponsors should be free from all liability for harm arising from their businesses.

Courts should impose tort liability upon Web site sponsors for creating unreasonable business models and hold them accountable for irresponsible and harmful business practices. To hold Web site sponsors accountable for creating socially irresponsible Web sites applies a “reasonableness” standard to Internet businesses that currently applies to offline businesses. What constitutes reasonable measures or reasonable business practices, however, should acknowledge and accommodate the differences between Internet and offline proprietorships, such as dramatically higher volume online.

Online harassment is often viewed through the prism of free speech. Attempts to curb the substance of what is being said online become mired in discussions of constitutional rights, and online “speech” is often analyzed in the same way as offline “speech.” Yet online communication is not the same as offline communication. Although harassment and bullying have always existed, when such behavior is conducted online, the consequences have different dimensions.

---

14 See, e.g., Carafano v. Metrosplash.com, Inc., 339 F.3rd 1119, 1123 (9th Cir. 2003) (noting that doubts should be “resolved in favor of immunity”); Zeran v. America Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997) (holding Communications Decency Act immunized interactive computer service provider that hosted message board, even though it refused to remove false statement after notice); Barrett v. Rosenthal, 146 P.3d 510, 529 (Cal. 2006) (noting that section 230 “does not permit” Internet service providers or users to be sued as “distributors”).
15 See Rebecca Tushnet, Power Without Responsibility: Intermediaries and the First Amendment, 76 Geo. Wash. L. Rev. 986, 1016 (2008) (noting that the Communications Decency Act uncouples ISP “property” ownership from responsibility). Tushnet proposes that Internet intermediaries’ immunity should be tied to limits on their ability to control speech. Id. at 1015.
16 See Fair Hous. Council of San Fernando Valley v. Roomates.com, LLC, 521 F.3d 1157, 1162 n.9 (9th Cir. 2008) (expressing concern with applying different rules for offline and online businesses). The Ninth Circuit stated: “[W]e must be careful not to exceed the scope of the immunity provided by Congress and thus give online businesses an unfair advantage over their real world counterparts, which must comply with laws of general applicability.” Id.
17 See discussion infra Part VI.
The anonymity of harassers, section 230 immunity, the ease of widespread digital dissemination, and the inability to contain and eliminate online information change the nature of harassment when it is conducted on the Internet. To apply a First Amendment analysis without recognizing the ways in which online communication differs from offline communication leaves many of the harms created by online harassment unaddressed.

Although this Article proposes several strategies that Web site sponsors can implement to reduce the incidence of online harassment, it does so with the awareness that any solution must be flexible enough to accommodate technological evolution. The objective of this paper is not to provide static solutions to online harassment; rather, it is to propose a new way of looking at what is, in fact, a new and evolving problem. This Article explains how online harassment is not “just like” harassment offline, and argues that to apply existing free speech doctrine without recognizing those differences ignores the norm-shaping impact of Internet communication at this stage of technology adaptation and accommodation. ¹⁸

Although my proposals do not eliminate all forms of online harassment on all Web sites, they do encourage moving away from the impulsive “anything goes” culture that prevails on some Web sites,¹⁹ to one that requires more reflection and accountability. My proposals thus seek to encourage First Amendment values rather than to chill expression, without falling prey to slippery-slope First Amendment absolutism. Thus, one objective of this Article is to reconcile the culture of the Internet with offline social norms of behavior.²⁰

This Article attempts to avoid solutions that are overbroad by specifically delineating the problems at issue. Part II identifies and describes the various types of conduct that fall under the umbrella definition of “online harassment.” It also provides a brief overview of current legal doctrines addressing the problem of online harassment.

Part III explains why existing legal remedies are inadequate to solve online harassment. Part IV proposes “reasonable measures” that Web site sponsors should take to reduce the incidence of online harassment on their sites. The proposals

¹⁸ This is not to say that there are not legitimate free speech concerns in Internet communication. For a discussion of these concerns, see Dawn C. Nunziato, The Death of the Public Forum in Cyberspace, 20 BERKELEY TECH. L.J. 1115 passim (2005) (lamenting the lack of public forums in cyberspace); Stacey D. Schesser, A New Domain for Public Speech: Opening Public Spaces Online, 94 CAL. L. REV. 1791 passim (2006) (advocating the formation of state-sponsored Web sites that would constitute public forums for free speech).

¹⁹ See supra note 9 and accompanying text.

²⁰ See generally Anne Wells Branscomb, Anonymity, Autonomy, and Accountability: Challenges to the First Amendment in Cyberspaces, 104 YALE L.J. 1639, 1647 (1995) (noting that “more and more often, confrontations are arising between the legal expectations of the real world and the developing ‘netiquette’ of the ‘netizens’ of cyberspaces” and acknowledging that “transferring legal norms from the real world may result in the application of rigid rules inappropriate to the cybercommunities and may jeopardize the full development of the information agora that the technology promises”).
focus on deterring, rather than penalizing, online harassment. The proposals also consider the ways in which Web-based businesses are different from brick-and-mortar businesses. In particular, accounting for the high volume of traffic handled by some Web proprietors, the proposals do not impose prescreening obligations or require proprietors to make difficult subjective decisions regarding whether to remove user-supplied content.

Part V summarizes and further develops an argument that I first proposed elsewhere: that tort law—in particular, a liability analogous to premises or business owner liability—may effectively be used to impose standards of conduct upon Web site sponsors. I recommend adopting at least some of the proposals in Part IV as “reasonable measures” to prevent foreseeable online harassment. To require adoption of an anti-cyberharassment policy is consistent with section 230, as it holds the Web site sponsor accountable for its own actions or omissions, not for the content posted by third parties. Part VI addresses the constitutional issues raised by anti-cyberharassment policies.

This Article concludes that the problem of online harassment necessitates a change in the way we currently view the role of Web site sponsors. Web site sponsors are proprietors of businesses, not state-sponsored public forums. Some Web site sponsors have accepted the responsibilities that come with proprietorship by creating safeguards and designing Web sites that discourage unlawful activity. By contrast, other Web site sponsors have exploited their section 230 immunity by intentionally adopting business models and designing their Web sites in ways that encourage online harassment. Although they may be immune from liability as publishers of harmful content on their Web sites, they should be held liable as proprietors for the creation of businesses likely to cause foreseeable harm to third parties.

II. WHAT IS CYBERHARASSMENT?

The term “online harassment” or “cyberharassment” is typically used to refer to Internet postings intended to embarrass, annoy, threaten, or bother another individual (as opposed to a social or political group or movement). Some words capture specific types of cyberharassment. “Cyberstalking,” for example, is the

---


22 *Id.* at 118. Courts have rejected the premises-liability argument. I discuss the leading case rejecting premises liability and why the court’s grounds for rejecting the theory was wrong in Part V.

23 See *infra* note 164 and accompanying text.

24 See *supra* note 9 and accompanying text.

25 See *supra* notes 13–15 and accompanying text.

26 Although harassment of a social or political group or organization, such as racial or religious minorities, constitutes cyberharassment, this Article focuses specifically on harassment targeted at individuals.
term most frequently used to describe the threatening, often anonymous stalking of an individual through chat rooms, e-mail, and other forms of instant communication.27 Some forms of online harassment are not threatening as much as they are annoying or humiliating. Some commentators distinguish “cyberharassment” from “cyberbullying” by defining cyberharassment as directed at adults and cyberbullying as directed at students or children.28 I use the term cyberharassment to include cyberbullying, but the range of conduct and communication makes the use of one term inadequate. The absence of precise terminology makes discussion of cyberharassment difficult. Consequently, proposals aimed at one type of conduct may be inappropriate (either under- or overinclusive) for other types of conduct.

The usefulness of the phrase cyberharassment as a broad, catch-all term makes it necessary to categorize the various types of conduct that fall under it. For example, cyberharassment covers both repeated and unwanted e-mail messages from known acquaintances, as well as threatening and aggressive blog postings from anonymous posters. It also covers distribution of video clips and digital images of a personal, embarrassing, or intimate nature. While all these examples share a commonality—the use of the Internet as the medium for distributing the communication—the maliciousness of the actions varies, as does the intended and likely effect upon the victim of the harassment. Accordingly, the solutions to prevent, deter, or punish such actions should also vary. Failure to delineate and categorize different types of conduct risks policy proposals and solutions that are overbroad or otherwise ill-suited for some problems, though appropriate for others.

To be useful, a characterization of forms of cyberharassment must recognize that words and images may be employed in a variety of ways and to different effects. Although online harassment can occur in closed arenas, such as through e-mail or closed group invitations, this Article is limited to online harassment that is conducted on publicly accessible Web sites, including those sites that require membership, so long as membership is nonselective and available to anyone who applies.29 Certain online communication (e.g., e-mails, list-serve communications,


29 For the sake of brevity, I will refer to Web sites that require membership, such as social networking sites, as “publicly accessible,” where the contents of those sites are available to all members and membership is nonselective. In other words, if the content is available only to invited viewers, then that site, or portion of that site, is not publicly accessible. If the content is accessible if one registers with the site, and registration is automatically granted, then that site is publicly accessible for the purposes of this Article.
and invitation-only Web sites that require passwords and whose contents are not searchable) should be accorded greater protection than publicly disseminated speech because closed-communication speech is already restricted in its manner of distribution. Furthermore, for reasons discussed in Part III, publicly viewable and searchable information has the potential to cause greater harm than restricted Web sites.  

This Part provides a description of the various forms of online harassment by dividing harassing conduct into two general categories—verbal and visual/auditory—which will be used in the remainder of the Article.

A. Verbal Cyberharassment

Words convey meaning online, but not necessarily in the same way as in the physical world. Words may be used to communicate intent to harm another. They may also be used simply as a way to express oneself, but even nonmalicious communications may incidentally harm another.

1. Online Threats

- A software developer is attacked on various blogs, including her own, and threatened with rape, mutilation, and strangulation, including a statement that it would be a “pity if you turned up in the gutter where you belong, with a machete shoved in that self-righteous little cunt of yours.”  

- A marketing consultant is threatened with bodily harm because of his blog posts.

The expression of intent to inflict harm upon a person that causes the person to reasonably fear for his or her safety constitutes an “online threat” when such expression is communicated through the medium of the Internet. Cyberstalking is a pattern of repeated, credible online threats. Systematic and organized online threats may constitute cyberterrorism, as further explained below.

30 Cyberharassment on closed Web sites also causes harm to its victims; however, to avoid being overbroad, this Article focuses on problems that are particular to publicly accessible Web sites. For example, “invitees” to password-protected sites are not members of the general public, and my analysis and application of tort law in Part V is consequently restricted to business owner liability to invitees of publicly accessible Web sites. Nevertheless, many of the issues raised in this Article could be applied to closed Web sites.

31 See Citron, supra note 12, at 64–65; see also Kathy Sierra, Creating Passionate Users, http://headrush.typepad.com (Apr. 6, 2007) (discussing why she has removed the post that generated threats).

32 Alex Pham, Cyberbullies’ Abuse, Threats Hurl Fear Into Blogosphere, L.A. TIMES, March 31, 2007, at C1.
2. **Online Insults**

   - The image of a boy with a rare congenital disease is posted on a Web site where users cruelly ridicule his appearance, calling his appearance “fucking hilarious,” “frighteningly akin to the Joker from the Dark Knight movie,” and a “grotesque sin.”

      Insults are words that are used to offend, deride, or embarrass another. Online insults differ from online threats because the target or subject of the insult does not feel threatened. He or she may feel embarrassed or offended, but does not feel frightened or in danger as a result of the insult. An online insult would include opinions about an individual, or an occurrence involving that individual.

3. **Online Gossip**

   - Women are encouraged to anonymously post information on a public Web site about their exes, such as whether they are promiscuous, have sexually transmitted diseases, and have illegitimate children.

   - An anonymous user posts on a popular social networking site that another family in the user’s neighborhood has a son—identified by name—who “has been to jail” and dates “underaged girls.”

      Gossip is the spreading of rumors or personal information about others. Gossip is distinguishable from insults in that it is presented as “factual.” Gossip includes both rumors that are later substantiated as accurate, as well as falsehoods.

4. **Online Confessions**

   - A blogger reveals explicit sexual information about her identifiable partners, including that one enjoyed being spanked, another was married, and that a third paid her for sex.

---

34 See DontDateHimGirl.com, http://dontdatehimgirl.com (last visited Sept. 1, 2009); see also Lizette Alvarez, (Name Here) is a Liar and a Cheat: Don’tDateHimGirl.com, N.Y. TIMES, Feb. 16, 2006, at G1.
36 See April Witt, Blog Interrupted; When Jessica Cutler Put Her Dirty Secrets on the Web, She Lost Her Job, Signed a Book Deal, Posed for Playboy and Raised a Ton of Questions About Where America is Headed, WASH. POST, Aug. 15, 2004, at W12.
The wife of a Broadway mogul reveals on a popular video-sharing site that she discovered her husband’s stash of porn and Viagra, and claimed that he was likely having an affair.37

Confessions are revelations of intimate details about oneself and one’s relationships with others. Online confessions include personal blogs and information posted on social networking sites. In addition, online confessions may be captured in the form of a video clip.

5. Cyberdeception

A thirteen-year-old girl commits suicide after communicating with a woman who was posing as a teenage boy through a fake MySpace account.38

A man posing as a woman posts an ad on a popular message board seeking a “brutal dom muscular male.” He then posts on his blog the names, pictures, e-mail addresses, and phone numbers of the men who respond.39

On the Internet, as the oft-quoted New Yorker cartoon states, nobody knows that you’re a dog.40 In some cases, the ability to disguise oneself and masquerade as someone else has led to tragic consequences.41 The ease of hiding one’s true identity and the use of communication tools to forge relationships make cyberdeception especially devious, particularly given that there is sometimes no crime for the resulting emotional wrongs.42

40 See Peter Steiner, On the Internet, Nobody Knows You’re a Dog, Cartoon, THE NEW YORKER, July 5, 1993, at 61.
41 See Jennifer Steinhauer, Verdict in MySpace Suicide Case, N.Y. TIMES, Nov. 27, 2008, at 25 (discussing how a forty-one-year-old woman disguised as a teenage boy harassed a teenage girl who subsequently committed suicide).
42 Id. (noting there was no existing crime under Missouri law for cyberdeception).
6. Cyberterrorism

- A female blogger and software developer suspends her blog and cancels public appearances after being attacked online, including having hackers reveal her home address and Social Security number.43

- A woman who had engaged in cyberdeception resulting in the suicide of a teenage girl finds herself the target of online mobs who reveal her e-mail address, satellite images of her home, and her phone numbers.44 She becomes the subject of death threats, including having a brick thrown through her kitchen window.45

In some cases, harassers take action beyond communicative activity. Cyberterrorism is the use of intimidation in a systematic way to achieve a particular objective, other than pure communication. Cyberterrorists may hack into a victim’s e-mail or online banking accounts, publicly reveal personal data such as Social Security numbers, and initiate online campaigns aimed at shutting down the victim’s personal Web site or blog.46

B. The Use of Images in Online Harassment

- Pedophiles swap images of children online.47

- School children surreptitiously snap pictures of their classmates undressing during gym class and post the photographs to a public Web site.48

- A group of children forces a classmate to engage in humiliating acts while videotaping him. The video is then posted to a public Web site for other classmates to view.49

---


44 See Schwartz supra note 39, at 26–27.

45 Id.

46 See Citron, supra note12, at 64–65; Valenti, supra note 43, at 16 (discussing how Kathy Serra, a blogger and software developer, was attacked on her own blog and on other Web sites when posters revealed her home address and Social Security number).

47 See Kurt Eichenwald, From Their Own Online World, Pedophiles Extend Their Reach, N.Y. TIMES, Aug. 21, 2006, at A1.

A woman claims that her ex-boyfriend has a sexually transmitted disease and posts his photograph and name on a publicly searchable social networking site.\textsuperscript{50}

A jilted boyfriend vengefully posts nude pictures of his ex-girlfriend online.\textsuperscript{51}

Online harassment may involve the use of images, such as photographs or videos. Although the method by which such images are captured may vary, for purposes of this Article’s analysis, all images—whether captured as a photograph, video, or document—are classified simply as images. Images may accompany an online confession. For example, an individual may post a video of herself talking about an impending divorce.\textsuperscript{52} In such cases, the words spoken by the individual should be viewed as an online confession and the video image (of the individual or others) should be analyzed as distinct from her confession. An image often reveals more information, and it therefore has the potential to be more damaging than a written description. For example, a blogger’s description of her lover’s face or naked body is likely to be less revelatory (and invasive to the lover’s privacy) than posting his picture. As discussed in Part VI, the posting of certain images should not be equated with speech protected by the First Amendment.

III. The Need for Alternative Approaches to Cyberharassment

Part III explains why existing remedies are inadequate to address the problem of online harassment. It then proposes reframing the problem of online harassment as a failure of business norms rather than as a constitutional right to expression.

A. An Overview of Existing Remedies

Laws currently address some, but not all, of the crimes that now fall under the umbrella definition of online harassment. This Part is not intended to be an exhaustive exposition of such remedies; nor is it a discussion of the various conceptions of privacy. Rather, it is intended to provide some necessary

\textsuperscript{49} Id.
\textsuperscript{50} See DontDateHimGirl.com, http://dontdatehimgirl.com (last visited Sept. 1, 2009).
\textsuperscript{52} One high-profile divorce video garnered over three million views. Trisha Walsh-Smith–The Video That Started it All!, http://www.youtube.com/watch?v=hx_WKxqQF2o (last visited Sept. 1, 2009).
background to frame the proposals offered in Part IV. Generally, the remedies currently available to victims (of some types) of online harassment can be grouped in three broad categories: (1) tort actions deriving from privacy, (2) nonprivacy tort claims such as defamation and intentional infliction of emotional distress, and (3) criminal or anti-stalking statutes.

1. Tort Actions Deriving from the Right to Privacy

The primary remedies for victims of online harassment derive from the right to privacy. Samuel Warren and Louis Brandeis first recognized privacy as a legal right in a groundbreaking article. Warren and Brandeis described the “right of privacy” as a natural development of the common law:

That the individual shall have full protection in person and in property is a principle as old as the common law; but it has been found necessary from time to time to define anew the exact nature and extent of such protection. Political, social, and economic changes entail the recognition of new rights, and the common law, in its eternal youth, grows to meet the demands of society.

As many scholars have noted, however, although courts have recognized the existence of a “right to privacy,” the parameters of such a right are vaguely defined.

The recognition of a right to privacy gave rise to several common law actions in tort, namely appropriation, false light, disclosure or wrongful


54 Samuel Warren & Louis Brandeis, The Right to Privacy, 4 HARV. L. REV. 193, 193 (1890). While the concept of privacy existed prior to publication of Warren and Brandeis’s article, the article is widely acknowledged as being the first to establish the legal foundation for such a right. For further discussion on privacy as a legal right, see Solove, supra note 53, at 481–83, and Katherine Strandberg, Privacy, Rationality, and Temptation: A Theory of Willpower Norms, 57 RUTGERS L. REV. 1235, 1267–68 (2005).

55 Warren & Brandeis, supra note 54, at 193.


57 RESTATEMENT (SECOND) OF TORTS, § 652A (1977) (“One who invades the right of privacy of another is subject to liability for the resulting harm to the interests of the other.”).

58 Id. § 652C (“One who appropriates to his own use or benefit the name or likeness of another is subject to liability to the other for invasion of his privacy.”).

59 Id. § 652E (“One who gives publicity to a matter concerning another that places the other before the public in a false light is subject to liability to the other for invasion of his
publication of private facts, and intrusion. Causes of action based upon privacy torts would be most appropriate where the plaintiff was the subject of online confessions, online gossip, and online insults.

2. Other Torts

Defamation law protects the interests of a person in his or her reputation. To establish liability for defamation, the plaintiff must show that the defendant made a false and defamatory statement that harmed the plaintiff’s reputation. Two types of defamation torts exist: libel and slander. Libel is the publication of defamatory statements by printed or written words. Slander is the publication of defamatory matter by spoken words or by any form of communication other than those covered by libel. Defamation-based causes of action would be appropriate where the plaintiff is the subject of online gossip, online insults, and online confessions, provided that the statements made were untrue. Ironically, the wild, juvenile nature of much online discourse may exculpate posters, as the context may indicate that it should not be taken seriously—although it may nevertheless tarnish the reputation of the subject of the online harassment.

In addition, the target of the online harassment may sue for intentional infliction of emotional distress, sometimes referred to as the tort of outrage, by proving that the poster engaged in extreme or outrageous conduct intending to cause severe emotional distress. While the intentional infliction of emotional distress may provide a basis for all kinds of online harassment, it is particularly useful with cyberdeception, which often has a different character from, and occurs more frequently than, offline cases of false identity.

privacy, if (a) the false light in which the other was placed would be highly offensive to a reasonable person, and (b) the actor had knowledge of or acted in reckless disregard as to the falsity of the publicized matter and the false light in which the other would be placed.

Id. § 652D (“One who gives publicity to a matter concerning the private life of another is subject to liability to the other for invasion of his privacy, if the matter publicized is of a kind that (a) would be highly offensive to a reasonable person, and (b) is not of legitimate concern to the public.”).

Id. § 652B (“One who intentionally intrudes, physically or otherwise, upon the solitude or seclusion of another or his private affairs or concerns, is subject to liability to the other for invasion of his privacy, if the intrusion would be highly offensive to a reasonable person.”).


See id. § 568.

See id.

See supra notes 2–4 and accompanying text.

RESTATEMENT (SECOND) OF TORTS § 46 (1977) (“One who by extreme or outrageous conduct intentionally or recklessly causes severe emotional distress to another is subject to liability for such emotional distress, and if bodily harm to the other results from it, for such bodily harm.”).
3. Crimes

In addition to torts, online threats and/or cyberstalking may also be a crime.\textsuperscript{68} Typically, the defendant must have engaged in behavior or a pattern of conduct with the intent to alarm, abuse, or frighten the victim.\textsuperscript{69} Acts of cyberterrorism may be criminalized under anti-hacking statutes.\textsuperscript{70}

B. The Inadequacy of Existing Remedies

Online harassment is distinct in both the process by which it occurs and the harms that it creates. While existing remedies discussed in the previous section may adequately address offline harassment, they are inadequate to deal with online harassment for several reasons.

First, posting is cheap and easy, which produces two effects with respect to online harassment. Offline publishers often have deep pockets, and the range of their distributive reach correlates with their financial means. On the Internet, however, widespread distribution is available to those without substantial financial resources. Consequently, even where a plaintiff prevails in a civil action against an online harasser, the odds are high that the plaintiff will not be able to recover significant damages.\textsuperscript{71}

Furthermore, online harassment affects private individuals in a very public manner by means that were previously infeasible. Because posting is cheap and easy, many forms of online harassment are more likely to involve a nonpublic figure than offline forms of the same conduct. For example, online publication of insults and gossip about nonpublic figures is much more common than publication of insults and gossip about nonpublic figures in traditional media. Unfortunately, because litigation is costly, many private individuals who are the target of online harassment do not have the financial resources to pursue legal remedies. For

\textsuperscript{68} See, e.g., CAL. CIV. CODE § 1708.7 (West 2009) \textit{(describing the tort of stalking, including cyberstalking, where the defendant engaged in a “pattern of conduct the intent of which was to follow, alarm or harass the plaintiff” and the plaintiff “reasonably feared for his or her safety, or the safety of an immediate family member”); CAL. PENAL CODE § 422 (West 2009) \textit{(making it a crime for a person “who willfully threatens to commit a crime which will result in death or great bodily injury to another person, with the specific intent that the statement, made verbally, in writing, or by means of an electronic communication device, is to be taken as a threat, even if there is no intent of actually carrying it out” thereby causing a person “reasonably to be in sustained fear for his or her own safety or for his or her immediate family’s safety”}).

\textsuperscript{69} See CAL. CIV. CODE § 1708.7 (West 2009); CAL. PENAL CODE § 422 (West 2009); \textit{see also MASS. GEN. LAWS ANN. ch. 265, § 43 (LexisNexis 2002); OHIO REV. CODE § 2917.21(B) (LexisNexis 2005); ALASKA STAT. § 11.61.120 (2007); TEX. PENAL ANN. § 42.07 (Vernon 2008); ALA. CODE § 13A-11-8 (2009).}

\textsuperscript{70} See, e.g., 18 U.S.C. § 1030 (2006) \textit{(criminalizing fraud and related activity in connection with computers).}

\textsuperscript{71} See, e.g., Margolick, \textit{supra} note 2.
example, the female law student plaintiffs in the AutoAdmit case could afford to bring a lawsuit against their aggressors because they were represented at no charge by one of the country’s leading litigation firms, and by one of the country’s leading intellectual property lawyers. It is unlikely that they would have done so otherwise; although the plaintiffs were seeking monetary damages, they would probably not have received much because many of the defendants were students and/or recent graduates and may have lacked the financial means to satisfy a substantial judgment.

Second, many harassing posts are anonymous. Anonymity removes many of the social controls that may have deterred offenders in the pre-Internet era. Anonymity also reduces accountability and accuracy. Anonymous information is simply not disseminated as easily offline as it is on the Internet. While one may argue that anonymously authored postings are not as credible as identified postings, the mere existence or prevalence of online gossip or online insults may have a negative effect even where such information is refuted or discredited. One study showed that repeated exposure to information made people believe the information was true, even where the information was identified as false. The “illusion of truth” appears to come from increased familiarity with the claim and decreased recollection of the original context in which the information was received.


73 See id.

74 See DANIEL J. SOLOVE, THE FUTURE OF REPUTATION: GOSSIP, RUMOR AND PRIVACY ON THE INTERNET 140 (2007) (noting “anonymous, people are often much nastier and more uncivil in their speech [because it] . . . is easier to say harmful things about others when we don’t have to take responsibility”). Solove adds that a gossiper risks harm to his or her own reputation, as well as the reputation of others: “If a person gossips about inappropriate things, betrays confidences, spreads false rumors and lies, then her own reputation is likely to suffer. People will view the person as untrustworthy and malicious. They might no longer share secrets with the person. They might stop believing what the person says.” Id. at 140–42.

75 Justice Scalia made this argument in his dissent in McIntyre v. Ohio Elections Commission, 514 U.S. 334, 382 (1995) (Rehnquist, C.J., joined by Scalia, J., dissenting) (noting “a person who is required to put his name to a document is much less likely to lie than one who can lie anonymously” and that anonymity “facilitates wrong by eliminating accountability”); see also Branscomb, supra note 20, at 1642–43 (noting anonymous or pseudonymous postings “relieve[] their authors from responsibility for any harm that may ensue [and that] . . . [t]his often encourages outrageous behavior without any opportunity for recourse to the law for redress of grievances”); SOLOVE, supra note 74, at 462–64.


77 Id.
Anonymity also reduces the likelihood of nonlegal measures or conciliatory efforts. If one can anonymously submit a post airing a grievance or a claim, one has less incentive to seek out the object of the post to determine its accuracy or to resolve the conflict giving rise to the grievance. Perceived wrongs can be redressed the coward’s way, by anonymously posting the rumor or incident for public opprobrium. If the victim of a post is unable to identify the poster, he or she is unable to resolve any conflicts or clarify any issues in a nonlegal manner. The victim of online harassment must initiate legal proceedings to unmask the identity of the poster as there may be no other way to negotiate with the poster or respond to the posting, other than by responding via a post. In fact, anonymity removes any opportunity to redress grievances in a nonpublic manner if the Web site sponsor is unwilling to intervene on the victim’s behalf. The victim must try to ignore the post (an admirable but perhaps unrealistic effort), react via a responsive post, or initiate an often costly and time-consuming lawsuit, which may draw even more attention to a humiliating or threatening post. As previously mentioned, the lawsuit may ultimately be fruitless because the poster is without significant resources or is a minor. Even if the subject of the post prevails in a lawsuit, posts may remain online and linked to by other Web sites during the drawn-out litigation, causing more emotional harm.

Third, near instantaneous, widespread dissemination and the impossibility of recapturing distributed postings put online harassment injuries in a class by themselves. There is no comparable injury in the offline world because there is no other method of distribution that is as inexpensive, accessible, widespread, and difficult—if not impossible—to retrieve. Online insults, for example, may not be libelous; yet, through their widespread distribution and permanence, they are harmful in a way that offline insults are not. Secrets and gossip have the potential to cause much greater reputational damage when spread online than when shared over a cup of tea among friends. Furthermore, anonymity enables less restrained disclosure as it frees the discloser from the stigma associated with salacious information.

Whether the Internet is different from the offline world, and thereby necessitates different rules, has been a recurring topic of discussion among

---

78 See infra notes 84–86 and accompanying text.
79 See infra notes 84–86 and accompanying text.
80 See infra notes 84–86 and accompanying text; see also Caroline E. Strickland, Note, Applying McIntyre v. Ohio Elections Comm’n to Anonymous Speech on the Internet and the Discovery of John Doe’s Identity, 58 WASH & LEE L. REV. 1537, 1539–41 (2001) (noting that without knowledge of online posters’ identities, a corporation that is the object of cybersmears has limited options and that, as a result, “Corporation X can only hope to obtain a damages award from the thus far unidentified parties,” by filing a lawsuit against “John Doe”).
81 See supra note 73 and accompanying text; see also Branscomb, supra note 20, at 1643 (“Law enforcement officials or lawyers seeking to file a civil suit might not be able to identify an individual to hold responsible.”).
Perhaps the more relevant issue, at least regarding the problem of online harassment, is not whether the Internet is inherently different as a mode of communication, but, rather, whether Internet-related conduct and the effects of such conduct is, and has been, treated differently from non-Internet-related conduct. The answer to that is an unequivocal yes. Users say and do things on the Internet they would not in the offline world, which is something psychologists refer to as the online “disinhibition effect.”[83] Web site sponsors routinely accept anonymous postings, whereas major newspapers generally require at least a contact name and telephone numbers, and often require additional biographical information about the author.[84] Offline gossipmongers typically do not spread their information wearing bags over their heads. The petty and malicious gossip of college students is not typical fodder for national print publications.

Even where anonymity is not an issue because the poster has self-identified, the Internet poses unique challenges that make traditional responses to harassment inadequate. Web site sponsors often disclaim all responsibility for harassing conduct that occurs on their Web sites, and many make no attempt to monitor or control uploaded content.[85] In contrast, if such conduct were to occur on physical property—written on a bathroom wall or posted on a bulletin board in a store, for example—social norms and fear of a lawsuit would compel the store owner to take some sort of action. Distributing false rumors in the physical world is a tort for which the publisher is liable, yet in cases where rumors are spread via the Internet, Web sites are immune from liability under section 230 of the CDA.[86]

---

82 See, e.g., Frank H. Easterbrook, Cyberspace and the Law of the Horse, 1996 U. CHI. LEGAL F. 207, 208 (noting it is not clear whether many features of existing law can be appropriately applied in the cyberrealm); cf. Lawrence Lessig, The Law of the Horse: What Cyberlaw Might Teach, 113 HARV. L. REV. 501, 502 (1999) (arguing “there is an important general point that comes from thinking in particular about how law and cyberspace connect”).


84 See, e.g., Op-Ed Guidelines for the Wall Street Journal, WALL ST. J., http://www.opinionjournal.com/guidelines (requiring the contact name and telephone number of authors); Submissions & Contributions, S.F. CHRON., http://www.sfgate.com/chronicle/submissions (requiring the contact name and telephone number of authors); Submitting an Article to Op-Ed, L.A. TIMES, http://www.latimes.com/news/opinion/oe-howtosubmitoped,0,5238591.story (requiring the contact name and telephone number of authors, as well as biographical information).

85 See supra note 9 and accompanying text.

The remedies currently available require a victim of online harassment to file costly, time-consuming, and often fruitless lawsuits. Although this may be a problem in the physical world, it is a greater problem in the virtual world because of the unique aspects—ease of publication, section 230 immunity for Web sites, anonymity, widespread distribution, and lack of control—of Internet publication.

Online harassment can be combated by changing the roles that Web site sponsors currently play and by imposing tort liability on those who fail to meet certain expectations. Web site sponsors maintain a proprietary interest in their Web sites, and we should expect them to conform to the standard of conduct expected of other proprietors. The interpretation of that standard, however, should recognize the differences between online and offline businesses. In other words, while we should expect proprietors to conduct their businesses with reasonable care, what constitutes reasonable care should take into consideration the differences between online and offline businesses. The proposals set forth in Part IV provide guidance as to what might constitute reasonable care on the part of Web site sponsors.

IV. PROPOSALS TO ADDRESS THE PROBLEM OF CYBERHARASSMENT

Given the harm created by nearly instantaneous widespread dissemination, the lack of editorial controls, and other obstacles to publication—and perhaps most significantly, the irretrievability and permanence of the effects of online harassment—the primary objectives of the proposals set forth in this section are deterrence and prevention. Tort remedies and criminal prosecutions provide scant comfort to victims of online harassment who lack the resources to pursue available remedies and/or wish to avoid further publicity. The remedies also often exacerbate the emotional trauma of having had personal details or images released to an audience of millions. The permanence of Internet distributions also makes rehabilitation much more difficult. One undergraduate student stated it was a challenge to continue his new life as a college student after he was identified by name as having appeared in a pornographic movie in a post on the Juicy Campus 2006) (noting section 230 “does not permit” Internet service providers or users to be sued as “distributors”).

87 Black’s Law Dictionary defines “proprietor” as “[a]n owner, esp. one who runs a business.” BLACK’S LAW DICTIONARY 1339 (9th ed. 2009).

88 Solove raises many of the same concerns expressed in this Article. His proposals focus on broadening the current definition and application of existing tort remedies to deter future instances of cyberharassment. See Solove, supra note 74, at 113. He also introduces remedies, such as alternative dispute resolution measures, that would limit the impact of cyberharassment. Id. at 124. Although I agree with many of Solove’s proposals, my proposals focus more on prevention of cyberharassment through primarily nonlegal measures.
Web site. The post also linked to a Web site that showed him engaging in explicit sexual acts with other men.

These proposals recognize that some forms of online harassment have the potential to cause more harm than others. They also take into account that at least some forms of online harassment potentially have social value, including expressive value, whereas other forms of online harassment do not. These proposals consider and balance the competing interests of both the poster and the subject of the post.

The proposals focus primarily on two aspects of Internet conduct that differentiate online harassment from physical-world harassment. The first is anonymity. While anonymous speech occurs in the physical world, it is more difficult to accomplish and less likely to occur than online. While anonymity has benefits—in particular for minority groups who may, for economic or social reasons, lack other forums for communication—it also has serious drawbacks. As discussed in Part III(B), anonymity may remove incentives for self-regulating behavior so that an anonymous poster may be more inclined to post damaging material. A poster may also be willing to post more vicious or less discreet information if he or she can do so anonymously. Disassociative anonymity is one of the factors causing online disinhibition. Anonymous users feel less vulnerable about being frank and do not feel responsible for their online conduct.

Finally, the notion of anonymity itself is misleading. In many cases, the identity of a user is accessible, even if such accessibility requires some effort. The government or a party to a lawsuit can subpoena user identification. In addition, the Web site sponsor often requires user registration, thereby identifying the poster internally, even if users are unaware of the poster’s identity. This Article’s proposals thus make plain for the user that anonymity is not permanent or secure, and may encourage users to reconsider impulsive, regrettable behaviors without overly restricting considered and desired communications.

My proposals also advocate a more responsible role for Web site sponsors. Because of the private regulatory nature of the Internet, Web site sponsors are in

90 Id.
91 Suler, supra note 83 (noting that “anonymity works wonders for the disinhibition effect”).
92 Id.
93 See supra notes 2–5 and accompanying text (discussing the AutoAdmit case).
94 See Branscomb, supra note 20, at 1643 (“Many providers of computer-mediated facilities do not permit genuine anonymity. They keep records of the real identity of pseudonymous traffic so that abusers can be identified and reprimanded.”). Branscomb, however, notes there is a trend toward using “anonymous remailers” who “provide a guarantee that messages cannot be traced back to their sources.” Id.
95 See Nunziato, supra note 18, at 1116 (noting the “vast majority of speech on the Internet today occurs within private places and spaces that are owned and regulated by private entities”).
the best position to address the problem of online harassment. Web site sponsors have borrowed the rhetoric of free speech as though they were public forums while taking full advantage of their sites as private businesses.\(^{96}\) As a matter of fairness and consistency, Web site sponsors should have a duty to take reasonable steps to prevent tortious and criminal conduct on their Web sites, at least in certain circumstances. These reasonable steps should include policies to deter online harassment that incorporate at least some of the proposals set forth in this Article.

\(A.\) Contractual and Architectural Constraints

One way Web sites may prompt posters to reconsider their postings is by implementing contractual and architectural restraints, many of which already exist on commercial Web sites.\(^{97}\) As many scholars have explained, the way choices are presented influences their selection or nonselection.\(^{98}\) Object design or architecture

\(^{96}\) As Dawn Nunziato notes:

Private regulation of speech on the Internet has grown pervasive, and is substantially unchecked by the Constitution’s protections for free speech, which generally apply only to state actors’ regulations of speech. At an earlier stage of the Supreme Court’s First Amendment jurisprudence, such private speech regulations might have been subject to the dictates of the First Amendment under the state action doctrine. The Supreme Court, however, has substantially limited the application of the state action doctrine in past decades, and courts have been unwilling to extend this doctrine to treat private regulators of Internet speech as state actors for purposes of subjecting such regulation to First Amendment scrutiny.

\(^{97}\) While some of the contractual restraints may prove to be unenforceable if the poster is a minor, they may nevertheless have a deterrent effect on undesirable behavior. Furthermore, a court might enforce a Web site’s terms of service against a minor. See A.V. v. iParadigms, 544 F. Supp. 2d 473, 480–81 (E.D. Va. 2008) (enforcing the terms of a clickwrap agreement against minor plaintiffs and stating that plaintiffs could not “use the infancy defense to void their contractual obligations while retaining the benefits of the contract”).

along with human foibles and limitations influence decisions and object use (or misuse). The contractual and architectural constraints proposed in this Part create structural barriers to speech without unreasonably restricting it. In other words, the constraints may serve in their modest way the purposes of time, place, and manner restrictions, as well as the purposes of physical, logistical, and normative barriers imposed upon speakers in the physical world.

1. Manifesting Assent to Web Site Policies

Web site sponsors can require posters to register with the Web site before gaining the ability to post comments. Although posters can easily create e-mail addresses and use false information, having to go through the motions of doing so may slow down the posting process or cause posters to reconsider their communications. As part of the posting process, a poster may be reminded of the Web site’s policy against online harassment. The poster may be asked to click “I agree” to the terms of the policy. The physical act of consent may remind the poster of the legally binding nature of agreement. In the alternative, or in addition to expressing agreement to the terms of the Web site’s online harassment policy, the poster may be asked, “Are you sure you want to post this?” A simple question requiring a pause may be annoying to some, but it may also cause posters to consider whether they really want to continue. Web site sponsors can also institute comment policies or moderate user comments.

2. Indemnification for User Misconduct

Additionally, Web site sponsors could discourage online harassment by requiring that all users contract to indemnify the sponsors from any harm the users cause. Part V advocates the imposition of proprietorship tort liability upon Web site sponsors for foreseeable harm caused by third parties. Web site sponsors, in turn, can reduce their risk of tort liability by contracting for indemnification with their users. To ensure enforceability, the agreement should be concise,
understandable, and conspicuous. Furthermore, the user should be required to manifest assent by clicking, rather than burying the terms in an interior “terms of use” page. Requiring active assent may make users realize that a Web site is serious about enforcing its anti-harassment policies and encourage users to consider the legal repercussions of their actions.

3. Community Controls

In addition, the Web site might incorporate user-generated controls on content. For example, some Web sites incorporate “report abuse” buttons to enable visitors to report offensive content or abusive conduct. Wikipedia enables users to update and delete content placed by others, and it has reportedly changed its policy so that new and anonymous users must have their content “flagged,” or approved by registered, reliable users. Craigslist incorporates an easy way to flag certain posts by categorizing the violation and enabling the user to click on the upper right-hand side of each listing. Amazon and eBay enable users to rate other users. All of these controls can be incorporated on other Web sites as well. A message board can have users rate the quality of a particular poster for the benefit of newcomers. The online version of the San Francisco Chronicle, for example, enables readers to click on an icon to indicate a “thumbs up” or “thumbs down” in rating a poster’s comments. Users can also report offensive posts and request that such posters be banned from the Web site.

4. Default to Identified Postings

Another architectural control that Web sites can implement is a default to identified postings rather than anonymous postings. Currently, much of Internet communication is structured to facilitate anonymous postings and to accord each posting equal weight. Instead, communication could be structured to default to


postings that identify the poster, unless the poster opts to post anonymously by clicking out of the default each time the user tries to post a message, in much the same way that companies currently require customers to opt out of receiving marketing information. It would be most effective if the ability to opt out required the user to navigate through another Web page, rather than merely checking a box on the same page. Of course, some posters would continue to falsify their registration information or take the extra step to opt out of identification, but some might not. The point is not to make identified postings mandatory, but to make identified postings easier than slightly more burdensome anonymous postings.

5. "Cooling Period"

A Web site sponsor can incorporate hurdles or procedures to slow down the posting process, encouraging posters to more carefully consider what they say before they press “Send.” Because anonymity and the ease and pace of the Internet may encourage impulsive behavior, one way to curb such behavior is to require a “cooling period.” A poster might be required to wait before a message or image is posted to a Web site. During this cooling period, the poster may choose to edit or remove the message or image from the posting “queue.” Some may object that instituting a cooling period will harm spontaneous discussion on message boards. This harm can be minimized by imposing a cooling period that reflects the type of harassment likely to occur on the Web site. Blog sites on political topics, for example, may require a shorter period of time, say ten minutes, whereas social networking sites may require a twenty-four-hour waiting period. Alternatively, a different waiting period may be instituted where the poster wishes to remain anonymous. Those users who choose to identify themselves may have a shorter waiting period, or no waiting period at all, whereas anonymous posters may be subject to a longer cooling period. Any waiting period at all encourages further reflection, which may have the additional benefit of better written and more thoughtful posts.

6. Warning Notices

Notices may also deter misconduct. A notice may inform a poster of the legal consequences of his or her actions. Some posters may be ignorant of what constitutes tortious or criminal behavior. Simply knowing that a defamatory or threatening post subjects the poster to a civil lawsuit may be enough to deter a poster who might not otherwise realize—either because of ignorance or because the issue is not at the mind’s forefront at the time of posting—the risks of his or her conduct. A reminder that a posting may have legal repercussions may prompt a poster to soften the language or reconsider the message.
A recent study illustrates the effectiveness of warning notices. Researchers created a MySpace page for “Dr. Meg” that identified her as a doctor. The researchers then found MySpace users who identified themselves as eighteen to twenty years of age, and who discussed sexual activity and substance use on their publicly viewable pages. “Dr. Meg” sent these users a note informing them of the risk of disclosing personal information along with a link to a Web site about sexually transmitted diseases. Three months later, the researchers learned that 42.1 percent of the MySpace pages had been changed (either by enhancing privacy settings or by deleting the references to sex and/or substance use) compared with 29.5 percent in the control group.

In addition to notifying the poster of what constitutes potentially criminal and tortious activity, the poster should be informed of the Web site’s policies governing user activity, especially because many such policies restrict more types of activity than are prohibited by law. This type of reminder is particularly important for social networking sites where users routinely violate policies that forbid harassing or annoying other members.

Notices could also inform posters of the circumstances under which their identities may be revealed. Some posters may not know, or may not consider at the time of posting, that “anonymity” is rarely secure and inviolable. Poster identities are usually known by the Web site sponsors and by Internet service providers. Poster identity may be revealed where a lawsuit has been filed pursuant to a subpoena, even where the Web site and the poster wish to maintain anonymity. In some cases, Web site policies may permit unmasking users under certain conditions, a few of which are suggested in the next section. Other posters may also reveal the identity of anonymous posters. For example, an anonymous blogger was recently identified as a law professor by another poster on a popular Web site. Informing posters of the tenuous nature of anonymity may cause some to modify their postings. To illustrate, it is unlikely that the students who posted offensive comments about female law students on the AutoAdmit message board

---

106 Megan A. Moreno, Reducing At-Risk Adolescents’ Display of Risk Behavior on a Social Networking Web Site: A Randomized Controlled Pilot Intervention Trial, 163 ARCHIVES OF PEDIATRIC & ADOLESCENT MED. 35, 35–41; see also Eric Nagourney, A Note to the Wise on MySpace Helps, N.Y. TIMES, Jan. 6, 2009, at D6 (discussing teenagers posting highly personal information that is accessible to future employers or online predators).
107 Moreno, supra note 106, at 36.
108 Id.
109 Id. at 37.
110 Id. at 38.
111 See supra notes 93–94 and accompanying text.
112 See supra notes 93–94 and accompanying text.
113 See supra notes 2–6 and accompanying text (discussing the AutoAdmit case).
would have done so if they had realized that their true identities might be revealed in the course of a lawsuit. They would have likely reconsidered how to phrase their posts, or they might have declined to post at all given the damage to their personal and professional reputations.

To be most effective, warnings and other notices should be concise and in plain English to be comprehensible to most users. It may be helpful for notices to include examples of what constitutes each type of prohibited conduct. Finally, notices should appear at the time of posting, rather than as a general message on the Web site or tucked in an interior page.

B. Ameliorating Anonymity’s Negative Effects

One of the problems associated with anonymity on the Internet is that it minimizes posters’ accountability for their actions. Anonymity encourages a lack of accountability because posters may feel more comfortable posting comments if their identities can remain hidden. There are, however, valid arguments in favor of anonymity. Stripping all posters of anonymity to curb online harassment is an overbroad measure because it threatens to stifle many socially beneficial forms of communication. Members of subcultures or minority social, racial, religious, and/or political groups may fear repercussions for posting unpopular views or for expressing beliefs, desires, or thoughts that do not conform to mainstream values or norms. An atheist may post remarks that question God’s existence. A teenage boy may seek support regarding his sexuality. A girl may ask for advice regarding an abortion. Anonymity enables all of them to share and receive important information. Given that one of our culture’s values is the protection of minority views, an absolute ban on anonymity is, for this reason, not encouraged.

Certain types of conduct merit more identity protection, however, than others. As previously discussed, anonymity may facilitate discussion for a variety of reasons. One of the most important is that it enables those afraid to speak to seek advice, support, and information from others without fear of repercussion. In some cases, however, anonymity is not useful because it enables and facilitates online harassment, often of members of the very same minority or disempowered groups. Currently, the only way a victim of online harassment can unmask the

---

117 See id.
118 See Christopher Wolf, Racists, Bigots and the Law on the Internet, ANTI-DEFAMATION LEAGUE, http://www.adl.org/Internet/Internet_law1.asp (last visited Sept. 1, 2009) (explaining how the Internet is “a relatively cheap and highly effective way for hate
identity of an anonymous poster is to file a lawsuit. In most cases, however, the process is time consuming and costly. Because it is time consuming, the damaging information remains online longer. The costliness of litigation is aggravated by the likelihood that the harassers are judgment-proof. Victims are consequently left without recourse. Therefore, one solution is to permit victims of certain types of online harassment to unmask harassing posters without filing lawsuits.

1. Easy Unmasking of Anonymity

The “easy unmasking” of anonymity proposed here is limited to situations where all of the following factors are present: the victim of the online harassment has been identified in a posting; the victim is a nonpublic individual; the victim has signed an affidavit swearing that he or she is the individual identified in the posting; and the victim sets forth facts establishing why easy unmasking is warranted. These limitations aim to limit easy unmasking to situations where the posts do not involve matters of legitimate public interest. The downside to these limitations is that there may be victims of online harassment (i.e., public officials) who are unable to easily unmask their anonymous harassers. Those individuals, however, may still seek to strip anonymity by filing a lawsuit and seeking recourse under the currently available remedies.\textsuperscript{119}

Several reasons support easy unmasking of a cyberharasser. The first is deterrence. If the barriers for removing anonymity are lowered, and the harasser is aware of that, he or she may be less inclined to post harassing information. The second is redress. If the victim knows the harasser’s identity, the victim can more effectively respond to the posted information and assess the credibility of the threat. Anonymous threats hover ominously online and permit the victim to imagine the most terrible culprits. A female law student who was the target of online threats on the law school admission message board, AutoAdmit, stated, “I . . . felt kind of scared because it was someone in my community who was

\textsuperscript{119} See discussion \textit{supra} Part III.A.
threatening physical and sexual violence and I didn’t know who.” In some instances, those fears are justified. In others, they may not be. Enabling victims of online threats to put a face on their tormentors helps the victims determine what they should do next, whether it is filing a police report or calling an underage harasser’s parents. Harassers may then feel social pressure to stop the harassing behavior without judicial intervention.

Some may argue that permitting the unmasking of the harasser subjects the harasser to attacks by the victim. It is important to keep in mind that easy unmasking would only be permitted where personal information about the victim, including the victim’s identity, was revealed online. Easy unmasking would not be permitted in situations where the victim was not identified to the public. Thus it would not occur in situations where one was speaking about an issue or problem, where the information was conveyed in generalities, or where the victim was not identified. Furthermore, easy unmasking would not be permitted where the gossip pertained to a company or business entity, or where the individual was a public figure. The concern about unmasking harassers in this context is then puzzling because the victim is only receiving the same type of information that the harasser felt comfortable revealing to the public about the victim. If the harasser feels it is acceptable to release to the public personal, identifying information about the victim, then the victim should be entitled to identifying information about the harasser. Turnabout is fair play and may reinforce the social norms that currently exist in the offline world. Although backbiting and gossip exist everywhere, in most cases the individual spreading the gossip must deal with the consequences of doing so. The gossiper develops a reputation for reliability or misinformation and may be sought out or shunned for the gossip. On the Internet, however, credibility is detached from identity where the gossiper’s identity is unknown. An unmasking policy enables a victim of online gossip to respond by revealing critical information about the harasser, such as ulterior motivations, which may undermine the harasser’s credibility and thereby minimize the damage from the online harassment. In some situations, such as where the victim and the harasser occupy the same social milieu, such as attending the same school or living in the same neighborhood, an unmasking helps redress the harassment by shaming the harasser. For example, in the AutoAdmit case, the degrading posts affected the academic performance and emotional well-being of the female subjects; the posters, on the other hand, continued about their daily lives comparatively unaffected. Easy unmasking would have allowed these female law students to socially shame their classmates, and would have made it more difficult for some of the posters to continue their messages unchecked. Some of the posters whose

---

121 Of course, this would not foreclose the business or public figure’s ability to seek recourse under existing legal remedies.
122 See Margolick, supra note 2.
identities were revealed as part of the lawsuit apparently had never even met the women. Of course, an easy unmasking policy is unavailing where the poster is the Web site sponsor itself (such as an anonymous blogger), or where the poster has used an anonymity service. But it may reduce some instances of online harassment where the poster has a reputation it wishes to preserve, or where the poster desires to post in a “truly anonymous” manner but does not know how to do so.

2. **Stigmatizing Anonymity**

For easy unmasking to be effective, the Web site must have an effective registration process. Many Web sites require posters to register their e-mail addresses. Other times, posters can be identified through their Internet protocol addresses. In some cases, however, a poster may use an anonymity provider or remailer using a forged identity, thus becoming “truly” anonymous. The Web site sponsor can deal with truly anonymous posters in one of several ways. It can ban truly anonymous posters from the Web site altogether. It can monitor the content from truly anonymous posters more closely, or require a longer posting time. It can also segregate anonymous postings to minimize their impact or reduce their ability to be searched.

One of the problems with anonymous postings is that readers are unable to assess the credibility of the poster. The current default for Internet postings seems to be anonymity. Changing the default to identified postings may make anonymous postings seem less credible. A poster continues to have the option of remaining anonymous; however, his or her postings can be segregated to the bottom of the message board or otherwise identified as written by someone who wished, for whatever reason, to remain anonymous. Message boards might even post a notice preceding the anonymous messages that underscores the possibility that they may lack authenticity or credibility. Web site sponsors could deprive anonymous posters of unique names, forcing them to identify only as “Anon 1,” “Anon 2,” etc., thus giving their posts less of a “headline” and no catchy alias without depriving them of a communication forum. For example, the AutoAdmit posters used juvenile pseudonyms such as “playboytroll,” “Pauliewalnuts,” “Whamo,” and “Spanky.” Depriving such posters of fanciful user names may reduce the online disinhibition effect. After having his identity revealed, “Whamo” claimed, “I

---

123 *Id.* (noting that one of the students was an undergraduate at the University of Iowa and another a Seton Hall graduate.)

124 The ISP, however, may be required by a court to reveal the blogger’s identity. See James Bone, *Rude Blogger Is Unmasked by Model*, THE TIMES (London), Aug. 19, 2009, at 8 (reporting that a New York supreme court judge ordered Google to reveal the identity of an anonymous blogger so plaintiff could file a defamation lawsuit).

didn’t mean to say anything bad . . . I said something really stupid on the . . . internet, I typed for literally, like, 12 seconds, and it devastated my life.”

Finally, as previously suggested, the Web site may be set up to allow other users to “rate” the quality and credibility of an anonymous user’s postings, putting other visitors on notice regarding the veracity or reliability of that particular anonymous user’s postings.

C. Notice and Takedown of Certain Postings

Because of the volume of traffic on many Web sites, prescreening is impracticable. Given both the quantity of postings as well as the nature of posts (which are often subjective or difficult to verify), Web site sponsors are ill-equipped to determine the lawfulness of content even after it has come to their attention. Yet, in certain cases, to require Web site sponsors to take down certain posts upon request would not impose an undue burden upon them. This Section argues that Web site sponsors should automatically remove content upon request in three situations: where the request is made by the original poster; where the post is a digital image of a naked person and the takedown request comes from the subject; and where the post is a digital image of a minor and the takedown request comes from the subject’s legal guardian.

1. Takedown Request by Original Poster

As previously discussed, online postings are often made in a heightened emotional state. In some cases, the poster may regret having made an online rant about another individual. The Web site sponsor should be required to remove the post upon request of the original poster. One alleged poster to a popular social networking Web site claimed that she posted information about her former partner that she later regretted. She claimed that the Web site refused to remove the post even after repeated requests. A failure to take down the original post upon

---

126 Margolick, supra note 2.
127 Facebook, for example, reportedly has more than 200 million unique users. See Michael Learmonth, Facebook Crushing MySpace in Traffic, ADVERTISING AGE (2009), http://adage.com/digital/article?article_id=134062. MySpace reportedly has 100 million unique users. Id.
128 See Cindy English, Don’t Date Him Girl! The Lawsuit . . ., http://www.cheating ways.com/cheaters/dont-date-him-girl-the-lawsuit (Apr. 28, 2007) (“They froze my account and will not delete the posting after i emailed 10 times for them to get rid of it. Yes I wrote the posting when i was hurt, angry, sad, I was a mess and i wanted him to hurt as much as I was. We are back together and this website is the only thing that is standing in our way of moving forward. Any suggestions on how to get this posting removed?????? I dont really have the money for an attorney but if it comes down to it i will have to do it but is asking them to remove a posting to much? They froze my account and left the posting! Crazy!!!”) (comment as originally posted by “Jesika”).
129 Id.
request should constitute unreasonable conduct. The Web site sponsor may be ill-positioned to verify the accuracy of posted content, but the poster is not. The poster may uncover additional information that makes the original post misleading or false, or the poster may have had time to reconsider what was originally posted. Unfortunately, the poster is often unable to remove the post without the Web site sponsor’s cooperation. To require Web site sponsors to remove content upon request of the poster does not require Web site sponsors to prescreen or to make subjective determinations of content accuracy. Furthermore, a Web site sponsor can redress the problem of repeat repentant posters by banning them from the site.

2. Takedown of Two Types of Digital Images (Nude Individuals & Nonpublic Figure Minors)

Digital images and videos arguably have the potential to cause the greatest harm to victims of online harassment. Images tend to stick in the minds of viewers. The popular perception is that pictures do not lie and that one picture is worth a thousand words.

On the contrary, images can distort the truth by presenting only part of the story. They can be digitally altered, or they may capture an image without its context, thus subverting its meaning. They can also expose and identify an individual in a way that mere words cannot, thus creating a greater danger to privacy and security.

In many cases, publication of images has social value. They may create an emotional impact that words alone cannot convey. The social benefit of images should be weighed against the unique harmful effects of the Internet. Given the harm of widespread simultaneous and permanent distribution, two types of images should be outright prohibited without written consent. The first category is images of nonpublic figure minors. Images of nonpublic minors should not be uploaded to publicly accessible Web sites without the written consent of their legal guardians. The second category is images of nude individuals who are identifiable.

130 Although minors who are public figures may suffer the same negative emotional consequences, they are arguably better prepared to deal with widespread recognition and notoriety. The public also has a more legitimate “right to know” where the image is of a public figure. Thus, although I may personally wish to extend the takedown upon request measure to images of children of public figures, I recognize that such an application likely would be unacceptable under U.S. laws. By contrast, in many European countries, the images of children of public figures cannot be published without parental consent. See, e.g., Stephen Howard, Rowling Wins Child Privacy Case, DAILY MAIL, May 8, 2008, at 12 (describing author J.K. Rowling’s suit for privacy violations stemming from covert, long-lens photographs of her toddler son); Alex Kingsbury, Washington Whispers, Photo of Obama and Spain’s First Children Causes a Stir, U.S. NEWS & WORLD REPORT, Sept. 25, 2009 (noting that privacy conventions in Spain prevent publication of photos of politicians); John Moore, Photo Oops: U.S. Posts Pic of Spain’s Goth First Daughters, ROCKY MTN. IND., Sept. 26, 2009 (noting that under Spanish law, the prime minister can prevent the media from publishing photos of his children).
Images of identifiable subjects who are completely or partially nude (i.e., the breasts, buttocks, and genitalia of a female subject and the buttocks and genitalia of the male subject) should not be permitted to be uploaded to public Web sites without the written consent of their subject.

Unauthorized postings of either type of image should be subject to immediate takedown after notice. Consents should be filed and maintained with the Web site sponsor for at least one year after the digital image has been removed from the Web site. Unlike other user-supplied content, to require Web site sponsors to take down these two types of images does not require them to make difficult subjective decisions. Whether a person is nude is not difficult to determine. Determining whether a person is younger than eighteen by looking at an image is more difficult. The burden on the Web site sponsor, however, is slight, and the sponsor in close cases might simply take down the image upon request of the individual or guardian, or ask for verification of the individual’s age.

As a recent study indicates, the brain continues to develop during the teenage years. During this period, the frontal cortex, or the “thinking” part, of the brain grows and synaptically prunes itself. The prefrontal cortex also undergoes change, which means that the part of the teenage brain responsible for controlling emotions and empathy is not yet where it will be in a few years. The video of the “Star Wars Kid” illustrates how even nonsexual images of minors can be used to cruel effect. A fourteen-year-old boy made a video of himself swinging a golf ball retriever around as if it were a lightsaber. His classmates uploaded the video to a video-sharing Web site, where it spread virally. The video remains popular and is often accompanied by abusive comments such as “what the hell is going through this kids mind, this kid must take it up the butt every single night,” “he is a fat nerd faggot kid that fights like a loser,” and “sad retarded fat kid. seriously what was he thinking the loser.”

Children are more emotionally vulnerable than adults to cruel behavior, and their peers are more likely to engage in it than are adults. Children have not yet developed a social and professional reputation to counter a negative online image. Furthermore, most children do not yet have the maturity and fortitude to ignore abusive comments accompanying embarrassing or cruel posted images. For example, after becoming the object of online scorn and ridicule, the “Star Wars kid” dropped out of school and enrolled in a children’s psychiatric ward. Given

132 Id.
133 Id.
135 Id.
a Web site’s lack of expressive interest, it should immediately remove an image of
a nonpublic minor upon the request of his or her legal guardian. 138

The prohibition against these two types of images is not as draconian as it
might seem. The images would be permissible and uploadable with authorization
from the subject of the image (or his or her legal guardian). The Web site would
not be required to respond until after a takedown request has been made. Many
news publications and Web sites already have special policies to address children’s
privacy and the publication of nude images139 and the proposal simply mirrors
existing social values. In modern American society, people don’t walk around in
public nude and most people don’t want strangers to see them in the buff.
Similarly, our society understands that children are more vulnerable than adults
and thus require greater protection and paternalism. Furthermore, with the use of
photo editing tools, the poster can easily crop out the prohibited figures from
otherwise permissible images, such as a group photograph. This proposed
prohibition does not prevent the capturing of the image, only its distribution on the
Internet. A photographer could still take pictures of children at the park and sell
them or publish them in a book; he or she could not, however, post them to a
publicly accessible Web site without the consent of the children’s guardians.

V. INCREASING WEB SITE ACCOUNTABILITY

Many types of online harassment might be curtailed or prevented if we altered
our expectations of Web site sponsors. Although debate continues to rage
regarding whether Web sites should be treated as tangible (real or personal)
property,140 legal duties and social norms and values govern the sense of
responsibility that most business owners feel regarding the conduct of patrons on
their physical property. Yet the legal duties and social norms governing the
responsibility of Web site sponsors over the conduct of their users is still evolving

138 The third-party poster can contest removal on First Amendment grounds, but the
image should not remain online during the adjudication process.
139 See, e.g., Roanoke Times News Standards and Policies, ROANOKE TIMES,
http://www.roanoke.com/newsservices/wb/xp-59614 (“We do not have specific
moratoriums against the publication of any type of accurate news picture. Pictures that
show extreme grief, graphic violence, dead people, nudity or other potentially offensive
content require careful consideration before publication. In these cases a photo editor, the
assistant managing editor, the managing editor or the editor must be consulted before
com/privacy (last visited Sept. 1, 2009) (“If a question, comment, story, joke or
opinion is published, only the student’s first name, grade and state/country appear on the
site.”). Newspapers likely adopted such policies to conform to the Children’s Online
140 See Dan L. Burk, The Trouble with Trespass, 4 J. SMALL & EMERGING BUS. L. 27,
39–54 (2000); Kevin Emerson Collins, Cybertrespass and Trespass to Documents, 54
CLEV. ST. L. REV. 41, 41–65 (2006); Jacqueline Lipton, Mixed Metaphors in Cyberspace:
Property in Information and Information Systems, 35 LOY. U. CHI. L.J. 235, 240–44
(2003).
and may be influenced by courts’ broad interpretation of section 230, which grants
Web sites immunity from liability for the conduct of their users. Yet section 230
immunity should not be interpreted to mean that Web sites should have no liability
whatsoever for the businesses they create.141

Arguments disclaiming Web site sponsor responsibility reflect a one-sided
and rather socially irresponsible notion of the role of Web site sponsors. Significantly, this view is at odds with expectations of offline business owners. This Part makes two different arguments for why Web site sponsors should adopt
the proposals set forth in the preceding section. Both arguments appeal to Web site
sponsors’ self-interest. The first argument, which is more carrot than stick, posits
that greater accountability enhances a Web site sponsor’s ability to control its
business and image. The second argument, which is more stick than carrot, contends that Web site sponsors should have, and may already have, liability under
tort law to address online harassment on their Web sites.

A. Encouraging Self-Regulation

Web site sponsors often express reluctance to regulate communication among
users for various reasons, including that such regulation undermines the nature of
their Web sites.142 Many Web sites already have policies in place that mirror some
or most of the proposals set forth in Part IV.143 The primary problem has been in

---

141 See Kim, supra note 21, at 116 (“The immunity that website sponsors . . . have as
publishers should not mean that they have no obligation whatsoever for the activity on their
website.”).
142 See supra note 9 and accompanying text.
143 See supra note 9 and accompanying text.
the reluctance of some Web site sponsors to enforce their own policies. Even Web sites whose very business models appear to encourage harassing behavior have policies prohibiting online harassment.144

Brian Leiter, a law professor and well-known blogger, recently explained the rationale underlying his selective open comments policy:

First, there are likely to be far more anonymous comments, and anonymity generally encourages irresponsible behavior . . . . Second, there would be a lot more spam . . . . Third, the quality of the threads is likely to be much more uneven . . . . I’d rather not have a site bearing my name be the repository for the kind of garbage that is typical on the blogs that do not moderate comments.


144 For example, the terms and conditions of Juicy Campus, a Web site that encouraged and disseminated college campus gossip with a promise of anonymity, formerly stated that users agreed not to post content that is “unlawful, threatening, abusive, tortious, defamatory, obscene, libelous, or invasive of another’s privacy.” See Eugene Volokh, Juicy Campus Lawyer Responds About the New Jersey Attorney General’s Investigation, http://www.volokh.com/posts/1207884421.shtml (Apr. 10, 2008, 23:27). After the Web site was investigated for consumer fraud by the New Jersey attorney general for failing to enforce its own terms and conditions, the Web site changed its terms to expressly disclaim any responsibility for monitoring content:

You acknowledge that JuicyCampus does not pre-screen Content. You agree that JuicyCampus is under no obligation to review all Content, or any Content, on any regular schedule or at all. You agree that JuicyCampus shall have the right (but not the obligation) to re-arrange, remove and/or restrict access to any Content on the Site at any time in its sole discretion, for any reason or for no reason. BY USING THE SITE, YOU AGREE THAT JUICYCAMPUS SHALL HAVE NO OBLIGATION TO MONITOR CONTENT ON THE SITE OR TO DELETE CONTENT FROM THE SITE, EVEN IF JUICYCAMPUS IS NOTIFIED THAT SUCH CONTENT VIOLATES THIS AGREEMENT. . . . User Conduct Guidelines. JUICYCAMPUS RESERVES THE RIGHT, BUT DISCLAIMS ANY OBLIGATION OR RESPONSIBILITY, TO REMOVE ANY CONTENT THAT DOES NOT ADHERE TO THESE GUIDELINES, IN ITS SOLE DISCRETION.
Web site sponsors may refrain from enforcing their policies because they fear user dissatisfaction with content regulation. The founder of AutoAdmit, who controlled the message board, reportedly feared that removing posts would prompt a mass exodus from his Web site, yet he admitted that his failure to do so meant he “lost his website” to “parasites” and “freaks.” The prevailing ethos of the Web is a libertarian one. Any attempt to restrict speech or activity tends to be greeted with cries of censorship, even though the entity seeking to regulate the conduct or the speech is a private actor. For example, a recent video of former Alaska governor and vice presidential candidate Sarah Palin’s former church, the Wasila Assembly of God, was removed from YouTube for “inappropriate content.” Internet users condemned YouTube’s actions as “censorship,” comparing the site’s decision with government censorship in China and advocating retaliatory measures against YouTube. Speech regulation by private entities, however, is not an infringement of free speech. The First Amendment prohibits the government, not private actors, from restricting speech. Yet many Web site users have grown
accustomed to thinking of the whole of the Internet as a public forum, rather than as privatized Web sites, and many view attempts to restrict activity as censorship.\textsuperscript{152} Web site sponsors are in the best position to regulate communication on their Web sites given concerns about governmental restrictions of speech.\textsuperscript{153} While it is popularly assumed that everyone is entitled to communicate on the Internet, everyone is not entitled to communicate everywhere. While the public/private distinction regarding speech has been firmly established in the physical world, there has been a shift in the attitude toward Internet communication, with many arguing that placing restrictions on postings to nongovernmental Web sites amounts to suppression of free speech by Web site sponsors.\textsuperscript{154} For example, AutoAdmit defended its controversial message board by saying it was simply a forum for free speech “where people can express themselves freely, just as if they were to go to a town square and say whatever brilliant or foolish thoughts they have.”\textsuperscript{155} This argument promotes a view that subjects Internet communication to different standards and rules than those that govern offline communication, where private actors are at liberty to regulate speech and conduct on their premises. This Article strongly rejects this sly political stance equating speech regulation on nongovernmental Web sites to censorship.\textsuperscript{156} As private actors, Web site sponsors of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.” U.S. Const. amend. I.

\textsuperscript{152} This may be due, in large part, to the absence of public forum for communication. See Nunziato, supra note 18, at 1117.

\textsuperscript{153} This is not to say that government regulation of content on the Internet would necessarily run afoul of the First Amendment. See Andrew Chin, Making the World Wide Web Safe for Democracy: A Medium-Specific First Amendment Analysis, 19 Hastings Comm. & Ent. L.J. 309, 313 (1997) (explaining how “the structural impact of the World Wide Web . . . on the distribution of power in public discourse may justify intervention by the state”).

\textsuperscript{154} See Branscomb, supra note 20, at 1641 (noting that “netizens” “assert what they call a First Amendment right of unencumbered access to whatever information they deem personally useful or desirable” and that, although it is “not accurate to describe this claim as a First Amendment right, clearly many Internet users’ developing expectation of freely flowing channels of information without censorship by outsiders cannot be ignored”).

\textsuperscript{155} Nakashima, supra note 2, at A1.

\textsuperscript{156} This is not to suggest that government censorship of the Internet is not a concern, only that it is not a concern when Web site sponsors themselves regulate the content. It is less of a concern in the United States than it is in other countries. Foreign government censorship of Web content has been the subject of recent news attention. See Thomas Crampton, World Business Briefing Europe: Turkey: YouTube Blocked Over Content Found Offensive, N.Y. Times, Mar. 8, 2007, at C1 (reporting that a court in Turkey ordered blockage of all access to YouTube after a video appeared on the Web site that was deemed insulting to Mustafa Kemal Ataturk, the founder of modern Turkey); Jane Spencer & Kevin J. Delaney, YouTube Unplugged: As Foreign Governments Block Sensitive Content, Video Site Must Pick Between Bending to Censorship, Doing Business, Wall St. J., Mar. 21, 2008, at B1 (discussing China and Turkey’s ban on access to YouTube). China banned
have greater freedom to shape social interactions on the Internet and to encourage behaviors that reflect prevailing physical-world social norms. To avoid this responsibility and to ignore that Internet norms are now being created is to default, Lord of the Flies-style, to the standards of conduct set by the trolls of the Internet.

The misperception regarding the private nature of Web sites may be partly responsible for the “anything goes” culture that prevails on some Web sites on the part of some users. Because some users regard the ability to post and participate on a Web site as a right, rather than a privilege, they may engage in conduct that ultimately damages the reputation and image of the Web site. MySpace, for example, received much negative public attention from a cyber deception incident that resulted in the suicide of a teenage girl. YouTube, eBay, and craigslist regularly fend off infringement claims arising from illicit content uploaded by users. Facebook struggles with trolls who misuse information posted on members’ pages.

Elevating expectations of Web site sponsor accountability adjusts user expectations of who controls Web site activity and content, and enhances the Web site sponsor’s ability to control its image and brand. The guidelines and standards

access to YouTube after video clips showing Tibetan monks being dragged through the streets by Chinese soldiers appeared on the site. Id.

As Danielle Citron observes, “[i]f we believe that the Internet is, and should remain, a Wild West with incivility and brutality as the norm, then those who are impervious to such conduct will remain online while the vulnerable may not. To that end, we may get more bull-headed, impervious posters and fewer thoughtful ones.” Citron, supra note 12, at 105.

See Schwartz, supra note 39, at 26 (defining a cyberspace “troll” as one who “intentionally disrupts online communities”).

See Susan Duclos, Indictment Handed Down in MySpace Hoax That Caused Child to Commit Suicide, DIGITAL J., May 16, 2008, http://www.digitaljournal.com/article/254805 (noting how “information obtained over the MySpace computer system [was used] to torment, harass, humiliate, and embarrass the juvenile MySpace member”).


Facebook’s policies expressly state that Facebook has no responsibility for misuse of user content by other users. See Facebook Privacy Policy, Facebook Principles, http://www.facebook.com/policy.php (last visited Sept. 1, 2009) (“You post User Content (as defined in the Facebook Terms of Use) on the Site at your own risk. Although we allow you to set privacy options that limit access to your pages, please be aware that no security measures are perfect or impenetrable. We cannot control the actions of other Users with whom you may choose to share your pages and information. Therefore, we cannot and do not guarantee that User Content you post on the Site will not be viewed by unauthorized persons.”).
established by Web sites in their user agreements and policies might then be taken more seriously by users as contractual requirements. For example, regardless of whether she read MySpace’s terms of use, Lori Drew knew that posing as a sixteen-year-old boy to deceive and manipulate another MySpace user was against the company’s policies, yet she did so.\textsuperscript{162} An extensive user policy prohibits eBay members from selling items that violate third-party rights.\textsuperscript{163} Although eBay may have instituted this policy to minimize its legal liability, its policy also ensures that it is viewed as a legitimate and reputable business. YouTube has a policy forbidding graphic violence, sexually explicit images, and “bad stuff like animal abuse, drug abuse, under-age drinking and smoking, or bomb making.”\textsuperscript{164} This policy helps set user expectations of what is permissible on the site and also of when YouTube will remove site content.\textsuperscript{165} By explaining its policies and managing expectations upfront, YouTube is better able to protect its image and brand.

Web site image and branding, in turn, affect mainstream acceptability and corporate sponsorship. For example, rather than being known as a site for porn video clips, YouTube is known for quirky and humorous clips on a wide range of topics, making it a more suitable site for advertisers. Juicy Campus went out of business in February 2009 due to a lack of advertising revenue.\textsuperscript{166} Whether the lack of advertising revenue was due to the negative publicity surrounding the Web site and an investigation by the New Jersey attorney general on consumer fraud charges, or simply the result of the economic downturn, was the subject of much speculation.\textsuperscript{167}

Marketed in the right way, some anti-cyberharassment measures may be viewed as features or Web site advantages. While some users may resist any imposition of restraints, others may appreciate their potential for deterring impulsive behavior. For example, Google recently introduced an optional Gmail feature called “Mail Goggles,” which incorporates contractual restraints to prevent embarrassing user behavior.\textsuperscript{168} The feature prompts a user with a pop-up window


\textsuperscript{165} See \textit{id}.


that asks, “Are you sure you want to send this?” if the user tries to send e-mail during certain times when the user is more likely to be inebriated (i.e., late night to early morning).169 Users are then required to solve several math problems before Google will permit the message to be sent.170

Web site sponsors should voluntarily adopt online harassment policies before Congress mandates them.171 In the aftermath of the MySpace case, legislators have proposed laws that would criminalize cyberharassment,172 although some fear that some of the proposed laws are overbroad.173 Several commentators have suggested that section 230 of the CDA should be amended so Web site sponsors receive a type of limited immunity based upon lack of actual notice, rather than absolute immunity.174 Others, however, have noted that a notice-and-takedown scheme may put Web site sponsors in the awkward position of having to make legal determinations of what constitutes defamatory or otherwise tortious material.175 Although the focus of this Article is on private law approaches to online harassment, it is worth noting that the success of any such approaches may have the effect of curbing government regulation.

My suggested anti-cyberharassment proposals are intended to deter impulsive, regrettable behavior by forcing posters to think before pressing “send.” The effect may be to evaluate and filter speech, but it is the poster, not a government actor, who is censoring communication. The proposals provide an opportunity for contemplation without imposing a ban on speech, leaving the decision whether to

---

169 Id.
170 Id.
171 Congress should also revisit section 230 of the Communications Decency Act to determine whether the definition of “interactive computer service” should be limited to Internet service providers and not Web site sponsors, and/or whether to qualify immunity. An interactive computer service is defined as “any information service, system, or access software provider that provides or enables computer access by multiple users to a computer server, including specifically a service or system that provides access to the Internet and such systems operated or services operated by libraries or educational institutions.” 47 U.S.C. § 230 (2006).
173 Mark “Rizzn” Hopkins, Should There be a Law Against Asshats Like Me?, http://mashable.com/2008/06/09/asshats/ (June 9, 2008) (fearing that pending legislation will have the effect of criminalizing posts criticizing celebrities and mainstream media).
174 See Solove, supra note 74, at 154 (arguing that once a Web site is notified about a cyberharassment problem, it should respond to the problem or be liable); see also Bradley A. Areheart, Regulating Cyberbullies Through Notice-Based Liability, 117 YALE L.J. POCKET PART 41, 43 (2007), available at http://thepocketpart.org/2007/09/08/areheart.html (arguing for ISP liability based upon a notice and takedown scheme based on actual notice).
175 See Citron, supra note 12, at 122; Lemley, infra note 182, at 801–02 (noting the problems with the copyright notice and takedown regime and proposing one based upon the trademark immunity statute).
post the communication in the hands of the poster. The government could arguably adopt even more onerous content-neutral restrictions without violating free speech principles.\(^{176}\) The government has chosen not to regulate the Internet, but that does not mean that if it were to do so its actions would be unconstitutional.\(^{177}\)

Although many Web sites work to minimize online harassment, some Web sites actively encourage it.\(^{178}\) Rather than seeking a more commercially acceptable image, these Web sites craft an alternative niche for themselves as a place where users can malign others.\(^{179}\) Although carrot-like incentives may work for those Web sites striving for mass market acceptance, some Web sites may need a more stick-like measure.

### B. Imposing Proprietorship Liability on Web Site Sponsors

One stick-like measure may be the imposition of tort liability on Web site sponsors. In a prior essay, I argued that proprietorship liability should be imposed upon Web site sponsors.\(^{180}\) I analogized the duty of Web site sponsors to brick-and-mortar businesses and advocated the imposition of liability similar to the “premises” liability imposed on offline businesses. The analogy to premises liability is not a perfect one given the differences between the Internet and the physical world, including the inability to draw secure boundaries and screen for potential harm. Nevertheless, the important similarity is that offline and online business owners establish, control, and benefit from their businesses. Web site sponsors are proprietors who exercise control over their businesses in many ways. Web site sponsors enforce their proprietorship over their Web sites by establishing terms of use in clickwrap or browsewrap agreements. Web sites may capitalize on their proprietorship by selling user information to advertisers. They may sell advertising space or products, such as T-shirts, directly on their Web sites. They may receive a percentage of revenues that their users receive from the sale of products on their Web sites. They may also profit indirectly by using their Web sites as marketing platforms to reach a broad audience. They may then sell

---

176 One such proposal is the Internet Community Ports Act, which would entail assigning ranges of “ports” or channels for information transmission to different purposes. See Cheryl B. Preston, *Making Family-Friendly Internet a Reality: The Internet Community Ports Act*, 2007 BYU L. Rev. 1471, 1475–78; Cheryl B. Preston, *Zoning the Internet: A New Approach to Protecting Children Online*, 2007 BYU L. Rev. 1417, 1468–69.

177 See Tushnet, *supra* note 15, at 988 (arguing that “Congress is free, within rather broad limits, to determine an appropriate intermediary liability regime”).

178 See *supra* note 9 and accompanying text.


180 Kim, *supra* note 21, at 116.
ancillary products or services, such as books or consulting services, to this audience.

Although some scholars and commentators argue that Web sites are property and should be treated as such, others disagree. Although those who disagree that Web sites are “just like” property must recognize that Web site sponsors maintain control over the content of the site (regardless of whether they choose to exercise that control) and are in the best position to prevent harm to other users on the site. Of course, any discussion of Web site sponsor liability must recognize the difficulties inherent in applying laws and norms developed with the physical world in mind to cyberspace, and proposed remedies must address those difficulties. Yet many of the criticisms of the Web-site-as-property view pertain to the difficulties of delineating boundaries of intangible works. These concerns are inapposite where the activity at issue is limited to what happens on the Web site.

This Article bypasses the complex issue of whether Web site ownership and intellectual property generally are akin to tangible property ownership for fear of distracting from the Article’s central issues. The topic has been discussed at length elsewhere. See Frank H. Easterbrook, Intellectual Property is Still Property, 13 HARV. J.L. & PUB. POL’Y 108, 111 (1990); I. Trotter Hardy, Not So Different: Tangible, Intangible, Digital, and Analog Works and Their Comparison for Copyright Purposes, 26 U. DAYTON L. REV. 211, 213 (2001) (asserting that “[t]hese assumptions of differences [between intangible intellectual property and tangible property] are wrong”); I. Trotter Hardy, Property (and Copyright) in Cyberspace, 1996 U. CHI. LEGAL F. 217, 246 (1996) (arguing that “it seems no harder to identify an informational work in cyberspace as a unit with “boundary lines” than it is to identify a similar informational work elsewhere”); Jacqueline Lipton, Mixed Metaphors in Cyberspace: Property in Information and Information Systems, 35 LOY. U. CHI. L.J. 235, 240–244 (2003) (critiquing the use of property metaphors in cyberspace); Michael J. Madison, Rights of Access and the Shape of the Internet, 44 B.C. L. REV. 433, 464–71 (2003) (explaining why trespass to intangibles is significantly different from trespass to chattels). But see Burk, supra note 140, at 28 (noting that proprietary interest in the Internet has “only the most tenuous of antecedents in the law of chattels”); Mark A. Lemley, Property, Intellectual Property, and Free Riding, 83 TEX. L. REV. 1031, 1032 (2004) (contending that “full internalization of positive externalities is not a proper goal of tangible property rights except in unusual circumstances”). There is a split of case authority on whether to treat Web sites and computer servers as “chattels.” See eBay Inc. v. Bidder’s Edge., Inc., 100 F. Supp. 2d 1058, 1073 (N.D. Cal. 2000) (holding that eBay’s servers were private property and finding that unauthorized access was a trespass); Register.com, Inc. v. Verio, Inc., 126 F. Supp. 2d 238, 248–51 (S.D.N.Y. 2000) (holding that an automated software robot searching a database without permission constituted a trespass to chattels). Contra Intel Corp. v. Hamidi, 71 P.3d 296, 311 (Cal. 2003) (holding that the tort of trespass to chattels did not encompass electronic communication that did not damage or impair the computer system).

182 YetMany of the criticisms of the Web-site-as-property view pertain to the difficulties of delineating boundaries of intangible works. These concerns are inapposite where the activity at issue is limited to what happens on the Web site.
Sponsors of publicly accessible Web sites are “proprietors,” whether the Web site is a blog, a gossip site, or a retail site. Along with ownership comes responsibility. Because the Web sites are publicly accessible (again, this Article is limited to content only on publicly accessible Web sites), they are more akin to businesses than to private residences. Although some sites sell products and are clearly “for profit,” other sites are less clearly commercial. Yet even sites that are not obviously retail-oriented have the ability to monetize their content in some way, such as by selling ancillary products like T-shirts, selling advertising or user data, or by using the site as a marketing vehicle to sell products, such as a novel, or services, such as consulting.\(^{183}\) In some cases, a Web site may intend to generate a large readership in the hopes of eventually selling out to a larger commercial entity. This Article uses the term “propriorship” or “proprietary interest” to refer to a Web site sponsor’s ability to capitalize upon the activity on a Web site, regardless of whether it in fact chooses to do so. It thus avoids the larger question of whether Web sites are the “property” of the Web site sponsor.

Tort law places upon a “possessor of land” or “premises occupier” a duty to exercise reasonable care to avoid foreseeable harm caused by the accidental, negligent, or intentionally harmful acts of third parties.\(^{184}\) This duty is not to ensure the safety of invitees, but to take “reasonable measures” to control the conduct of third parties, or to give adequate warning to enable invitees to avoid harm.\(^{185}\) Accordingly, the owner is liable for negligence only where the owner failed to take reasonable care to discover the occurrence of dangerous conduct by third parties, or where the owner failed to exercise reasonable care to provide appropriate precautions. For example, an owner may be found to have failed to exercise

---


\(^{184}\) The Restatement (Second) of Torts states:

> A possessor of land who holds it open to the public for entry for his business purposes is subject to liability to members of the public while they are upon the land for such a purpose, for physical harm caused by the accidental, negligent, or intentionally harmful acts of third persons or animals, and by the failure of the possessor to exercise reasonable care to (a) discover that such acts are being done or are likely to be done, or (b) give a warning adequate to enable the visitors to avoid the harm, or otherwise to protect them against it.

See *RESTATEMENT (SECOND) OF Torts* § 344 (1977).

\(^{185}\) Murphy v. Penn Fruit Co., 418 A.2d 480, 482 (Pa. Super. Ct. 1980); Exxon Corp. v. Tidwell, 867 S.W. 2d 19, 21 (Tex. 1993) (noting that a landowner has “a duty to protect invitees on the premises from criminal acts of third parties if the landowner knows or has reason to know of an unreasonable risk of harm to the invitee”).
reasonable care for failure to provide security personnel or adequate lighting in a parking lot. A California court held that the owners of a restaurant could be found negligent for failing to protect customers where the layout of the restaurant required customers to stand in front of parking spaces with low barriers.

The critical factor in determining whether there was a duty of care (a prerequisite to a finding of negligence) is foreseeability—whether the business owner knew or should have known that the harm was likely to occur on the premises. For example, in one case, a tavern customer verbally threatened the plaintiff and was escorted out of the tavern. The tavern employees then allowed the abusive customer to re-enter the tavern, and the customer attacked the plaintiff. The court found that the tavern had a duty to exercise reasonable care to control the customer’s conduct to prevent harm to the plaintiff, and that the duty arose “by reason of the defendant’s knowledge that an assault . . . was ‘about to occur’” in the tavern.

Two different tests determine foreseeability. The first is the “prior similar incidents” test, under which a duty arises when incidents similar to the harm at issue should have put the business owner on notice. The second test is the “totality of the circumstances” test, whereby the court examines not just whether there were prior similar incidents, but also factors such as whether the business was located in a high-crime area. The Restatement (Second) of Torts also considers the character of the business in determining foreseeability:

Since the possessor is not an insurer of the visitor’s safety, he is ordinarily under no duty to exercise any care until he knows or has reason to know that the acts of the third person are occurring, or are about to occur. He may, however, know or have reason to know, from past experience, that there is a likelihood of conduct on the part of third persons in general which is likely to endanger the safety of the visitor, even though he has no reason to expect it on the part of any particular

---

186 See Murphy, 418 A.2d at 482; see also RESTATEMENT (SECOND) TORTS § 344 (1977).
188 But see Michael J. Yelnosky, Business Inviters’ Duty to Protect Invitees from Criminal Acts, 134 U. PA. L. REV. 883, 883–84 (1986) (arguing that courts should adopt an unqualified duty-to-protect rule that would require all business inviters to take reasonable steps to prevent crime on their premises).
190 Id.
191 Id. at 658; see also Bartosh v. Banning, 251 Cal. App. 2d 378, 384 (1967) (holding that one who operates a bar “must act as a reasonable man to avoid harm from the negligence of other persons who have entered the premises or even from intentional attacks on the part of such third persons”).
193 Id. at 548–49.
194 Id.
individual. If the place or character of his business, or his past experience, is such that he should reasonably anticipate careless or criminal conduct on the part of third persons, either generally or at some particular time, he may be under a duty to take precautions against it, and to provide a reasonably sufficient number of servants to afford a reasonable protection.\footnote{Restatement (Second) of Torts § 344 cmt. f (1977); see also Barker v. Wah Low, 97 Cal. Rptr. 85, 88 (1971) (noting that Restatement principles “have been recognized and applied” in California). Some courts have expressly rejected the Restatement view that the character of the defendant’s business may be considered in a foreseeability analysis. See Timberwalk Apartments, Partners, Inc. v. Cain, 972 S.W.2d 749, 758–59 (Tex. 1998).}

A duty analogous to possessors of land may be imposed upon Web site sponsors who fail to exercise reasonable care to provide appropriate precautions to prevent or minimize online harassment on their Web sites. This duty would arise where the Web site sponsor has notice of harm or the likelihood of harm to invitees by third parties. The Web site sponsor may have notice either because of similar prior incidents or because the nature of the Web site made such harm foreseeable.\footnote{See Doe v. MySpace, 474 F. Supp. 2d 843, 851 (W.D. Tex. 2007) (“The court declines to extend premises liability cases to the internet context, especially where, as here, the Defendant provides its services to users for free.”); see also Goddard v. Google, Inc., No. C 08-2738 JF, 2008 WL 5245490, at *4–5 (N.D. Cal. Dec. 17, 2008) (citing to Doe v. MySpace to reject characterization of claim as one of receiving “tainted funds” rather than hosting content).}

In Doe v. MySpace, the court expressly declined to apply premises-based liability to the Internet context.\footnote{Doe v. MySpace, 528 F.3d 413, 416–17 (5th Cir. 2008).} There, the minor plaintiff sued the online social networking site MySpace after she was sexually assaulted.\footnote{Id. at 416.} One of the plaintiff’s claims contended that MySpace failed to implement basic safety measures to prevent sexual predators from communicating with minors on MySpace.\footnote{Id. at 418–20.} The court rejected the argument, stating the plaintiff cited no precedent for treating a Web site as a virtual premises.\footnote{Id. at 418.} Yet the court’s opinion reveals it was primarily concerned that the remedy the plaintiff sought would impose an undue burden on MySpace’s business:

Plaintiffs allege MySpace can be liable under a negligence standard when a minor is harmed after wrongfully stating her age, communicating with an adult, and publishing her personal information. To impose a duty under these circumstances for MySpace to confirm or determine the age of each applicant, with liability resulting from negligence in performing or not performing that duty, would of course stop MySpace’s business in
its tracks and close this avenue of communication, which Congress in its wisdom has decided to protect.\textsuperscript{200}

In other words, the court appears to base its decision on the reasonableness of MySpace’s safety procedures given the vast amount of traffic on its Web site, and the minor’s knowing violation of those procedures, instead of explaining why premises-based liability is wholly inapplicable on the Internet.\textsuperscript{201} But consider this: if MySpace had no safeguards,\textsuperscript{202} and 50 percent of teens using MySpace had been assaulted by sexual predators (even though they accurately stated their age during registration), would the court have reached the same conclusion? Taking the court’s statements regarding blanket immunity at face value, MySpace would have no obligation to improve its business model or adopt safety measures to protect further assaults because it was merely the “publisher” of communications between the minor victim and the assailant.\textsuperscript{203} Yet such a conclusion would be socially unacceptable.

Given the difference between physical world business premises and Web site premises, the analogy of “possessors of land” to Web site sponsors is limited to just that—an analogy. A theory of liability based upon Web site proprietorship must recognize the differences between Web-based businesses and businesses with physical locations.

Internet-based businesses may have millions of weekly site visitors, compared with brick-and-mortar businesses, which may have dozens; accordingly, the ability to successfully police virtual premises may be more elusive. A determination of what constitutes reasonable proprietorship conduct should take into account the vast amount of Web site traffic. While one attack in an underground parking lot may suffice to put a brick-and-mortar store on notice of the existence of a harmful condition, and thus create an obligation to remedy the condition, one harmful incident of online harassment would not be enough to establish foreseeability on a Web site with thousands of daily postings.

\textsuperscript{200} MySpace, 474 F. Supp. 2d at 851.

\textsuperscript{201} The district court merely “decline[d] to extend premises liability cases to the internet context particularly where, as here, the Defendant provides its services to users for free. Plaintiff has cited no case law indicating that the duty of a premises owner should extend to a website as a ‘virtual premises.’” Id.

\textsuperscript{202} MySpace sets the default for users age fourteen and fifteen at “private” rather than “public” so that their profiles are not searchable or viewable by anyone other than their named friends. Joint Statement on Key Principles of Social Networking Sites Safety, http://www.ag.state.mn.us/PDF/PressReleases/SocialNetworkingSitesSafety.pdf (last visited Sept. 1, 2009).

\textsuperscript{203} MySpace, 474 F. Supp. 2d at 849 (“It is quite obvious the underlying basis of Plaintiffs’ claims is that, through posting on MySpace, Pete Solis and Julie Doe met and exchanged personal information which eventually led to an in-person meeting and the sexual assault of Julie Doe. . . . No matter how artfully Plaintiffs seek to plead their claims, the Court views Plaintiffs’ claims as directed toward MySpace in its publishing, editorial, and/or screening capabilities.”).
Although the cases addressing business owner liability concerned physical harm to invitees, there is nothing in the underlying rationale of these cases that would preclude nonphysical harm. It makes sense to permit recovery for incorporeal injuries that are unique to, and arise from, the incorporeal nature of the business premises. Furthermore, where nonphysical harm, such as defamation, is conducted on physical premises, plaintiffs can sue defendants directly for republication or dissemination of defamatory matter and have no need to rely upon a premises liability theory. But this remedy is unavailable to plaintiffs alleging online harassment because of the immunity granted to Web site sponsors under the Communications Decency Act.

Businesses have also been held liable for their conduct and business activities, not simply their control over premises. For example, business owners have been found liable for injuries caused by third parties during the course of promotional activities. In one case, the defendant planned to drop table tennis balls from an airplane as part of a promotion. Each ball contained a certificate entitling the holder to a prize from the defendant store. The plaintiff went to the site of the promotion and was injured by the crowd that rushed to retrieve the fallen balls. The Supreme Court of Alabama held that “when a proprietor or storekeeper causes a crowd of people to assemble pursuant to a promotional activity, then that person owes a duty to exercise reasonable care commensurate with foreseeable danger or injury to protect those assembled from injuries resulting from the . . . crowd[.]”

A business may be found liable even where the promotional activities occur off-site. For example, a radio station was held liable for the wrongful death caused

---

204 See Hellar v. Bianco, 244 P.2d 757, 759 (Cal. Ct. App. 1952) (holding tavern owner could be held liable for republication of defamatory matter regarding plaintiff by failing to remove such matter from men’s room wall after having reasonable opportunity to do so); Fogg v. Boston & L.R. Co., 20 N.E. 109, 109–10 (Mass. 1889) (defendant railroad held liable for posting in office libelous newspaper extract about plaintiff railroad broker for forty days). Describing what constitutes publication, the Restatement (Second) of Torts states:

(1) Publication of defamatory matter is its communication intentionally or by a negligent act to one other than the person defamed.
(2) One who intentionally and unreasonably fails to remove defamatory matter that he knows to be exhibited on land or chattels in his possession or under his control is subject to liability for its continued publication.

RESTATEMENT (SECOND) OF TORTS § 577 (1977). But cf. Scott v. Hull, 259 N.E.2d 160, 161 (Ohio Ct. App. 1970) (finding failure of building owner or agent to remove defamatory graffiti on the outside of building was not grounds for defamation suit because building owner and agent did not engage in “positive acts,” such as inviting public into premises).

206 Id. at 68.
207 Id.
208 Id. at 71.
by a radio station listener who was participating in a promotional activity. The radio station conducted a contest that rewarded the first contestant to locate a traveling disc jockey. Two minors, in pursuit of the disc jockey, negligently forced the decedent’s car off the highway. In determining whether the defendant radio station owed a duty to the decedent arising out of its broadcast of the giveaway contest, the California Supreme Court noted that a “number of considerations may justify the imposition of duty in particular circumstances, including the guidance of history, our continually refined concepts of morals and justice, the convenience of the rule, and social judgment as to where the loss should fall.”

As previously mentioned, Web site sponsors under section 230 are immune from liability for content posted by users on their Web sites. Congress, in passing this legislation, intended to encourage the development of the Internet and to protect “good Samaritan” ISPs from liability for blocking or screening obscene material. Court decisions, however, have applied the immunity

\[\text{See Weirum v. RKO Gen., 539 P.2d 36, 45–47 (Cal. 1975).}\]
\[\text{Id. at 37.}\]
\[\text{Id.}\]
\[\text{Id. at 39.}\]
\[\text{See supra notes 13–15 and accompanying text.}\]
\[\text{See Communications Decency Act, 47 U.S.C. § 230(b)(1) (2006) (“It is the policy of the United States: (1) to promote the continued development of the Internet and other interactive computer services and other interactive media.”).}\]
\[\text{Id. Section 230 further provides:}\]

(c) Protection for “good samaritan” blocking and screening of offensive material

(1) Treatment of publisher or speaker

No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.

(2) Civil liability

No provider or user of an interactive computer services shall be held liable on account of—

(A) any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or

(B) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1).

\[\text{Id. The Ninth Circuit recently elaborated that in passing section 230, Congress intended to allow interactive computer services to “perform some editing on user-generated content without thereby becoming liable for all defamatory or otherwise unlawful messages that they didn’t edit or delete. In other words, Congress sought to immunize the removal of}\]
provision too broadly.216 As a consequence, the very section intended to protect
Web sites that screen and remove offensive content is now being used as a shield
by Web sites to actively encourage the posting of such content.217 For example, the
founder of Juicy Campus, a now-defunct Web site that encouraged its college
student users to post gossip, stated on the site blog: “Juicy Campus is the provider
of an interactive computer service.”218 He cited section 230 to support his claims
that “Juicy Campus is immune from liability from content posted by users.”219
Another Web site encourages users to anonymously post gossip about others, and
even suggests they create profiles for other people.220 The purpose of one Web site
is specifically to provide a forum where women can post negative information
about the men they have dated.221 Other Web sites encourage users to post sexually
graphic videos and photographs of their ex-lovers.

The imposition of proprietorship liability on Web site sponsors furthers the
legislative intent of section 230. The intent underlying section 230 immunity, at
least as interpreted by courts, is to both permit Web site sponsors to monitor
content and relieve them of the burden of doing so.222 This intent is understandable
given that, at least on some Web sites, the amount of content would make
screening and filtering an onerous responsibility.223 The imposition of

---

216 See, e.g., Zeran v. America Online, Inc., 129 F.3d 327, 332 (4th Cir. 1997)
(holding that the Communications Decency Act immunized interactive computer service
provider that hosted message board, even though it refused to remove false statement after
notice); see also Lyrissa Barnett Lidsky, Silencing John Doe: Defamation and Discourse in
230(c) have broadened its ambit far beyond merely protecting ‘Good Samaritan’ editorial
control. As interpreted, section 230 gives ISPs complete immunity from liability for
defamatory content initiated by third parties, even if the ISP consciously decides to
republish the defamatory content. The practical effect of these interpretations of section
230 of the CDA is to leave Internet defamation victims with no deep pocket to sue. The
defamed plaintiff can no longer sue the intermediary who republished a defamatory
communication. Instead, the plaintiff must go to the source and sue the person who
originated the defamatory communication, even if that person is an unknown John Doe.”).

217 See supra notes 13–15 and accompanying text.

218 Matt Ivester, Hate Isn’t Juicy. A Letter from the Founder of JuicyCampus.com,

219 Id.

(“On GossipReport.com you can anonymously talk about anyone you want. Instead of
creating a profile about yourself, you can create a profile about someone else. Get in the
loop. Go Gossip!”).

221 DonDateHimGirl.com, http://dontdatehimgirl.com/home/ (last visited Sept. 1,
2009).

222 See supra notes 13–15 and accompanying text.

223 See Brief of Defendant-Appellee at 4, Chi. Lawyers’ Comm’n for Civil Rights
Under Law, Inc. v. Craigslist, Inc., 519 F.3d 666 (7th Cir. Dec. 4, 2007) (No. 07-1101)
proprietorship-based tort liability would not mean an obligation to screen or filter content, as the Web site sponsor would not be liable for information posted by third parties. Nor would it require Web site sponsors to determine whether a particular post is defamatory or otherwise criminal or tortious. While it may not be reasonable to expect a Web site sponsor to filter numerous messages prior to posting, it is reasonable to expect it to establish and enforce policies to discourage online harassment. The standard for liability would be negligence based upon the nature of the business (including volume of traffic), not strict liability. Negligence on the part of the Web site sponsor then would mean that it failed to take reasonable steps to prevent or deter foreseeable online harassment.

The determination of reasonableness and foreseeability should consider the nature of the Web site, including the number of daily visitors and the size of the business. A leanly staffed business such as craigslist, for example, which receives millions of weekly visitors, should not be expected to prescreen or to make content-based decisions regarding user postings. Web sites like Dontdatehimgirl.com (whose motto is “don’t date him until you check him out first”) or EncyclopediaDramatica (whose motto is “in lulz we trust”), on the other hand, that actively encourage users to anonymously and impulsively (i.e., without a cooling period or registration requirement) post inflammatory material about private individuals, are merely exploiting their section 230 immunity.

Given Congress’s objectives in implementing section 230 and the vast amount of traffic on many Web sites, Web site sponsors generally should not be required to implement procedures to prescreen or make subjective determinations of the lawfulness of user-supplied content. A Web site sponsor, however, should be liable for the harm that results from its negligence in setting up its Web site and for a business model that fails to incorporate reasonable steps or safeguards to prevent or minimize foreseeable harassment.

To establish reasonable conduct, Web site sponsors should adopt anti-cyberharassment policies and procedures, such as those set forth in Part IV, to prevent or reduce the likelihood of online harassment. A complete absence of such policies or a business model that encourages harassing behavior would indicate a breach of this duty. Similarly, a failure to enforce anti-cyberharassment policies might indicate negligence.

---

224 Craigslist, http://www.craigslist.org/about/factsheet (last visited Sept. 1, 2009) (stating that craigslist has thirty employees).
225 Id. (stating that craigslist receives more than twenty billion page views each month, and that fifty million people (including forty million in the United States) use the Web site’s services each month).
In addition, a Web site should be liable for its conduct in responding to harassment complaints. For example, one of the female law students in the AutoAdmit case e-mailed one of the Web site’s founders and requested that he take down certain offensive posts about her.\(^{228}\) He responded in an AutoAdmit post and warned that if he kept receiving similar requests, he would post them on the message board.\(^{229}\) While not liable for the content of users’ posts, AutoAdmit should be liable for its own conduct and for responding to the takedown request in an unreasonable manner (i.e., through intimidation and public humiliation). Another Web site reputedly\(^{230}\) asks requestors to submit a fee before considering any takedown request.\(^{231}\) In addition, the appeal process is allegedly a sham, and the requesting party’s pleas for mercy are uploaded onto the Web site for further ridicule and “lulz.”\(^{232}\) The plaintiff would have to establish harm before charging that the Web site sponsor’s negligence was responsible for the harm; however, it would be the Web site’s own actions for which it would be liable. The possibility of being subject to tort liability may prod Web site sponsors who are otherwise unwilling to self-regulate into adopting policies and procedures to reduce or eliminate online harassment.

C. Imposing Proprietorship Liability Conforms to Objectives of Tort Law

The call for greater Web site sponsor accountability is a call for a normative shift in the way we view online harassment. Requiring a certain level of accountability on the part of Web site sponsors is not particularly shocking. They are currently held accountable to a certain extent for copyrighted materials on their site.\(^{233}\) Web site sponsors adopt a hypocritical position by claiming their Web sites are public forums and that they are constrained by free speech concerns from taking any action to prevent online harassment, while at the same time treating their Web sites as “private property” for commercial gain by, for example, selling

\(^{228}\) Margolick, supra note 2.

\(^{229}\) Id.

\(^{230}\) The author received an anonymous letter from a victim of the Web site regarding the takedown appeal process at this particular Web site. The author was unable to confirm the takedown appeal process at this particular Web site without submitting a fee and undergoing it herself, which she for obvious reasons was unwilling to do.

\(^{231}\) Id.

\(^{232}\) “Lulz” is the term used to describe laughing at another’s expense. Mattathias Schwartz, The Trolls Among Us, N.Y. TIMES, Aug. 3, 2008, at MM24 (defining “lulz” as “[a] corruption of ‘LOL’ or ‘laugh out loud,’ ‘lulz’ means the joy of disrupting another’s emotional equilibrium”).

\(^{233}\) Web sites may take down material if they are served with a takedown notice under the Digital Millennium Copyright Act (“DMCA”) that the material is infringing rather than waiting for a court to definitively make such a decision. See Vernor v. Autodesk, Inc., 555 F. Supp. 2d 1164, 1165 (W.D. Wash. 2008) (plaintiff sued Autodesk because it sent an infringement notice to eBay, where the plaintiff was selling copies of Autodesk software products, and eBay suspended the auction in response to the notice and eventually suspended the plaintiff’s eBay account without waiting for a court to resolve the matter).
advertising. Of course Web site sponsors are at liberty to cater to users’ preferences and desires, but in doing so they should be held liable for users’ abuses. The business models of certain Web sites are specifically intended to encourage behavior that is likely to result in online harassment. One Web site encourages posters to submit gossip and states not only that all posts are anonymous but that users can create profiles of other individuals and post using those profiles. It encourages posters to submit “pics and videos to really tell the story!” These Web sites should be viewed as having made a calculated business decision to permit “high risk” activity on their Web sites, and their liability should reflect that calculation. Web site sponsors that encourage harmful behavior, even if they do so slyly, should be held accountable for the ill effects resulting from their business models. On the other hand, Web site sponsors that have implemented policies to deal with online harassment and that enforce such policies should not be held liable for the conduct of their users.

The imposition of proprietorship liability upon Web site sponsors furthers the objectives of tort law. It deters antisocial conduct and compensates those injured by such conduct. It allocates the risk of injury to the party in the best position to avoid its occurrence and absorb the loss. The burden on the Web site sponsor in adopting an online harassment policy is less than the likelihood of injury from failure to adopt such a policy, especially given that the duty does not arise unless the injury was foreseeable.

Some critics may argue that my proposed model of liability lacks defined parameters and will make it difficult for Web site sponsors to ascertain what steps they must take to avoid liability. As a result, the argument goes, innovation may suffer as businesses decline to pursue new ventures for fear of being sued. But entrepreneurship has never come with guarantees, and imposing tort liability upon Web site sponsors is no different from tort liability imposed upon offline

---

234 For example, Juicy Campus’s slogan was “C’mon. Give us the juice. Posts are totally 100 percent anonymous.” See Alexandria Phillips, College Web Page Spreads Rumors, THE TRIPLET.ORG, Jan. 16, 2009, http://media.wwt.com/thetriangle.org/ (in search box, enter title of article). The New York Times described Juicy Campus as “a dorm bathroom wall writ large, one that anyone with Internet access can read from and post to.” Richard Morgan, A Crash Course in Online Gossip, N.Y. TIMES, Mar. 16, 2008, at ST 7.

235 See supra note 220.

236 Id.

237 See Doug Lichman & Eric Posner, Holding Internet Service Providers Accountable, 12 SUP. CT. ECON. REV. 221, 221–28 (2006) (arguing that immunity for intermediaries is difficult to defend on policy grounds and is inconsistent with conventional tort law principles).


239 Lawrence v. Bainbridge Apartments, 957 S.W.2d 400, 405 (Mo. Ct. App. 1997) (noting the objective of tort law is to place the cost of the injury on the party in the best position to avoid the risk and absorb the loss).

240 This is an articulation of the formula, B<PL, famously stated by Learned Hand in United States v. Carroll Towing Co., 159 F.2d 169, 173 (2d Cir. 1947).
businesses. All businesses should be encouraged to act responsibly and in this regard, Web site sponsors should be treated no differently from offline businesses. As previously noted, however, what is reasonable should depend upon the context, and a determination of reasonableness should factor in the ways that online businesses are different from offline businesses. Finally, a “reasonableness” standard best reflects social norms and accommodates technological evolution, making it especially attractive in the dynamic online context.

The proposals in this Article will not banish all online harassment, nor will they stop the most determined trolls; however, the proposals do force Web site sponsors to recognize there are measures they can take to curb abusive and damaging conduct on their Web sites. Cyberharassment is just too easy on the Internet. Perhaps most important, these proposals recognize and treat online harassment as the social problem that it is, rather than assuming all conduct and postings on the Internet are constitutionally protected “speech.”

The underlying objective of these proposals is to more closely align the social norms that exist on the Internet with those in the physical world. While many free speech advocates decry any sort of regulation as “censorship,” that term is wrongly applied where the content regulators are private entities. In fact, proactive steps on the part of Web sites may have the effect of forestalling government intervention efforts that could be more sweeping and invasive than actions taken by Web site sponsors, and certainly more restrictive than the foregoing proposals. For example, the federal grand jury indictment and conviction of Lori Drew in the high-profile MySpace cyberharassment case has generated much consternation among legal experts.241 Because there was insufficient evidence to bring charges under state criminal statutes, federal prosecutors brought the indictment under the Computer Fraud and Abuse Act.242 The indictment stated the Act was violated when Drew violated MySpace’s terms of service.243 Several legal scholars argued that the


243 Id.
prosecution set a dangerous precedent because few people read Web site terms of service. 244

VI. CONSTITUTIONAL SCRUTINY OF PROPOSED ANTI-CYBERHARASSMENT POLICIES

The First Amendment is often used as a defense to online harassment claims. 245 Part VI addresses the First Amendment doctrine in the online context generally, and then specifically as it pertains to the “reasonable measures” proposed in Part IV.

A. The Awkwardness of Applying First Amendment Doctrine to Online Harassment

The First Amendment prohibits the government from impinging on the freedom of speech. The objective of this prohibition was to prevent censorship and to encourage a free marketplace of ideas, which in turn, leads to knowledge and truth. 246 Yet there are limitations on the free speech right. These limitations include defining what constitutes “expression.” Obscenity, for example, is deemed to have no real expressive value and is not considered “speech” protected under the First Amendment. 247 Certain crimes and torts, such as verbal assault, defamation, and perjury, are directed purely at certain types of speech. 248

Whether words in a given context are protected as “speech” under the First Amendment may be analyzed in terms of the public/private distinction. Although

---

244 See id.; cf. Nick Akerman, The Law Fits the Crime, NAT’l L.J., May 26, 2008, at 1; Kim, supra note 162. In Korea, the suicide of a popular actress, reportedly distressed over Internet rumors, has prompted the governing party to promote a law to punish online insults that would be tougher than existing laws. See Choe Sang-hun, Korean Star’s Suicide Reignites Debate on Web Regulation, N.Y. TIMES, Oct. 13, 2008, at B7.


247 See Davenport v. Wash. Educ. Ass’n, 551 U.S. 177, 188–89 (2007) (noting that “speech that is obscene or defamatory can be constitutionally proscribed because the social interest in order and morality outweighs the negligible contribution of those categories of speech to the marketplace of ideas”).

248 See E. Walter Van Valkenburg, The First Amendment in Cyberspace, 75 Or. L. Rev. 319, 319 (1996) (“[P]rotections afforded by the First Amendment . . . are far from absolute. Certain categories of communication are subject to extensive regulation and, in some cases, outright prohibition. Defamation, for example, can give rise to civil liability, as can communication that violates rights to privacy or publicity. Pornography and other forms of obscenity are, under well established law, outside the protections of the First Amendment.”).
public speech is said to be accorded more protection than private speech, harms that are framed as “public” harms are also weighted more heavily than those that are framed as “private” harms. For example, where speech is labeled as obscene and not protected expression, the harm is to community norms. Defamation injures one’s reputation, which is the way others—the social community—think about the plaintiff, not the personal injury that it has caused the plaintiff.

In addition to criminal laws that limit speech, there are competing rights that limit free speech. Copyright, trade secret, and trademark law limit what one can say and/or how one can say it. With each of these, the right holder’s interest is pecuniary and therefore “public” because it affects the marketplace, rather than “private,” where the injury would be limited to the affected individual.

The Internet poses unique challenges and requires us to rethink the way we define rights and harms when it comes to speech. Where speech is “obscene,” courts have asked whether there was any expressive value and whether and to what extent community norms of decency were offended. Yet how do we evaluate speech against community norms where both the community and the norms are uncertain? What is the community for purposes of Internet speech? If poster

---

249 See Connick v. Myers, 461 U.S. 138, 146 (1983) (“When employee expression cannot be fairly considered as relating to any matter of political, social, or other concern to the community, government officials should enjoy wide latitude in managing their offices, without intrusive oversight by the judiciary in the name of the First Amendment.”).

250 Another way to consider the public/private distinction is as a power dynamic, given that those with a “public” interest have more socioeconomic power. For example, the movie industry is more powerful than an individual, and individuals belonging to minority groups have even less power. As Richard Delgado and Jean Stefancic have pointed out, “[p]owerful actors . . . have always been successful at coining free speech ‘exceptions’ to suit their interest.” Richard Delgado & Jean Stefancic, Ten Arguments Against Hate-Speech Regulation: How Valid?, 23 N. KY. L. REV. 475, 484 (1996). For further discussion on this issue, see Davenport, 551 U.S. at 188–89 (noting “speech that is obscene or defamatory can be constitutionally proscribed because the social interest in order and morality outweighs the negligible contribution of those categories of speech to the marketplace of ideas”).


252 Cf. Van Valkenburg, supra note 248, at 324 (observing that “existing First Amendment doctrine should translate relatively well to Cyberspace”).

253 See Davenport, 551 U.S. at 118 (citation omitted).

254 See Branscomb, supra note 20, at 1652 (discussing the problem of deciding “which local community’s standards should apply—that of the uploading provider, that of the downloading user, or the community standards maintained by and within the virtual community on the electronic network” in an online pornography case).
“X” describes sexual acts he would like to do to poster “Y,” such communication may be considered obscene in certain communities but not in others. The unrestrained nature of online discourse may lower the bar of socially acceptable speech on a particular Web site, which then spills over to other Web sites. Often what is considered obscene in an offline context is considered the norm on a particular Web site or chat room. Significantly, the norms on the Internet are now being shaped, and the extent to which we permit certain types of behavior affects what those norms are.

Uncivil and even unlawful conduct is perceived as a right. For example, stolen digital images of a movie star were posted online. The actor had taken his computer to be serviced, and the employees of the computer services firm had helped themselves to the contents of his digital photo album. When the police conducted an investigation, many online comment posters decried a crackdown on the right to free speech. But when did publication of stolen personal photographs constitute a free speech right? Many of the existing limitations on the right to speech do not apply in the context of the Internet. Courts have permitted reasonable time, place, and manner restrictions even with protected speech. The Internet blurs the public/private distinction and makes traditional time, place, and manner restrictions inapposite. For example, X, our hypothetical online harasser, is at home typing in the middle of the night, not on a street corner during the working day.

The purpose of time, place, and manner restrictions is to provide for public safety and to maintain order. Yet a secondary effect of the restrictions is to filter out those speakers who have only an impulsive or trivial interest in making a statement. In the offline world, there are physical barriers and pragmatic limitations to speech that are absent in the virtual world. For example, a speaker

255 Cf. John Fee, Obscenity and the World Wide Web, 2007 BYU L. REV. 1691, 1691 (arguing it is “constitutionally permissible to apply the traditional test for obscenity, the Miller standard, to the Internet, including its reference to ‘contemporary community standards,’ without any requirement for a more particular definition”). Fee argues against using a different standard for the Internet because doing so would “tip the scales” in obscenity cases toward defendants. Id. at 1692.

256 But as Danielle Citron notes, “anonymous message-board postings are not immune from defamation liability simply because they are too outrageous to be believed.” Citron, supra note 12, at 108.

257 See Branscomb, supra note 20, at 1641.


259 Id.

260 Id.

261 This example is even more striking considering the actor was Chinese and the comment posters appeared to be from China, which does not have free speech rights as in the United States. See id.

262 Harry Surden, Structural Rights in Privacy, 60 SMU L. REV. 1605, 1606 (2007) (noting that society relies upon latent structural constraints to inhibit unwanted conduct in a
may want to make a statement about the war in Iraq, but she may not feel strongly enough about making that statement to drive down to City Hall and join a demonstration on a rainy day. In addition, an individual who wishes to express an opinion in the physical world must find an outlet that finds that opinion newsworthy, or at least valuable to its readers. That opinion is thus subject to some sort of editorial screening, and the circulation of that publication often correlates with the selectivity of that screening process. Of course, an individual may also self-publish an opinion via leaflets or by shouting that opinion through a bullhorn. In those cases, the circulation of that opinion will be limited in both territory and duration. Leaflets are not preserved (if they are even read), and the existence of an opinion shared using a bullhorn does not survive its transmission. In addition, distribution of leaflets and the use of bullhorn announcements are subject to the time, place, and manner restrictions mentioned above.263

Some may argue that the very benefit of the Internet is its democratic accessibility.264 While this may be so, it does not diminish the argument that Internet communication is not treated the same as communication in the physical world. The physical barriers and legal restrictions placed on speech in the physical world serve as a means by which to screen out impulsive speakers and false or misleading information; Internet postings are not subject to the same built-in delay or editorial process. Postings can be made in the heat of emotion, without deliberation or a second opinion, and without a cooling period. The susceptibility of Internet communications to impulsive behavior is made even more significant given that the age of users appears to correlate strongly with frequency and complexity of online activity.265 Data collected from a survey conducted in August 2008 by the Internet and American Life Project of the Pew Research Center notes that 38 percent of people age 65 and older use the Internet compared with 91 percent of those ages 18 to 29, 86 percent of those ages 30 to 49, and 74 percent of

way that is functionally comparable to the law, and discussing how these latent structural constraints are vulnerable to dissipation due to emerging technologies).

263 Lucero v. Trosch, 121 F.3d 591, 599 (11th Cir. 1997) (holding abortion protestors were properly prohibited from demonstrating, including using bullhorns, within 200 feet of the residences of abortion clinic staff).

264 See SOLOVE, supra note 74, at 20, (arguing that “[b]logs are more egalitarian than the mainstream media” because “[y]ou don’t need connections to editorial page editors to get heard. If you have something interesting to say, then you can say it”). Solove discusses the thrill of the blog:

The blog I posted on was visited thousands of times a day. A lot of people were reading. What made this so exciting was that I’d never had any success getting an op-ed published. I had tried many a time, but the editors just wouldn’t give me a plot of valuable space on their pages. Suddenly I no longer needed them.

Id. at 5.

265 See infra notes 266–268 and accompanying text.
those ages 50 to 64. Even within a narrow age range, such as college students, the younger students were found to be more likely to post their creations online. Not surprisingly, some online posters may experience “poster’s regret,” a wasted emotion when the damage has already been done.

The proposals in Part IV contemplate how Internet communication differs from communication in the physical world. The public/private distinction has affected free speech analysis in a way that is inapplicable to Internet speech. Much of free speech analysis considers the location of where speech occurs and whether the defendant has a “commercial” interest in the speech. Commercial is typically understood to mean that the defendant has a pecuniary stake in the action rooted in the speech defense, such as a copyright or a valuable trade secret. With respect to online harassment, the competing interests are between the speaker and the object of the speech, and to a lesser extent, the site visitor’s “right to know.”

Much of the harm caused by distribution might be avoided if we limited anonymity. Anonymity has been tied to free speech, but limiting anonymity in the context of online harassment has positive “public” effects. Restricting anonymity increases reliability of information and encourages accountability, respects the public’s interest in the source of the information, heightens the expressive value of the speech, and reinforces social/community norms.

The private harm created by online harassment also has harmful public effects. Cyberharassment affects the employability of victims of harassment, undermines community norms, and ignores the malleability of such norms, thus leading to the “lowest common denominator” effect (i.e., a speaker’s conduct lowers the standard of civility on a Web site). It affects the way we interact with others, introducing a type of distrust and paranoia into personal relationships that may ultimately make such intimacy difficult, if not impossible. It also affects our expectations and our sense of what is normal and acceptable, creating incremental changes to our culture whether we acknowledge it or not. Because so many of the existing available tort remedies depend upon normative standards, such as

269 S. Elizabeth Malloy, Anonymous Blogging and Defamation: Balancing Interests of the Internet, 84 WASH. U. L.R. 1187, 1192 (2006) (noting that the Internet is different from other mediums in its ease of access, permanence, and pervasiveness).
270 For a more comprehensive discussion of the costs and benefits of anonymous speech, see Lidsky & Cotter, supra note 116, passim.
reasonableness, they may fail to protect the very values that those standards were intended to reflect. While some types of online harassment may now seem shocking or unexpected, repeated reports of such conduct may diminish that effect. For example, pictures of nude celebrities or celebrities having sex were shocking when they first appeared on the Internet; now, such images are commonplace and easy to find. A failure to explicitly address online harassment may dull our sensitivity to it. Resignation to cultural changes, however, should not be confused with acceptance or receptiveness. In the end, we must acknowledge that some types of communication or speech are just not expressive (or not expressive of anything that our society and judiciary have deemed worthy of protection).

The standard First Amendment response to “bad speech” has been “more speech.” This response makes certain assumptions about “bad speech” that are inaccurate or unproven. First, it assumes that “bad speech” is in fact constitutionally protected speech. It also assumes that “more speech” will be accorded the same platform as “bad speech,” and that any response will be distributed widely enough to blunt any ill-effects from the “bad speech.” In fact,

271 See Wolf, supra note 118 (“[T]he best antidote to hate speech, ADL maintains, is more speech.”). Kurt Opsahl, a staff attorney at the Electronic Frontier Foundation, notes that for those who feel slandered, “[T]he cure to bad speech is more speech.” Nakashima, supra note 2.

272 Professor Andrew Chin makes the convincing argument that because of concerns about balkanization and concentration of power exaggerated by the use of external Web links, the Internet is “one of many fora where social structure constrains public discourse.” Andrew Chin, Making the World Wide Web Safe for Democracy: A Medium-Specific First Amendment Analysis, 19 HASTINGS COMM. & ENT. L.J. 309, 313 (1997). Chin uses the following illustration:

[S]uppose that there are two perspectives, A and B, with respect to a particular political issue. Perspective A is held by 40% of the public and Perspective B is held by 10%, with the remaining 50% undecided. Each perspective is represented by a number of sites on the Web, proportional to its level of support in the population. Suppose that proponents of B believe that their perspective will be persuasive to anyone who engages in a deliberative comparison between A and B. Web sites for B therefore include many links to Web sites for A. On the other hand, proponents of A may believe that the best way to protect their lead in the polls is to avoid any reference to B. Because there are many more A sites than B sites on the Web, publishers of A sites can be confident that their perspective will be seen by the undecided reader.

As a result of these strategies, Web sites for A actually garner more than four times as many hits as Web sites for B among exploring readers. . . . This situation is analogous to the plight of marginalized groups in conventional public discourse: the minority group, in order to survive, must understand the dominant perspective sufficiently to deconstruct and criticize it, whereas the mainstream group may benefit unjustly from its ignorance of minority perspectives.
a site constructed by someone unfamiliar with how search engines retrieve results may get little or no attention at all.\(^{273}\) College students have been found to be heavily influenced by the order in which Google search results were presented.\(^{274}\) These students exhibited substantial trust in Google’s ability to rank results by their relevance to the query even where the abstracts were less relevant to their query.\(^{275}\) Thus, to effectively respond to bad speech, one must have the resources and know-how to effectively distribute that response as widely as the initial posting.

Finally, the “more speech” rejoinder fundamentally misunderstands the nature of “bad speech.” Although speech that expresses unpopular or reprehensible views may indeed be addressed or diluted by “more speech,” speech that threatens or personally attacks an individual is quite different. Responding to personal attacks by an anonymous poster through “more speech” adds fuel to the fire and may result in more vicious and repeated attacks. For example, when the parents of a boy with Keppen-Lubinsky syndrome (a rare congenital disorder), protested that images of their son were being distorted and circulated on the Internet, one Web site provided the following response:

Instead of locking this grotesque sin against the natural order in a dark basement like they should have, his family maintains a website chock full of lulzy photographs of the abomination as if they were actually proud of pushing the bawling skinbag out of the mother's obviously cursed twat.

His parents, however, were operating under the false assumption that the internet is a happy place where everyone’s feelings are respected and validated. It comes as no surprise then that the proud parents were shocked, SHOCKED, when they discovered that photos of their son had exploded across the internet into a wide variety of lulzy photoshoops and macros.\(^{276}\)


\(^{274}\) See Bing Pan et al., In Google We Trust: Users’ Decisions on Rank, Position, and Relevance, 12 J. COMPUTER-MEDIATED COMM. 801, 816 (2007).

\(^{275}\) Id.

\(^{276}\) Encyclopedia Dramatica, supra note 33. I include these statements to illustrate the aggression that a response to harassing posts might elicit. Although these statements might arguably constitute parody and/or protected expression, they might also constitute
The Web site continues its abuse by calling the boys’ parents “un-intelligent un-washed Guidos (as if there are any other kind), [who] have difficulty grasping the basics of the English language.” The insults are accompanied with doctored photographs of the boy and hurtful captions mocking his condition.

In addition to aggravating attacks, a requirement that the victim of an attack respond to the attack with “more speech” further degrades the victim by forcing him or her to engage the attacker. A response by the victim may also lend the initial attack more legitimacy and draw more unwanted attention to the post.

One could argue that recipients of online harassment should simply ignore the harassment and “grow a thicker skin.” The problem with this suggestion is not simply that it is unrealistic, but also that it encourages us as a society to cultivate insensitivity and apathy as a norm, and to shun mutual respect and civility as core values. Inaction to the problem of online harassment does not simply maintain the status quo. On the contrary, it protects and reinforces a standard of conduct that would be intolerable in the physical world. Even differentiating the “virtual world,” or “cyberspace,” from the “offline world,” or the “physical world,” is problematic because it assumes the existence of two different universes or realities. As this Article reflects, I am highly critical of this assumption. I nevertheless use the terms to distinguish Internet activity from non-Internet activity, a distinction that is important given the unique dimensions of harassment where the Internet is used as a medium of communication. In other words, although the Internet is not a different universe, communication distributed via the Internet has different effects from communication distributed through other mediums. This Article is concerned

intentional infliction of emotional distress, particularly when viewed in conjunction with the digitally altered images of the boy.

277 Id.
278 Id.
279 For example, law professor and blogger Ann Althouse criticized the filing of the AutoAdmit lawsuit by stating: “So this is the 21st century? Where courts award punitive damages for offensive words and pictures? Isn’t ‘the scummiest kind of sexually offensive tripe’ exactly what we always used to say people had to put up with in a free country? Man, that was so 20th century!” Ann Althouse, Yale Law Students Sue Over the ‘The Scummiest Kind of Sexually Offensive Tripe’ at AutoAdmit, http://althouse.blogspot.com/2007/06/yale-law-students-sue-over-scummiest.html (June 12, 2007, 15:52). In reference to commentary about Althouse’s post, another law professor and blogger, Glenn Reynolds, added:

Well, I’m pretty thick-skinned about Internet trash-talk—when I teach libel I give my students a few choice search terms and let them see what people have said about me. They’re usually appalled, but I’ve never sued anyone, and the list of things about which I might actually sue is awfully short. Besides, once you get past the puppy-blending stuff, who’s going to believe much of anything they read?

with the effects of Internet communication without meaning to suggest that those
effects are confined to the Internet. The remainder of Part VI explains how and
why the strategies proposed in Part IV, with a few modifications, survive
constitutional scrutiny.

B. Contractual and Architectural Constraints are Content-Neutral

The contractual and architectural constraints proposed in Part IV are intended
to curb impulsive and regrettable behavior by forcing the poster to think before
pressing send. The effect may be to evaluate and filter speech, but it is the poster,
and not a government actor, who is censoring communication. These contractual
and architectural constraints serve some of the same functions as time, place, and
manner restrictions and structural barriers. They provide an opportunity for
contemplation without imposing a ban on speech, leaving the decision whether to
post the communication in the hands of the poster. A state-mandated policy might
arguably adopt even more onerous content-neutral restrictions without violating
free speech principles.\(^{280}\) I do not advocate broader restriction, but raise the
possibility simply as a point of comparison—as an example of how far the
government could go in the absence of self-imposed regulation on the part of both
Web site sponsors and users.\(^{281}\) My proposed contractual and architectural
constraints, by contrast, merely encourage the exercise of best judgment and self-
control.

C. Easy Unmasking Proposals Can Be Limited to Survive Constitutional Scrutiny

Although the Supreme Court has recognized that the First Amendment
protects a general right to anonymity, it has done so in limited contexts and only
after weighing governmental interests against individual interests.\(^{282}\) The Supreme
Court has never ruled on an absolute right to anonymity,\(^{283}\) nor has it ruled on the

\(^{280}\) See Perry Educ. Ass’n v. Perry Local Educators’ Ass’n, 460 U.S. 37, 44–55
(1983).

\(^{281}\) Content-neutral government regulations are subject to a less rigorous analysis than
content-based regulations. See Turner Broad. Sys., Inc. v. F.C.C., 520 U.S. 180, 213–15
(1997).

\(^{282}\) Talley v. California, 362 U.S. 60, 64–65 (1960) (striking down as unconstitutional
a city ordinance prohibiting the distribution of handbills without identifying the sponsors);
banning distribution of anonymous political campaign literature was not justified by state
interest in preventing fraud and libel and informing voters).

\(^{283}\) In Talley v. California, the city ordinance prohibited all anonymous leafleting, 362
U.S. at 60. In McIntyre v. Ohio Elections Commission, the text of the state statute contained
no language limiting its application to fraudulent, false, or libelous statements, and
therefore the Court held it was overbroad. 514 U.S. at 357.
issue of anonymous speech regarding nonpublic matters. In fact, language in the two Supreme Court cases recognizing a constitutionally protected right to anonymity suggest it would not extend to false or fraudulent speech. Furthermore, although Internet anonymity has been recognized by lower courts, the Supreme Court has never addressed the issue of whether anonymity on the Internet is constitutionally protected, much less decided upon the parameters of any such presumed right.

“Easy unmasking” policies do not mandate identification as a precondition to posting, nor do they prohibit anonymous speech. Furthermore, policies where unmasking is limited to cases of online threats, gossip, and confessionals typically would not involve protected speech. Easy unmasking in these three situations is unlikely to affect protected speech because they involve threats or defamatory statements. Even if the gossip or confessionals involve events or facts that are true, they may be invasive of privacy because they publicly disseminate private information about private individuals. Finally, governmental regulations or court decisions may avoid First Amendment challenges by making easy unmasking policies voluntary, but also encouraging their adoption. For example, the adoption of an unmasking policy may be evidence of reasonable care, but a failure to adopt such a policy in and of itself would not constitute negligence.

D. Prohibition on Certain Digital Images is Narrowly Tailored to a Legitimate Government Interest

There are anticipated objections to the outright prohibition of digital images of non-public figure children and of nude individuals. The first objection is that they limit the expression of the poster. In the case of images of nude individuals, many of these images would not be considered fully protected speech because they would be obscene or pornographic and could be regulated under the government’s police power. The expressive nature is debatable as the poster is not the subject

---

284 The decision in Talley addressed an economic boycott, 362 U.S. at 60, and the decision in McIntyre dealt with political speech, 514 U.S. at 334. See also Strickland, supra note 80, at 1558 (asserting that a “thorough examination of both McIntyre’s facts and the Court’s analysis indicates that the assumption that the case extends to all anonymous Internet speech is conclusory and may be incorrect,” and that “wholesale application of McIntyre in the context of a cybersmear claim is misapplied”).

285 See McIntyre, 514 U.S. at 1523 (noting that the right to anonymity may be abused when it shields fraudulent conduct and that the state may punish fraud directly); Talley, 362 U.S. at 538 (expressly noting that the Court was not deciding on ordinances limited to prevent fraud).


287 Of course, Web site sponsors could apply unmasking strategies to a wider range of conduct.

288 See Heideman v. S. Salt Lake City, 348 F.3d 1182, 1195–96, 1183 (10th Cir. 2003) (holding ordinance banning nudity within sexually oriented business was not subject to strict scrutiny because prohibition was on a form of conduct and applied to all such
in the photograph; the act of expression then is the uploading of another’s image. 289 Even where speech is affected, the government has a compelling state interest in protecting citizens’ most basic privacy. There is indisputably a privacy interest at stake where unauthorized, widespread distribution of nude images is concerned, especially where the subject is not a public figure.

The government has a compelling state interest in protecting minors from abuse and exploitation. Some Missouri lawmakers are considering an outright ban on sex offenders’ permission to take photographs of children primarily because of the ease with which such photographs can be shared with other sex offenders on social networking sites 290 Georgia lawmakers have also considered a similar measure. 291 Yet recent studies indicate that online harassment of teens by their peers is a bigger threat than sexual predation online. 292 Cyberharassment by peers includes not just images taken without the consent of the subject, but images taken by the subject him or herself, which are then misused by the recipient.

According to a recent study, one in five teenagers and one in three young adults use cell phones and online technology to send or post nude or seminude images of themselves. 293 Seventy percent of those who admitted to sending sexual images of themselves sent them to a boyfriend or a girlfriend. 294 A teenager sending a sexy image to a significant other is not thinking the relationship might meet an unpleasant end. In the aftermath of a breakup, teenagers and young adults may not be thinking rationally. Raging hormones, a bruised ego, and a brain not yet fully formed in the crucial ability to empathize and control emotions make possession of an ex’s naked digital image a powerful and dangerous weapon. With a few clicks of a mouse, a jilted and heartbroken teen can get the type of revenge that was unimaginable a few years ago. Although young love does not last forever, a digital image of one’s naked teen self might, haunting someone like a digital

---

291 Id.
293 Survey, National Campaign to Prevent Teen and Unwanted Pregnancy & Cosmogirl.com, Sex and Tech: Results from a Survey of Teens and Young Adults 1, http://www.thenationalcampaign.org/sextech/PDF/SexTech_Summary.pdf.
294 Id. at 4.
scarlet letter and limiting one’s future career choices and life opportunities. News stories and court cases indicate young men are more likely to exact this type of revenge upon their ex-girlfriends. In addition, more than 80 percent of survey respondents said they thought the reason girls send sexy images is to get a guy’s attention (compared with 60 percent who thought guys did so to get a girl’s attention) and about 40 percent said they thought girls sent sexy images under pressure from a guy (compared with 18 percent who thought guys did so under pressure from a girl).

The restriction on speech imposed by this Article’s proposed ban on these two categories of digital images is narrowly tailored to the specific harms created by widespread and easy Internet distribution. It does not prohibit existing avenues for distribution, such as print publication. It is also limited to publicly accessible Web sites and would not preclude someone from posting photographs to an invitation-only, password-required, photo-sharing Web site. The restriction on speech—if these images would qualify as protected speech—applies only where the necessary consent is missing. As previously mentioned, the poster can always digitally alter or crop out an objecting subject’s image in an otherwise permissible image, such as a group picture. Clear parameters could be established on response times or the number of requests required before determining that the Web site sponsor has acted unreasonably. Concerns that a Web site might be inundated with illegitimate takedown requests might be tempered if requestors were fined for failing to provide verification of their identities and standing to request the takedown. Finally, a court might conclude that reasonable conduct entails only taking down images of nude nonpublic figures, rather than any unauthorized nude image.

See Laure Manaudou, Chi. Trib., Aug. 11, 2008, http://www.chicagotribune.com/sports/olympics/chi-laure-manaudou-080811-hf,0,6552632.story (describing how nude photos of Laure Manaudou, an Olympic swimmer from France, appeared on Web sites after a tempestuous break-up with her boyfriend). The photos appeared to have been taken by a lover in “an intimate moment,” but her boyfriend disclaimed any responsibility. Id. Manaudou reported that, “[w]henever I typed Laure Manaudou on the Internet, it was horrible. I felt humiliated.” Id.; see also Barnes v. Yahoo!, No. Civ. 05-926-AA, 2005 WL 3005602 (D. Or., Nov. 8, 2005), rev’d, No. 05-36189, 2009 WL 1232367 (9th Cir., May 7, 2009) (addressing plaintiff’s complaint over ex-boyfriend posting nude pictures of plaintiff to Yahoo’s online profiles).

Constitutional law scholar Erwin Chemerinsky has noted U.S. Supreme Court cases “try to strike a balance: They give more weight to speech that is relevant to the political process and of public interest; they give more weight to reputation when a person has not voluntarily entered the public domain and when the matter is not of public concern.” CHEMERINSKY, supra note 150, at 1055. Another potential objection to a ban is that the image is the property of the individual who captured it, since that individual owns the copyright. This argument should also be dismissed because copyright ownership should not be allowed to trump the privacy and security interests at stake in these two limited circumstances. Copyright is granted to encourage creators to create for the benefit of society; yet it is not socially beneficial to have unauthorized pictures of minors or unauthorized pictures of naked children or adults circulating on the Internet. It is unlikely
VII. CONCLUSION

Although there are difficulties in monitoring content, Web sites should not then have no obligation to deter or minimize online harassment on their Web sites. Whether one concludes that Web sites are sui generis or “just like property” is beside the point. Web site sponsors have a proprietary interest in their sites and should have some accountability for what happens as a result of the forum they provide (even if not the same liability that mainstream media publishers have). This Article proposes several simple measures Web sites can take to deter online harassment without having to pre-screen or make subjective decisions about user-supplied content. By suggesting actions that Web site sponsors can readily adopt, this Article aims to shift the normative expectations of Web site sponsor behavior and to reframe the problem of online harassment as one of a failure of business norms rather than of free speech gone wild.
GOOD FAITH, FIDUCIARY DUTIES, AND THE BUSINESS JUDGMENT RULE IN DELAWARE

Clark W. Furlow*

TABLE OF CONTENTS
I. INTRODUCTION ............................................................................................... 1062
II. THE IMPORTANCE OF GOOD FAITH ............................................................... 1063
III. GOOD FAITH IS NOT AN INDEPENDENT FIDUCIARY DUTY ....................... 1066
IV. DIRECTORS’ DUTIES AND GOOD FAITH ....................................................... 1068
   A. The Statutory Duties of Directors ............................................................. 1068
   B. The Fiduciary Duties of Directors .......................................................... 1069
   C. Good Faith and the Duty of Loyalty ...................................................... 1069
V. THE JURISPRUDENCE OF BAD FAITH ............................................................. 1071
   A. Decisions Motivated by an Improper Purpose ........................................ 1071
   B. Decisions Motivated by Subjective Bad Faith ........................................ 1072
   C. Intentional Violation of Law .................................................................... 1072
   D. Intentional Dereliction of Duty ............................................................ 1073
      1. Good Faith and the Duty to Monitor .................................................... 1073
      2. Good Faith and Careless Decision-Making ......................................... 1075
      3. Dereliction of Duty versus Gross Negligence .................................... 1077
   E. Lyondell Chemical Co. v. Ryan ................................................................. 1080
   F. The Proper Role of Good Faith in Fiduciary Analysis ............................. 1082
VI. GOOD FAITH AND THE BUSINESS JUDGMENT RULE .................................... 1083
   A. The Presumption of Care and Loyalty ...................................................... 1083
   B. Rebutting the Presumption of Care ........................................................... 1084
   C. Rebutting the Presumption of Loyalty .................................................... 1085
      1. Financial Interest .................................................................................. 1085
      2. Lack of Good Faith ............................................................................. 1086
   D. Establishing an Improper Motive ............................................................ 1088
      1. Incidental Benefit to the Directors’ Affiliates ....................................... 1088
      2. Irrational Business Judgments ............................................................. 1091
VII. PROPOSED MODE OF JUDICIAL REVIEW OF GOOD FAITH ...................... 1093
VIII. CONCLUSION ............................................................................................. 1095

* © 2009 Clark W. Furlow, is an Associate Professor of Law at Stetson University College of Law and serves at the College’s Associate Dean of its Tampa Law Center. Professor Furlow wishes to thank the Honorable Andrew G.T. Moore II, Justice of the Delaware Supreme Court, retired, for his thoughtful comments and criticisms. He also wishes to express his gratitude to Sally A. Milano (she knows why), to Brian R. Dettman for his assistance, and to Stetson University College of Law for providing financial support for this Article.
I. INTRODUCTION

The Delaware courts frequently refer to a corporate director’s duty of good faith. However, they struggle to define it. One Delaware judge described this struggle as a “fog of hazy jurisprudence.” Delaware’s inability to offer a clear, consistent conception of good faith is significant because, under Delaware statutory law, a director is not entitled to protection from personal liability for a decision that was not taken in good faith.

This Article argues that any effort to define good faith as an independent fiduciary duty is doomed to failure because the term describes a state of mind—the honesty with which one holds a belief—whereas fiduciary duties define the way directors are to conduct themselves in office.

---

1 See, e.g., Aronson v. Lewis, 473 A.2d 805, 812 (Del. 1984) (holding that directors of a Delaware corporation are presumed to act “on an informed basis, in good faith and in the honest belief that the action taken was in the best interests of the company”), overruled by Brehm v. Eisner, 746 A.2d 244 (Del. 2000); Cede & Co. v. Technicolor, Inc. (Cede II), 634 A.2d 345, 361 (Del. 1993) (referring to the duties of loyalty, care, and good faith as a “triad” of fiduciary duties).

2 See, e.g., Cede II, 634 A.2d at 361–66 (containing two lengthy sections on the duty of loyalty and the duty of care, but not defining good faith).

3 In re Walt Disney Co. Derivative Litig. (Disney IV), 907 A.2d 693, 754 (Del. Ch. 2005). The long-running Disney stockholder derivative litigation involved an effort by stockholders to hold Disney’s directors personally liable for hiring and then firing Michael Ovitz as the company’s president under circumstances that allowed Mr. Ovitz to receive a severance package valued at $140 million after fourteen months of lackluster performance. Id. at 699–733. The litigation produced five significant judicial decisions: In re Walt Disney Co. Derivative Litig. (Disney I), 731 A.2d 342, 380 (Del. Ch. 1998) (granting the defendants’ Rule 23.1 motion to dismiss for plaintiffs’ failure to allege with particularity facts that would excuse their failure to make pre-suit demand); Brehm v. Eisner (Disney II), 746 A.2d 244, 267 (Del. 2000) (affirming the dismissal of plaintiffs’ complaint and remanding with leave to file an amended complaint); In re Walt Disney Co. Derivative Litig. (Disney III), 825 A.2d 275, 291 (Del. Ch. 2003) (denying defendants’ Rule 23.1 motion to dismiss and holding the plaintiffs had adequately alleged bad-faith breaches of fiduciary duty that would excuse demand); In re Walt Disney Co. Derivative Litig. (Disney IV), 907 A.2d 693, 778–79 (Del. Ch. 2005) (holding, in post-trial decision in favor of all defendants, that the evidence failed to demonstrate bad faith); and In re Walt Disney Co. Derivative Litig. (Disney V), 906 A.2d 27, 75 (Del. 2006) (affirming Disney IV).

4 See DEL. CODE ANN. tit. 8, § 102(b)(7) (2001) (allowing a Delaware corporation to include a provision in its certificate of incorporation shielding its directors from personal liability from certain breaches of fiduciary duty, provided the breach did not involve an act or omission “not in good faith”).

5 Good faith is defined as “[a] state of mind consisting in (1) honesty of belief or purpose, (2) faithfulness to one’s duty or obligation, (3) observance of reasonable commercial standards of fair dealing in a given trade or business, or (4) absence of intent to defraud or to seek unconscionable advantage.” BLACK’S LAW DICTIONARY 713 (8th ed. 2004); see also Elizabeth A. Nowicki, A Director’s Good Faith, 55 BUFF. L. REV. 457, 482–86 (2007) (reviewing uses of the phrase “good faith” in fiduciary and other contexts
The duty of loyalty defines what the directors are to seek to accomplish—i.e., the best interests of the corporation.\textsuperscript{7} The duty of care defines how they are to pursue that goal—i.e., by using “that amount of care which ordinarily careful and prudent men would use in similar circumstances.”\textsuperscript{8} Good faith, on the other hand, describes the state of mind of a director who is acting in accordance with her duty of loyalty. The duty of loyalty requires that a director’s corporate decision be based on a good-faith belief that it will serve the best interests of the corporation.\textsuperscript{9} It is the absence of that belief that justifies imposing personal liability upon a director.

Part II explains the importance of understanding the role of good faith in the analysis of directors’ fiduciary duties. Part III explains that good faith does not function as an independent fiduciary duty. Part IV examines the statutory and equitable sources of directors’ corporate duties and explains that good faith is the state of mind with which a director acts in compliance with her duty of loyalty. Part V traces the effort of Delaware courts to define good faith as the absence of bad faith and concludes that bad faith does not define good faith; rather, bad faith is defined as the absence of good faith. Part VI examines the role of good faith in the context of the business judgment rule. Finally, Part VII offers a definition of good faith consistent with the purpose of section 102(b)(7) of the Delaware Code, traditional understandings of directors’ fiduciary duties, and the business judgment rule.

II. THE IMPORTANCE OF GOOD FAITH

Delaware has long expected corporate directors to perform their duties in good faith.\textsuperscript{10} The business judgment rule, by which Delaware courts review

---

\textsuperscript{6} \textit{Black’s Law Dictionary} 543 (8th ed. 2004) (defining “duty” as a “legal obligation that is owed or due to another and that needs to be satisfied”).

\textsuperscript{7} The duty of loyalty requires directors “to protect the interests of the corporation committed to his charge,” and at the same time it requires the director “to refrain from doing anything that would work injury to the corporation, or to deprive it of profit or advantage which his skill and ability might properly bring to it, or to enable it to make in the reasonable and lawful exercise of its powers.” Guth v. Loft, Inc., 5 A.2d 503, 510 (Del. 1939).

\textsuperscript{8} \textit{Disney IV}, 907 A.2d at 749 (quoting Graham v. Allis-Chambers Mfg. Co., 188 A.2d 125, 130 (Del. Super. Ct. 1963)). This requires that the directors “consider all material information reasonably available.” \textit{Id.} (quoting \textit{Disney II}, 746 A.2d at 259).

\textsuperscript{9} Guttman v. Huang, 823 A.2d 492, 506 n.34 (Del. Ch. 2003) (“A director cannot act loyally towards the corporation unless she acts in the good faith belief that her actions are in the corporation’s best interest.”).

\textsuperscript{10} See, e.g., Skeen v. Jo-Ann Stores, Inc., 750 A.2d 1170, 1172 (Del. 2000) (citing Malone v. Brincat, 722 A.2d 5, 10 (Del. 1998)) (stating that directors are “fiduciaries who owe good faith to the company and its stockholders”); Warshaw v. Calhoun, 221 A.2d 487, 493 (Del. 1966) (holding that directors are presumed to act in good faith); \textit{Guttman}, 823
business decisions, creates “a presumption that in making a business decision the
directors of a corporation acted on an informed basis, in good faith and in the
honest belief that the action taken was in the best interests of the company.” 11 The
Delaware courts frequently recite this explanation of the business judgment rule, 12
but the phrase “good faith” is usually included as something of a rhetorical grace
note; it has never provided the basis for the court to decide the matter before it.

Good faith took on greater significance in 1986, when, in response to the
Delaware Supreme Court’s decision in Smith v. Van Gorkom, 13 the Delaware
legislature adopted section 102(b)(7). 14 The statute allowed a Delaware corporation
to include a provision in its certificate of incorporation shielding its directors from
personal liability for breach of the duty of care. 15 The statute accomplishes this by

---

11 Aronson v. Lewis, 473 A.2d 805, 812 (Del. 1984) (emphasis added); see also
Kaplan v. Centex Corp., 284 A.2d 119, 124 (“‘The acts of directors are presumptively acts
taken in good faith and inspired for the best interests of the corporation . . . .’” (quoting
Corp., 126 A. 46, 48 (Del. Ch. 1924) (“[T]he directors of the defendant corporation are
clothed with that presumption which the law accords to them of being actuated in their
conduct by a bona fide regard for the interests of the corporation whose affairs the
stockholders have committed to their charge.”).

12 See, e.g., Smith v. Van Gorkom, 488 A.2d 858, 872 (Del. 1985); see also Citron v.
Fairchild Camera and Instrument Corp., 569 A.2d 53, 69 (Del. 1989) (holding that the
board’s decision to approve the sale of the company was protected by the business
judgment rule).

13 488 A.2d 858. In Van Gorkom, the court held that a violation of the duty of care
would support the imposition of personal, monetary liability on independent, disinterested
directors who believed in good faith that they were serving the best interests of their
company and its stockholders. Id. at 864. As a consequence, liability insurance for
corporate directors became difficult to purchase and, when available, expensive. See
EDWARD P. WELSH, ANDREW J. TUREZYN, & ROBERT S. SAUNDERS, FOLK ON THE
DELAWARE GENERAL CORPORATION LAWS §102.14 (5th ed. 2008). This created a concern
that corporations might not be able to attract and retain qualified directors. Id. The
Delaware legislature adopted section 102(b)(7) to address these problems.

14 DEL. CODE ANN. tit. 8, § 102(b)(7) (2001). For a discussion of the process by
which the Delaware legislature adopted section 102(b)(7), see Christopher M. Bruner,
Good Faith, State of Mind, and the Outer Boundaries of Director Liability In Corporate
corporation law section of the Delaware State Bar Association, the body that drafted the
statute and recommended its adoption to the General Assembly of the State of Delaware).

15 Section 102(b)(7) provides that the certificate of incorporation of a Delaware
corporation may contain

[a] provision eliminating or limiting the personal liability of a director to the
corporation or its stockholders for monetary damages for breach of fiduciary
duty as a director, provided that such provision shall not eliminate or limit the
liability of a director: (i) For any breach of the director’s duty of loyalty to the
permitting exculpation for every “breach of fiduciary duty,” with four exceptions. The important exceptions were (1) a breach of the fiduciary duty of loyalty, and (2) a breach of fiduciary duty involving an “act[ ] or omission[ ] not in good faith.”16 Thus, directors of Delaware corporations would continue to be subject to personal liability for breach of the duty of loyalty and for “acts or omissions not in good faith,” but they could be protected from personal liability for a breach of the duty of care. This raised an important question: could a plaintiff transform an act or omission that violated the duty of care into a claim that would support the imposition of personal liability by showing that the careless act or omission was “not in good faith?”

The analysis of this question got off to a poor start when, in 1993, the Delaware Supreme Court seemed to suggest that good faith functioned as an independent fiduciary duty. In Cede & Co. v. Technicolor, Inc. (“Cede II”),17 the court explained that a plaintiff could rebut the business judgment rule’s presumption that directors had based their decision on adequate information (in compliance with their duty of care) and a belief that the decision would serve the company’s best interests (in compliance with their duty of loyalty)18 by “providing evidence that directors, in reaching their challenged decision, breached any one of the triads [sic] of their fiduciary dut[ies]—good faith, loyalty or due care.”19

This “triad” of fiduciary duties appeared to elevate the concept of good faith to the status of an independent fiduciary duty on par with the traditional duties of loyalty and care.20 Subsequent decisions repeated or paraphrased this language, bringing the weight of authority in support of this “triadic” definition of directors’ fiduciary duties.21 However, none of these cases involved allegations of a breach of

16 Id. at 360. The court noted the business judgment rule “creates a presumption that in making a business decision, the directors of a corporation acted on an informed basis [i.e., with due care], in good faith and in the honest belief that the action taken was in the best interest of the company.” Id. at 360 (quoting Aronson v. Lewis, 473 A.2d 805, 812 (Del. 1984)).


19 Emerald Partners v. Berlin, 787 A.2d 85, 91 (Del. 2001) (“[A] shareholder plaintiff has the burden of proving that the board of directors, in reaching its challenged decision, violated any one of its triad of fiduciary duties: due care, loyalty, or good faith.”); Cinerama Inc. v. Technicolor, 663 A.2d 1156, 1164 (Del. 1995) (“[A] shareholder plaintiff
the duty of good faith, and, therefore, none discussed the contours of this apparent duty of good faith. As a consequence of this “triadic” view of fiduciary duties, the phrase “good faith” appeared to have taken on dual significance. As a matter of judge-made law, it appeared to be an independent fiduciary duty, and as a matter of statutory law, it was a duty the violation of which would necessarily expose a director to personal liability.

III. GOOD FAITH IS NOT AN INDEPENDENT FIDUCIARY DUTY

The first serious consideration of the meaning of good faith came in the Court of Chancery’s decision in *In re Caremark International, Inc. Derivative Litigation* ("*Caremark*”). That decision was thought by some to endorse the view that good faith functioned as an independent fiduciary duty. The court held that a board’s sustained and systematic failure to monitor its corporation’s compliance with governing law would be evidence of an absence of good faith that would avoid the protection of an exculpation provision in the company’s certificate of incorporation. However, a closer reading of *Caremark* reveals the court did not apply the concept of good faith as if it were an independent fiduciary duty. Rather, the court held that a lack of good faith could be demonstrated by a board’s failure to undertake a preexisting duty—the statutory duty to monitor. The duty to monitor implicates the duty of care. It requires directors to exercise reasonable care in designing a monitoring system. But the court did not hold that a failure to exercise care would demonstrate a lack of good faith. Rather, the court held that a lack of good faith would be evinced by a failure to even

assumes the burden of providing evidence that the board of directors, in reaching its challenged decision, breached any one of its triad of fiduciary duties: good faith, loyalty, or due care.”). The term “triadic” was coined by academic commentators. See Eisenberg, *supra* note 20, at 5–6, 11–12 (arguing in favor of a triadic conception of fiduciary duty).

In one case, the Delaware Supreme Court made remarks that seemed to suggest that “bad faith” would be required to establish a breach of the duty of good faith. *Cinerama*, 663 A.2d at 1164. In *Cinerama*, the court noted that *Cede II* involved only duty of care and duty of loyalty claims because the plaintiff had abandoned its claim of “bad faith.” *Id.* (citing *Cede II*, 634 A.2d at 359).

*Id.* at 970. For a discussion of the duty to monitor, see *infra* notes 75–84 and accompanying text.

**DEL. CODE ANN. tit. 8, § 141 (2001)** (providing that the business and affairs of a Delaware corporation are under the management and direction of its board of directors).
attempt to perform the statutory duty to monitor. 28 A failure to perform a statutory duty to oversee the company’s business affairs cannot be motivated by a belief that such a failure would serve the company’s best interests. Thus, a systematic and sustained failure to perform this statutory duty would constitute a breach of the duty of loyalty. 29

The idea that good faith functioned as an independent duty was called into question in Chancellor Chandler’s decision in Disney IV. 30 There, he acknowledged that an intentional dereliction of duty fell short of the board’s obligation to act in good faith. 31 In the next sentence, however, the court made it clear that good faith was a component of the duty of loyalty and not a separate duty, stating that “[d]eliberate indifference and inaction in the face of a duty to act is, in my mind, conduct that is clearly disloyal to the corporation. It is the epitome of faithless conduct.” 32

The idea that good faith functions as an independent duty was put to rest by the Delaware Supreme Court’s recent decision in Stone v. Ritter. 33 There the court explained that oversight liability, which rests on a failure to act in “good faith,” was a violation “of the fundamental duty of loyalty.” 35 Thus, the court said, “good faith” is within the duty of loyalty. 36 Stone also rejected Cede II’s triadic formulation of fiduciary duties, explaining that “the obligation to act in good faith does not establish an independent fiduciary duty that stands on the same footing as the duties of care and loyalty. Only the latter two duties, where violated, may directly result in liability . . . .” 37

28 Caremark, 698 A.2d at 970.
29 In Guttman v. Huang, the court said that the Caremark opinion “articulates a standard for liability for failures of oversight that requires a showing that the directors breached their duty of loyalty by failing to attend to their duties in good faith.” 823 A.2d 492, 506 (Del. Ch. 2003) (footnote omitted). “Put otherwise, the decision premises liability on a showing that the directors were conscious of the fact that they were not doing their jobs.” Id.
30 907 A.2d 693 (Del. Ch. 2005).
31 Id. at 755.
32 Id. (emphasis added).
33 911 A.2d 362, 369–70 (Del. 2006). Stone was a derivative action in which plaintiffs sought to hold the corporation’s directors liable for criminal fines and penalties paid by the corporation for its violations of federal banking laws. Id. at 364–66. The plaintiffs alleged these violations of law were the result of the directors’ failure to implement an effective monitoring system. Id. at 364. The corporation’s certificate contained an exculpation provision shielding the directors from personal liability for breach of the duty of care. Id. at 367.
34 Id. at 369.
35 Id. at 370.
36 Id.
37 Id.
IV. DIRECTORS’ DUTIES AND GOOD FAITH

A duty is a “legal obligation that is owed or due to another and that needs to be satisfied.” 38 A duty defines conduct. An affirmative duty imposes an obligation on the director to do something, and a negative duty imposes an obligation on the director to refrain from doing something. 39

Directors of Delaware corporations have two layers of duties. The first is imposed by statute and the second is imposed by principles of equity.

A. The Statutory Duties of Directors

The Delaware General Corporation Law requires the board of directors to manage the business and affairs of its corporation. 40 These management duties are divided into two broad areas: decision making and supervision. 41 The board’s decision-making function imposes on the board the duty to decide how the company will run its business. The board makes the big decisions that chart the corporation’s future. 42 The board’s supervisory function imposes on the board the duty to monitor the performance of the corporation’s officers and employees and to make sure they comply with the law. 43

38 BLACK’S LAW DICTIONARY 543 (8th ed. 2004); see also Nowicki, supra note 5, at 510 (discussing the concept of duty).

39 For example, the duty of loyalty imposes an affirmative obligation that requires directors “to protect the interests of the corporation committed to his charge,” and at the same time it imposes a negative obligation that requires the director “to refrain from doing anything that would work injury to the corporation, or to deprive it of profit or advantage which his skill and ability might properly bring to it, or to enable it to make in the reasonable and lawful exercise of its powers.” Guth v. Loft, Inc., 5 A.2d 503, 510 (Del. Super. 1939).

40 DEL. CODE ANN. tit. 8, § 141 (2001). Delaware Code provides, in pertinent part:

The business and affairs of every corporation organized under this chapter shall be managed by or under the direction of a board of directors, except as may be otherwise provided in this chapter or in its certificate of incorporation. If any such provision is made in the certificate of incorporation, the powers and duties conferred or imposed upon the board of directors by this chapter shall be exercised or performed to such extent and by such person or persons as shall be provided in the certificate of incorporation.


42 “Legally, the board itself . . . [is] required to authorize the most significant corporate acts or transactions: mergers, changes in capital structure, fundamental changes in business, appointment and compensation of the CEO, etc.” Id. at 968.

43 See id.
B. The Fiduciary Duties of Directors

The second layer of duties consists of fiduciary duties imposed on the directors by traditional principles of equity. This layer requires the board to perform its statutory duties in a way that is consistent with the directors’ fiduciary duties of loyalty and care. The duty of loyalty defines what directors are to seek to accomplish while performing their statutory duties. The duty of care defines how directors are to pursue that goal. The duty of loyalty requires directors’ decision making to be motivated by an intention to serve the best interests of the corporation and its stockholders. The duty of care requires directors to make decisions in an informed and deliberate manner, and to use reasonable prudence in performing their monitoring function.

C. Good Faith and the Duty of Loyalty

Good faith, unlike the duties of loyalty and care, defines neither the goal toward which directors’ actions should be aimed nor the way directors are to pursue that goal. The duties of loyalty and care define conduct. Directors can be told to be loyal, and they can be told to be careful, but they cannot be told to be “good faith.” The reason is this: unlike the duties of loyalty and care, “good faith” does not define conduct. Rather, it defines the state of mind.

“Good faith” is “[a] state of mind consisting in (1) honesty in belief or purpose, (2) faithfulness to one’s duty or obligation, (3) observance of reasonable commercial standards of fair dealing in a given trade or business, or (4) absence of intent to defraud or to seek unconscionable advantage,” and is alternatively termed “bona fides.” Generally, it is an honest, faithful, sincere, and reasonable belief one is doing the right thing.

In the context of directors’ decision-making function, the duty of loyalty requires that a board decision be intended to serve the best interests of the corporation and the stockholders. But in deciding what course of action is most...

---

44 See Revlon, Inc. v. MacAndrews & Forbes Holdings, 506 A.2d 173, 179 (Del. 1986). The Delaware General Corporation Law does not define directors’ duties to the corporation or its shareholders. Rather, it leaves these duties to be defined under traditional principles of equity by the judges of the Delaware Supreme Court and the Delaware Court of Chancery.


46 Van Gorkom, 488 A.2d at 872–73.

47 Stone v. Ritter, 911 A.2d 362, 369 (Del. 2006); Caremark, 698 A.2d at 971.

48 BLACK’S LAW DICTIONARY 713 (8th ed. 2004); see also Nowicki, supra note 5, at 481 (quoting BLACK’S LAW DICTIONARY).

49 See Nowicki, supra note 5, at 482–83.

50 When the context of the transaction requires directors to focus on the welfare of the corporation, they must decide what course of action will serve the corporation’s best interests. When the context of the transaction requires them to focus on the welfare of the
likely to serve the best interests of the corporation, directors must make assessments about the future. Directors can never be certain their decision will work out as planned. Because directors are not expected to know with certainty what will happen in the future, the law allows directors to base their decisions on their good-faith beliefs about how things are likely to turn out in the future. The duty of loyalty, then, demands that directors’ decisions be based on a good-faith belief that the chosen course of action is the best one for the corporation. Thus, good faith is the state of mind characteristic of directorial decision making that complies with the duty of loyalty.51

The directors’ belief that a decision will benefit the corporation is developed by gathering information relevant to the decision at hand. This process implicates the duty of care. Care and loyalty relate to one another as do means and ends. Loyalty defines the end to be served by the decision—the best interests of the corporation or its stockholders. Care is the means by which the decision is made. “Good faith” is the state of mind, the belief, that animates the decision.

It could be argued that a director who failed to exercise reasonable care in forming her decision lacked the informational foundation required to entertain a reasonable belief regarding the company’s best interests. But this would transform every breach of the duty of care into a breach of the duty of loyalty. That would be inconsistent with 102(b)(7), which allows a corporation to protect directors from liability for breach of the duty of care, but not a breach of the duty of loyalty.52

The better approach is to focus on the belief of the director at the time of the decision making. A negligent director can still hold an honest belief that the decision is in the best interests of the company. Moreover, a decision-making process found to be deficient by a court in hindsight may have been followed by directors who believed, in good faith, that they were doing the right thing. And even a director who is subsequently found to be negligent may have reached the challenged decision with a good-faith belief that it would serve the best interests of the corporation.53


51 In Gutman v. Huang, 823 A.2d 492, 506 n.34 (Del. Ch. 2003), the court stated: “A director cannot act loyally towards the corporation unless she acts in the good faith belief that her actions are in the corporation’s best interest.” See also Cede & Co. v. Technicolor, Inc., 634 A.2d 345, 361 (Del. 1993) (defining good faith as loyalty); In re Gaylord Container Corp. S’holders Litig., 753 A.2d 462, 475 (Del. Ch. 2000) (connecting good-faith belief to the duty of loyalty).


53 In Smith v. Van Gorkom, 488 A.2d 858, 873 (Del. 1985), the court observed that the directors had acted in good faith, even though their decision-making process was deemed to be grossly negligent.
V. THE JURISPRUDENCE OF BAD FAITH

Delaware courts have concluded that “at least in the corporate fiduciary context, it is probably easier to define bad faith rather than good faith.” But the effort to define “good faith” as the absence of “bad faith” tends to narrow the scope of good faith to the absence of particular categories of extremely bad conduct deemed to be in bad faith. This approach tends to deemphasize directors’ affirmative duty to base their decisions and actions on the best interests of the corporation and the stockholders. Good faith requires more than simply avoiding bad conduct.

However, the Delaware Supreme Court has recently defined “bad faith” in a way that avoids this problem. It has identified four categories of conduct that constitute bad faith: (A) conduct motivated by a purpose other than the best interests of the corporation, (B) conduct motivated by subjective bad faith, (C) conduct that involves an intentional violation of law, and (D) dereliction of duty. Each will be discussed in the following sections.

A. Decisions Motivated by an Improper Purpose

The first category refers to conduct motivated by a purpose other than the best interests of the corporation. The breadth of this category prevents Delaware’s definition of bad faith from unduly narrowing the meaning of good faith. The phrase “conduct motivated by a purpose other than the best interests of the corporation” defines the universe of conduct that would violate the duty of loyalty. The court’s use of the word “motivation” implicitly recognizes the connection between a good-faith belief and the duty of loyalty. The duty of loyalty requires that directors’ actions be motivated by a “good-faith” belief that their actions will serve the best interests of the corporation. Actions motivated by any purpose other than serving the best interests of the corporation would, therefore, be a violation of the duty of loyalty.

In effect, Delaware’s definition of bad faith reverses the relationship between good faith and bad faith. Delaware does not define “good faith” as the absence of a narrowly defined set of circumstances that demonstrate bad faith. Rather, Delaware defines bad faith as any conduct not motivated by a good-faith belief that the conduct will serve the best interests of the corporation. In other words, bad faith

\[54\] Disney IV, 907 A.2d 693, 753 (Del. Ch. 2005); see also Disney V, 906 A.2d 27, 63–67 (Del. 2006) (holding that “bad faith” was a violation of the duty of good faith).

\[55\] See Nowicki, supra note 5, at 530–32 (discussing directors’ good faith).

\[56\] Id.

\[57\] See Stone v. Ritter, 911 A.2d 362, 369 (Del. 2006); Disney V, 906 A.2d at 67.

\[58\] See Disney V, 906 A.2d at 64.

\[59\] See Stone, 911 A.2d at 369; Disney V, 906 A.2d at 67.

\[60\] See Stone, 911 A.2d at 369 (citing Disney V, 906 A.2d at 67).

\[61\] Guttman v. Huang, 823 A.2d 492, 506 n.34 (Del. Ch. 2003).
refers to any conduct that violates the duty of loyalty. Thus, in Delaware, the phrase “bad faith” is simply a shorthand way of referring to the absence of good faith, a circumstance that necessarily violates the duty of loyalty. The remaining three categories of “bad faith”—intent to harm the corporation, violation of law, and dereliction—are merely particular examples of conduct contrary to the best interest of the corporation, which therefore violates the duty of loyalty. The jurisprudence in this area adds detail to the understanding of good faith.

B. Decisions Motivated by Subjective Bad Faith

In *Disney V*, the court said that “bad faith” included “subjective bad faith,” which it defined as fiduciary conduct motivated by an actual intention to do harm to the company or its stockholders. The court said the proposition that such conduct constitutes a violation of fiduciary duty is so obvious that it requires no further explanation. But the apparent simplicity of this proposition conceals the more important point that such a motivation would also make the decision a violation of the duty of loyalty. As discussed above, the duty of loyalty requires directors to base each of their decisions on a good-faith belief that the decision will serve the corporation’s best interest. A decision intended to harm the corporation can hardly be described as being motivated by such a belief. Accordingly, a decision motivated by intent to harm the corporation is a violation of the duty of loyalty.

C. Intentional Violation of Law

In *Disney V*, the court also stated that a decision to intentionally violate the law reflects “bad faith.” Again, this does little to advance the analysis. Some argue that the corporation’s best interests should be defined in terms of a single-

---

62 See Stone, 911 A.2d at 370.
63 Id.
64 Conduct that is intended to harm the corporation is clearly not consistent with good faith. The intentional violation of law and dereliction of duty are also examples of conduct that will harm the corporation and, thus, are contrary to the duty of loyalty.
65 The court stated, “The first category involves so-called ‘subjective bad faith,’ that is, fiduciary conduct motivated by an actual intent to do harm. That such conduct constitutes classic, quintessential bad faith is a proposition so well accepted in the liturgy of fiduciary law that it borders on axiomatic.” *Disney V*, 906 A.2d at 64.
66 Id.
67 See Nagy v. Bistricer, 770 A.2d 43, 49 n.2 (Del. Ch. 2000) (“By definition, a director cannot simultaneously act in bad faith and loyally towards the corporation and its stockholders.”); see also *In re ML/EQ Real Estate P’ship Litig.*, No. 15741, 1999 WL 1271885, at *4 n.20 (Del. Ch. Dec. 21, 1999). In *Guttman v. Huang*, 823 A.2d 492, 506 n.34 (Del. Ch. 2003), the court observed “there is no case in which a director can act in subjective bad faith towards the corporation and act loyally.”
68 906 A.2d at 67.
minded pursuit of profit. Under this view, a board could reasonably decide to bend the rules if the potential gain that could be achieved by illegal conduct outweighed the cost of paying the fine, discounted by the probability of getting caught.

This calculus would justify making illegal payments to doctors in exchange for prescribing medications sold by the corporation or bribing foreign officials to build the company’s business in a foreign nation. The problem with this calculus is that the duty of loyalty requires the board to serve the best interests of the corporation within the boundaries of the law. The board should try to win the game, but it must also play by the rules. If a company sustains a loss as a result of board-authorized illegal conduct, the directors should be liable for the net loss caused by their illegal decision. Moreover, it is not necessary to define intentional violation of the law as “bad faith” (i.e., “not in good faith”) to make sure such culpable directors are held personally liable for their illegal conduct. Section 102(b)(7) expressly excludes acts involving “intentional misconduct” or a “knowing violation of law” from the scope of exculpation.

D. Intentional Dereliction of Duty

1. Good Faith and the Duty to Monitor

Caremark was the first judicial effort to define “good faith” for purposes of determining whether a board might be protected from personal liability by a section 102(b)(7) exculpation provision in the corporation’s charter. The issue before the court was whether to approve a settlement of a stockholder derivative action seeking to hold the corporation’s directors personally liable for criminal fines paid by the corporation to resolve an indictment in which the corporation had been charged with violating a federal law. The plaintiff argued that the alleged violations of law resulted from the board’s failure to actively monitor the

---

69 Stephen M. Bainbridge et al., The Convergence of Good Faith and Oversight, 55 UCLA L. Rev. 559, 592–94 (2008) (arguing that pursuit of the corporation’s best interests can, in some cases, justify minor violations of law).

70 Id.

71 In In re Caremark International, Inc. v. Derivative Litigation, 698 A.2d 959, 962 (Del. Ch. 1996), the corporation was accused of violating the Anti-Referral Payments Law, which makes it illegal for a provider of medical services, like Caremark, to pay physicians for referrals.


73 Disney V, 906 A.2d at 67.

74 DEL. CODE ANN. tit. 8, § 102(b)(7)(ii) (2001); see also Bainbridge et al., supra note 69, at 591(referring to section 102(b)(7)'s “exculpation of ‘a knowing violation of law’”).

75 See 698 A.2d at 970.

76 Id. at 960–61.
employees’ compliance with the law. The defendants argued that they were protected from financial liability for this alleged breach of their duty of care by a 102(b)(7) exculpation provision in the corporation’s charter.

In assessing the strength of the plaintiff’s claim, the court observed that:

[A] director’s obligation includes a duty to attempt in good faith to assure that a corporate information and reporting system, which the board concludes is adequate, exists, and that failure to do so under some circumstances may, in theory at least, render a director liable for losses caused by non-compliance with applicable legal standards.

To emphasize the significance of the phrase “good faith” in the quoted passage, the court inserted a footnote that warned: “Moreover, questions of waiver of liability under certificate provisions authorized by 8 Del. C. § 102(b)(7) may also be faced.”

Because Caremark’s certificate of incorporation contained an exculpation provision, even gross negligence in the board’s performance of its statutory duty to monitor would not be enough to support the imposition of personal liability. But, the court explained, a sustained or systematic failure to exercise oversight would be evidence of a lack of good faith.

[I]n my opinion only 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. Such a test of liability—lack of good faith as evidenced by sustained or systematic failure of a director to exercise reasonable oversight—is quite high. But, a demanding test of liability in the oversight context is probably beneficial to corporate shareholders as a class, as it is in the board decision context, since it makes board service by qualified persons more likely, while continuing to act as a stimulus to good faith performance of duty by such directors.

By connecting oversight to the duty of care, Caremark appeared to suggest that egregiously reckless performance of duty would be sufficient to show that the board had not made a good-faith effort to satisfy its duty of care, and this absence of good faith would allow the imposition of financial liability on the ultra-careless directors. Some, reading Caremark through the lens of Cede II’s triadic conception

77 Id. at 964.
78 Id. at 965.
79 Id. at 970.
80 Id. at 970 n.27.
81 See id. at 965, 970.
82 Id. at 971.
of fiduciary duties, thought that Caremark had defined an independent fiduciary duty of good faith that would be breached by ultra-careless decision making.

2. Good Faith and Careless Decision Making

The next step came when plaintiffs sought to use Caremark’s theory that dereliction of duty constituted bad faith as the basis for a claim to recover damages from the directors in the context of a challenge to the directors’ performance of their duties in the decision-making context. Judicial review of a board’s performance of its fiduciary duties in the decision-making context focuses on the directors’ duties of loyalty and care. Under the business judgment rule, the board is presumed to have exercised due care and to have based its decision on a belief that it would serve the best interests of the corporation. To successfully challenge the board’s decision the plaintiff must introduce facts that call the loyalty or the diligence of a majority of the decision-making directors into question. But in an action to recover financial damages and where the corporation’s certificate contained an exculpation provision, even a successful challenge to the board’s compliance with its duty of care would be unavailing because the directors would be shielded from personal liability.

To get around the exculpation provision, plaintiffs sought to extend Caremark’s abdication-of-duty theory from the oversight context to the decision-making context. The logic was this: Caremark held that a sustained and systematic failure to act where there is a duty to act is evidence of a lack of good faith. A sustained or systematic failure to satisfy the duty of care is, therefore, evidence that the directors lacked good faith. Thus, a sustained and systematic failure to use due care in the decision-making process would not be protected from liability by a section 102(b)(7) exculpation provision.

---

83 See supra notes 20–21 and accompanying text.
84 This argument was advanced by the plaintiffs in the Disney litigation. See Disney V, 906 A.2d 27, 62–63 (Del. 2006).
85 Furlow, supra note 50, at 539–43.
86 Id. at 539.
87 Id. at 541–42.
88 There is another theory that might allow plaintiffs to place a claim beyond the reach of a section 102(b)(7) exculpation provision. Section 102(b)(7)(ii) states, in pertinent part, that an exculpation provision may not shield a director from personal liability for breach of fiduciary duty involving “intentional misconduct.” Del. Code Ann. tit 8, § 102(b)(7)(ii) (2001). The word “misconduct” is broad. A breach of fiduciary duty (including the duty of care) would fall within the meaning of “misconduct.” Thus, a director who decides (an intentional act) to be grossly negligent in the performance of his directorial duties would not be protected by a section 102(b)(7) provision. Stated more affirmatively, he could be held personally liable for financial damages caused by his intentional gross negligence. It might be easier to convince a court that a director’s gross negligence was intentional (and not inadvertent) than to convince a court that the director completely abdicated his duty of care.
This view appeared to receive some support in Chancellor Chandler’s decision in Disney III. The plaintiffs’ amended complaint catalogued numerous defects in the process by which the board had decided to hire Michael Ovitz to serve as Disney’s president and the process by which, after fourteen months of lackluster performance, the board allowed Mr. Ovitz to resign under circumstances that entitled him to a $140 million severance package. The defendants moved to dismiss on the ground that the amended complaint merely alleged that the directors’ decision-making process failed to satisfy their duty of care, a breach of duty from which they were protected from personal liability by an exculpation provision in the Disney certificate of incorporation.

The court rejected this argument, holding that the plaintiffs’ allegations did more than merely describe negligent or grossly negligent decision making. Rather, the allegations suggested “that the Disney directors failed to exercise any business judgment and failed to make any good faith attempt to fulfill their fiduciary duties to Disney and its stockholders.” Such allegations, said the court, support the conclusion that the defendant directors “abandoned all responsibility to consider appropriately an action of material importance to the corporation.” This, the court added, “puts directly in question whether the board’s decision-making processes were employed in a good faith effort to advance corporate interests. In short, the new complaint alleges facts implying that the Disney directors failed to ‘act in good faith and meet minimal proceduralist standards of attention.’” The court explained:

[T]he facts alleged in the new complaint suggest that the defendant directors consciously and intentionally disregarded their responsibilities, adopting a “we don’t care about the risks” attitude concerning a material corporate decision. Knowing or deliberate indifference by a director to his or her duty to act faithfully and with appropriate care is conduct, in my opinion, that may not have been taken honestly and in good faith to advance the best interests of the company. . . . Viewed in this light, plaintiffs’ new complaint sufficiently alleges a breach of the directors’ obligation to act honestly and in good faith in the corporation’s best interests . . . .

Thus, it appeared that a conscious disregard of the duty of care in the context of decision making was an act or omission, not in good faith, from which the directors would not be protected by an exculpation provision. In other words, a

---

89 Disney III, 825 A.2d 275, 279 (Del. Ch. 2003).
90 Id. at 286.
91 Id. at 278.
92 Id.
93 Id.
94 Id.
95 Id. at 289.
willful breach of the duty of care was worse than gross negligence and would, therefore, support a claim for monetary damages against the directors.

In his post-trial opinion, Disney IV, Chancellor Chandler found that, as a matter of fact, the directors had not acted in bad faith. However, the decision again suggested that a conscious disregard of the duty of care would support a claim for damages against the directors: “Upon long and careful consideration, I am of the opinion that the concept of intentional dereliction of duty, a conscious disregard for one’s responsibilities, is an appropriate (although not the only) standard for determining whether fiduciaries have acted in good faith.”

3. Dereliction of Duty Versus Gross Negligence

In Disney V, the Delaware Supreme Court acknowledged that intentional dereliction of duty was a type of bad faith that should not be entitled to exculpation. But, regarding an alleged dereliction of the duty of care, the court was careful to draw a distinction between “gross negligence” (the degree of carelessness required to establish a breach of the duty of care) and “intentional dereliction” of the duty of care, which would constitute bad faith. The court explained that the distinction between “gross negligence” and “bad faith” was required by section 102(b)(7), which only allows corporations to shield directors from liability for gross negligence and does not allow them to shield directors from acts or omissions not in good faith. Thus, intentional dereliction of the duty of care would be treated as conduct that cannot be shielded from personal liability.

As a general matter, the distinction between dereliction of the duty of care and gross negligence makes logical sense. There is a difference between choosing to

---

96 Disney IV, 907 A.2d 693, 760–72 (Del. Ch. 2005).
97 Id. at 755.
98 Disney V, 906 A.2d 27, 66 (Del. 2006).
99 Id. at 64 (placing gross negligence in a separate category from intentional dereliction of duty). The court noted that an intentional failure to be informed because of hostility to the company would be intentional bad faith. Id. at 65 n.104. Such ignorance would be a breach of the duty of care, but the intent to harm the company would also make it bad faith. Id. at 64. What the court overlooks is that the motive also makes it a violation of the duty of loyalty.
101 Stone v. Ritter, 911 A.2d 362, 370 (Del. 2006). To establish liability for director oversight, the plaintiff must show one of two things. Either (1) that the directors utterly failed to implement any reporting or information system or controls, or (2) having implemented such a system, the directors consciously failed to monitor its operation, and they knew they were not discharging their fiduciary obligations. Id. The court explained: “Where directors fail to act in the face of a known duty to act, thereby demonstrating a conscious disregard for their responsibilities, they breach their duty of loyalty by failing to discharge that duty in good faith.” Id.
102 Gross negligence in the context of a director’s fiduciary duties requires “reckless indifference to or deliberate disregard of” the interests of the corporation or its...
exercise no care at all, and exercising so little care that the directors’ actions are deemed to be grossly negligent. But when a court reviews the directors’ compliance with their duty of care in the decision-making process, the concept of intentional dereliction of duty has little, if any, application.

The idea that dereliction of duty constitutes bad faith (or an absence of good faith) was developed in cases dealing with the board’s duty to monitor. That duty, by definition, involves an activity which occurs over a period of time. Accordingly, it is easy to determine whether there has been a sustained and systematic failure to perform the duty to monitor. But the decision-making process is different from the monitoring process. The decision-making process occurs at a board meeting. Judicial review of a board’s decision-making process necessarily focuses on the care with which the board reached the challenged decision. To make a decision, the board must hold a meeting. At the meeting, there must be a sufficient number of directors to form a quorum. After at least a cursory discussion of the issue at hand, the board reaches a decision. The board may have been grossly negligent, but it would not have “knowingly and completely failed to undertake [its] responsibilities.” It would take an extreme set of facts to conclude that the directors who participated in that process had completely abdicated their duty of care.

One case that presented such extreme facts was *Sealy Mattress Co. of New Jersey, Inc. v. Sealy, Inc.* However, that case did not address the issue of good faith. It was decided before bad faith had become an important issue, and it


103 See supra notes 78–87 and accompanying text.

104 Section 141(f) permits a board to take action by unanimous written consent without a meeting. Del. Code Ann. tit. 8, § 141(f) (2001). A director could affix her signature to a consent without having given the matter any consideration at all. Such an act would be a clear example of an intentional disregard of the duty of care.

105 Unless the corporation’s certificate of incorporation or its bylaws require more than half, a majority of the total number of directors constitutes a quorum. Id. A quorum must be present for the board to transact business. Id.


107 Id. at 243 (quoting In re Lear Corp. S’holders Litig., 967 A.2d 640, 654–55 (Del. Ch. 2008). In Lear, the court said:

When a discrete transaction is under consideration, a board will always face the question of how much process should be devoted to that transaction given its overall importance in light of the myriad of other decisions the board must make. . . . Courts should therefore be extremely chary about labeling what they perceive as deficiencies in the deliberations of an independent board majority over a discrete transaction as not merely negligence or even gross negligence, but as involving bad faith.

967 A.2d at 654.

108 532 A.2d 1324 (Del. Ch. 1987).
involved an application for injunctive relief; thus, there was no need to fashion a
tory to avoid the corporation’s exculpation provision. The Sealy case involved
a cash-out merger in which Sealy’s parent corporation sought to eliminate Sealy’s
sole minority stockholder. The directors of Sealy, all of whom were employed
by the parent corporation, approved the merger in a telephone meeting that lasted
between fifteen and thirty minutes. They reviewed no financial information
before the meeting, and they had no financial information available to them at the
meeting. The directors claimed they were functioning not in a fiduciary capacity,
but in a ministerial capacity to carry out the parent corporation’s bidding. The
resolution by which they approved the merger had been prepared for them by the
parent corporation’s lawyer. The merger price was set by the parent corporation,
and the directors of the subsidiary made no effort to determine whether the price
was within the range of fairness. The court found that they “were completely
uninformed and made no effort to inform themselves of the material facts.”

These days, if a plaintiff sought to recover damages from the directors based
on similar facts, the court would probably find that the directors had completely
disregarded their duty of care. Because the directors could not have had a good-
faith belief that the minority stockholders’ best interests would be well served by
their complete abdication of their duty of care, their knowing dereliction of duty
would constitute bad faith and a breach of the duty of loyalty. But, short of such
extreme and unlikely facts, it would seem that so long as the directors held a
meeting and discussed—even in a very superficial way—the matter before them,
they could not be said to have abdicated their duty of care. Their conduct would
surely be grossly negligent, but they would not have intentionally disregarded their
duty of care. Thus, the concept of dereliction of duty would seem to have rare
application in the context of judicial review of the directors’ decision-making
process.

---

109 Id. at 1326 (noting plaintiff sought injunctive relief).
110 Id.
111 Id. at 1330.
112 Id. (noting the directors only had a copy of the resolutions).
113 Id. at 1337.
114 Id. at 1333.
115 Id.
116 Id.
117 In Disney III, Chancellor Chandler said dereliction of the duty of care would be
established by proof that “the defendant directors knew that they were making material
decisions without adequate information and without adequate deliberation, and that they
simply did not care if the decisions caused the corporation and its stockholders to suffer
injury or loss.” 825 A.2d 275, 289 (Del. Ch. 2003).
118 Lyondell Chem. Co. v. Ryan, 970 A.2d 235, 240–41 (Del. 2009); In re Lear Corp.
E. Lyondell Chemical Co. v. Ryan

The folly of attempting to shoehorn the dereliction-of-duty concept into a review of a board’s decision-making process is illustrated by the Court of Chancery’s decision\(^{119}\) that was reversed by the Supreme Court in *Lyondell Chemical Co. v. Ryan*.\(^{120}\) In that case, the plaintiffs, who were stockholders of Lyondell Chemical Co., sought to impose personal liability on the directors who approved the sale of the company via a cash merger.\(^{121}\) The plaintiff alleged the directors failed to take appropriate steps to achieve the best available price, in violation of their *Revlon* duties.\(^{122}\) Ten of the eleven directors/defendants were, in the court’s words, “well-credentialed, independent directors.”\(^{123}\) The price at which the challenged merger was accomplished was conceded to be within the range of fairness.\(^{124}\) But the plaintiffs argued that the board did not act aggressively enough in its efforts to achieve an even higher price and it capitulated too easily to the acquirer’s insistence on certain deal protective measures in the merger agreement.\(^{125}\)

These arguments, which focused on the process by which the board negotiated and approved the merger, challenged only the board’s compliance with its duty of care, a breach of duty from which the directors were protected by a 102(b)(7) exculpation provision in the corporation’s certificate.\(^{126}\) Nonetheless, the court denied the defendants’ motion for summary judgment, concluding that the directors (1) failed to negotiate aggressively enough, (2) failed to perform a pre-signing market check, and (3) agreed too easily to protection measures that would hobble a post-signing market check.\(^{127}\) These facts, said the Court of Chancery, were sufficient—for summary judgment purposes—to support the inference that


\(^{120}\) 970 A.2d 235 (Del. 2009).

\(^{121}\) The merger had been consummated, so equitable remedies were no longer available. An award of financial damages against the directors was the only remedy still available to plaintiffs. *Ryan*, 2008 WL 2923427 at *1 n.4.

\(^{122}\) Id. at *2. “Revlon duties,” so called, are defined in *Revlon, Inc. v. MacAndrews & Forbes Holdings*, 506 A.2d 173, 179 (Del. 1986). See Furlow, supra note 50.


\(^{124}\) Id. at *17 (noting that the deal price was “fair” and was in the middle of the various ranges of fairness developed by the corporation’s financial advisors).

\(^{125}\) The merger agreement included a “no shop” provision that prevented Lyondell from launching an active post-signing market check. The board also acquiesced in the purchaser’s demand to include a termination fee of about 3 percent if the deal did not close. Id. at *8, *16–17.

\(^{126}\) Lyondell’s certificate of incorporation included a section 102(b)(7) exculpation provision, so the plaintiff was forced to ground his *Revlon* claim on a failure to act in good faith in violation of the duty of loyalty. Id. at *18.

\(^{127}\) Id. at *15–16.
the board had intentionally disregarded its *Revlon* duties and thus acted in bad faith, a violation of its duty of loyalty.128

The Delaware Supreme Court took the unusual step of allowing the defendants to submit an interlocutory appeal from the Court of Chancery’s denial of their motion for summary judgment.129 Then, in an en banc decision, the Supreme Court reversed the trial court’s decision and remanded with directions that summary judgment be entered in favor of the defendants.130

The Supreme Court reaffirmed the proposition that the directors’ duty of loyalty required them to perform their decision-making function in good faith.131 The court also acknowledged that an intentional dereliction of a known duty would reflect an absence of good faith, which would constitute a violation of the duty of loyalty.132 The Delaware Court of Chancery misapplied these principles because it failed to recognize that deficiencies in the process by which a board undertakes to sell a corporation merely involve issues related to the board’s duty of care.133 But that is not the same as intentionally disregarding *Revlon* duties.

*Revlon* and its progeny stand for the proposition that when a board undertakes to sell a company, it assumes a duty “to ‘[g]et] the best price for the stockholders at a sale of the company.’”134 But *Revlon* did not define particular ways in which a board must perform its *Revlon* duties.135 Rather, the *Revlon* doctrine recognizes “there is no single blueprint that a board must follow to fulfill its [Revlon] duties.”136 Accordingly, “[n]o court can tell directors exactly how to accomplish that goal, because they will be facing a unique combination of circumstances, many of which will be outside their control.”137

The process followed by the Lyondell board may have been flawed, but “there is a vast difference between an inadequate or flawed effort to carry out fiduciary duties and a conscious disregard for those duties.”138 Only a complete and utter failure to seek confirmation that the deal under consideration represents the best reasonably available price would implicate the kind of bad faith constituting a

128  *Id.* at *19.


130  Lyondell Chem. Co. v. Ryan, 970 A.2d 235, 244 (Del. 2009).

131  *Id.* at 240 (discussing *Disney V*, 906 A.2d 27 (Del. 2006)).

132  *Id.* (discussing *Stone v. Ritter*, 911 A.2d 362 (Del. 2006)).

133  *Id.* at 243–44.

134  *Id.* at 242 (quoting *Revlon*, Inc. v. MacAndrews & Forbes Holdings, 506 A.2d 173, 182 (Del. 1986)).

135  *Id.* at 242–43 (holding the trial court’s conclusion that the directors must follow one of several courses of action was erroneous and explaining the trial court’s decision was based on the incorrect view that *Revlon* required directors to follow one of three courses of action: an auction, a pre-signing market survey, or market research that provides a precise knowledge of the company’s value).

136  *Id.* at 242–43 (quoting *Barkan v. Amsted Indus.*, Inc., 567 A.2d 1279, 1286 (Del. 1989)).

137  *Id.* at 242.

138  *Id.* at 243.
breach of the duty of loyalty. Indeed, that the process was undertaken at all reflects, at the least, an attempt to comply with the duty of care. The board may have failed to satisfy its duty of care, but it did not disregard its duty of care. Even a careless board can form a good-faith belief that its decision will serve the best interests of its constituents.

F. The Proper Role of Good Faith in Fiduciary Analysis

Good faith is the state of mind with which a director performs her duty of loyalty. The duty of loyalty requires directors to base their decisions on a good-faith belief that the decisions serve the best interests of the corporation. Any decision that is not based on such a belief is a violation of the duty of loyalty. Thus, a decision motivated by a desire to harm the corporation is made with “subjective bad faith,” and this improper motive means the decision is a violation of the duty of loyalty. By the same token, a decision to cause the corporation to intentionally violate the law cannot be motivated by a good-faith belief that illegal conduct will serve the corporation’s best interests. A decision to disregard the statutory duty to monitor the corporation’s activities can hardly be thought to be motivated by a good-faith belief that such dereliction of duty will serve the corporation’s best interests. Similarly, a director’s decision to exercise no care at all in performing her decision-making function could hardly be thought to be motivated by a good-faith belief that such dereliction of duty would serve the corporation’s best interests. However, it will be an exceedingly rare case that presents such extreme facts.

The point that remains is that in judging a director’s compliance with her duty of loyalty, the court must determine whether her actions (in the monitoring context or the decision-making context) were based on a good-faith belief that they would serve the best interests of the corporation. This necessarily requires the court to draw an inference from objective facts about the director’s subjective motivation.

139 Id. at 244 (noting only a knowing and complete failure to undertake their responsibilities would be a breach of directors’ duty of loyalty).

140 An example of this circumstance is provided by McPadden v. Sidhu, in which the court concluded that the process by which the board reached the decision to sell one of the corporation’s divisions demonstrated “gross negligence” by the corporation’s directors, a violation of the duty of care from which the directors were shielded by the section 102(b)(7) exculpation provision in the corporation’s certificate of incorporation. 964 A.2d 1262, 1273–75 (Del. Ch. 2008). The court also concluded that, notwithstanding their carelessness, the directors had not made the decision in bad faith. Id. at 1277. The court explained that “a board of directors may act ‘badly’ without acting in bad faith.” Id. at 1263.

141 Stone v. Ritter, 911 A.2d 362, 370 (Del. 2006) (noting that a director’s duty of loyalty requires that her actions be based on a good-faith belief that they will serve the best interests of the corporation).

142 Id.

The process by which that is accomplished is governed by the business judgment rule.  

VI. GOOD FAITH AND THE BUSINESS JUDGMENT RULE

Business decisions require boards to make judgments about the future of corporations. Because the future is uncertain, directors cannot be held to a standard that would subject them to personal liability for decisions that turn out badly. The business judgment rule allows boards to take calculated business risks without fear of incurring personal liability. The rule accomplishes this end by preventing courts from reviewing the substantive merits of business decisions made with due care and based on a good-faith belief that the decisions would serve the corporation’s best interests. To successfully challenge a board decision, a plaintiff must prove the decision was made in violation of the directors’ fiduciary duties.

A. The Presumption of Care and Loyalty

The business judgment rule functions as a procedural rule and as a substantive rule of law.  

The vague verb “showing” is used here to embrace both the plaintiff’s burden of pleading such facts in the complaint and proving such facts at trial.

The rule requires the plaintiff to rebut this presumption by showing facts that give reason to doubt the directors’ performance of these duties. If the plaintiff fails to produce such evidence, the presumption that the directors acted with care and loyalty stands unrebutted. At this point the “substantive” aspect of the rule requires the court to defer to a business judgment made by these presumptively careful and loyal directors, provided their decision is not completely irrational. To rebut the

---

144 Furlow, supra note 50, at 539–44.
147 Cede, 634 A.2d. at 361; Aronson v. Lewis, 473 A.2d 805, 812 (Del. 1984); Orman v. Cullman, 794 A.2d 5, 19–20 (Del. Ch. 2002).
148 The vague verb “showing” is used here to embrace both the plaintiff’s burden of pleading such facts in the complaint and proving such facts at trial.
149 See Cinerama, Inc. v. Technicolor, Inc., 663 A.2d 1156, 1162 (Del. 1995); Cede, 634 A.2d. at 361. Although the rule speaks of this as imposing a burden on the plaintiff to rebut the presumption, it is really nothing different from the ordinary requirement that a plaintiff must establish facts to prove the elements of a breach of duty.
150 E. Norman Veasey, The Defining Tension in Corporate Governance in America, 52 BUS. LAW. 393, 394 (1997) (noting that under the business judgment rule, it is difficult to demonstrate that a decision was irrational); see also David Rosenberg, Galactic Stupidity
presumption that the directors complied with their fiduciary duties, a plaintiff may challenge the directors’ performance of their duties of care and loyalty.

B. Rebutting the Presumption of Care

The duty of care requires directors to “inform themselves, prior to making a business decision, of all material information reasonably available to them,”$^{151}$ and to act with requisite care in deliberating on their decision.$^{152}$ This requires directors to take an active and direct role in the decision-making process.$^{153}$ To establish a breach of the duty of care that will support a claim for damages, the plaintiff must show that the director acted with gross negligence.$^{154}$

Proof of a breach of the duty of care rests on objective facts. The court will look at the amount of time available to directors to prepare for the meeting,$^{155}$ the extent of the directors’ preparation for the meeting,$^{156}$ time spent by the directors at the meeting,$^{157}$ the type and quality of the advice available to the directors,$^{158}$ the directors’ participation in the meeting,$^{159}$ and the documents the directors

and the Business Judgment Rule, 32 J. CORP. L. 301, 321–22 (2006) (arguing that review for rationality should be expanded to reach extremely stupid decisions).


$^{152}$ See Cede, 634 A.2d at 367.


$^{154}$ See Van Gorkom, 488 A.2d at 873. Of course, there are few instances in which courts hold directors personally liable for damages resulting from breach of the duty of care. Van Gorkom is noteworthy because it was so unusual. See Lawrence A. Hamermesh, Why I Do Not Teach Van Gorkom, 34 GA. L. REV. 477, 489–91 (2000) (explaining that Van Gorkom is unusual because claims to recover damages from directors based on a breach of the duty of care are rare). Moreover, most corporations have section 102(b)(7) provisions in their certificates of incorporation. However, breach of the duty of care remains important in claims seeking injunctive relief and other equitable remedies that do not involve the imposition of financial damages on directors.


$^{156}$ See Cede, 634 A.2d at 371 (holding that the directors’ failure to inform themselves of all reasonably available information material to the decision at hand was a violation of the duty of care).

$^{157}$ See Van Gorkom, 488 A.2d at 874 (finding that a two-hour meeting was insufficient to satisfy the duty of care).

$^{158}$ See Sealy Mattress Co. v. Sealy, Inc. 532 A.2d 1324, 1337 (Del. Ch. 1987) (finding that a decision made without advice from knowledgeable experts violated the duty of care).

$^{159}$ See id. (finding the directors’ failure to consider facts relevant to the merger violated the duty of care).
reviewed. Such proof, standing alone, is sufficient to establish a breach of the duty of care. What remains is to establish that this lack of care caused damage to the corporation or its stockholders. In the tort context, the burden would be on the plaintiff to prove damages proximately caused by the defendant’s negligence. But in the fiduciary duty context, the grossly negligent directors must prove that their grossly negligent decision making caused no harm. They do this by proving that the challenged transaction was entirely fair to the corporation.

C. Rebutting the Presumption of Loyalty

The evidence required to rebut the presumption that the directors complied with their duty of loyalty is different from evidence required to establish a breach of the duty of care. The duty of loyalty requires a director to base her decision on a good-faith belief that it will serve the best interests of the corporation. Thus, a breach of the duty of loyalty turns on the directors’ subjective motivation. Unlike a breach of the duty of care, which can be established by direct evidence, a finding that a director breached her duty of loyalty requires the court to draw inferences about the director’s subjective motivations from surrounding circumstances and the quality of the decision being attacked by the plaintiff. The plaintiff must introduce facts from which the court can infer there is reason to doubt the director’s decision was based on a good-faith belief that it would serve the corporation.

1. Financial Interest

Classically, two fact patterns create reason to doubt the good-faith basis of a director’s decision. The first involves a director who has a direct, or indirect, financial interest in the pending decision that conflicts with the corporation’s

---

160 See Van Gorkom, 488 A.2d at 874 (finding the directors’ failure to review merger documents violated the duty of care).
161 See Cede, 634 A.2d at 367; Van Gorkom, 488 A.2d at 874.
162 See Cede, 634 A.2d at 367.
164 See id. at 1302.
165 Cf. Cede, 634 A.2d at 367.
166 Id. at 361; Guttman v. Huang, 823 A.2d 492, 506 n.34 (Del. Ch. 2003).
167 Citron v. Fairchild Camera & Instrument Corp., No. 6085, 1988 WL 53322, at *15 (Del. Ch. May 19, 1988). The court said: “As in other contexts, however, this inquiry into a subjective state of mind necessarily will require inferences to be drawn from overt conduct—the quality of the decision made being one notable possible source of such an inference.” Id.
interests.\textsuperscript{168} The second involves a director who is financially beholden to a person whose interests in the transaction conflict with those of the corporation.\textsuperscript{169} Both involve situations in which the director has a personal financial interest that gives reason to doubt the director’s ability to make an impartial assessment of the corporation’s best interests. Proof that the director’s impartiality has been compromised is not, however, sufficient to establish that the compromised director violated her duty of loyalty.\textsuperscript{170} It is always possible the compromised director may have subordinated her self-interest to the interests of the corporation and approved the challenged decision with a good-faith belief that it would well serve the corporation.\textsuperscript{171} In that circumstance, the duty of loyalty would have been served.\textsuperscript{172} In recognition of this possibility, the court will allow the compromised directors an opportunity to prove that the transaction was entirely fair to the corporation. If they succeed, the court infers that the decision was motivated by the best interests of the corporation in compliance with the directors’ duty of loyalty. If they fail, the court infers that the decision was based on a purpose other than the best interests of the corporation, namely the conflicting interest that placed the directors’ impartiality in question. Thus, the court concludes the directors breached their duty of loyalty.

2. **Lack of Good Faith**

Lack of independence or lack of disinterestedness are not the only circumstances that give reason to doubt whether the challenged decision was based on the directors’ good-faith belief that it would serve the best interests of the corporation. In *Stone*, the court made it clear that duty of loyalty claims are not

\begin{itemize}
  \item \textsuperscript{168} This would be the case where the director stands on both sides of the transaction, see *Nixon v. Blackwell*, 626 A.2d 1366, 1374 (Del. 1993), or where the director has a financial interest in the party on the other side of the corporation. Even in an arm’s-length transaction, a director may have a material self-interest that will compromise his objectivity. *Cede*, 634 A.2d at 362.
  \item \textsuperscript{169} This is where directors are employed by a party interested in the transaction to be approved by the board. See *Weinberger v. UOP, Inc.*, 457 A.2d 701, 710 (Del. 1983) (noting that directors who were employed by, or served on the board of, the other party to a proposed merger had a conflict of interest). This also occurs where a director is beholden to an interested party. See *Heinemann v. Datapoint Corp.*, 611 A.2d 950, 955 (Del. 1992) (holding financial dealings between a director and an interested party are sufficient to compromise the director’s independence).
  \item \textsuperscript{170} *Cede*, 634 A.2d at 362.
  \item \textsuperscript{171} Id.
  \item \textsuperscript{172} It is no violation of the duty of loyalty to place the best interest of the corporation ahead of one’s own self-interest. In *Guttmann v. Huang*, 823 A.2d 492, 506 n.34 (Del. Ch. 2003), the court acknowledged it was possible that a director whose ability to make an impartial decision regarding the corporation’s best interest nonetheless acts in subjective good faith “(e.g., if the director is interested in a transaction subject to the entire fairness standard and cannot prove financial fairness).” Id.
limited to claims involving conflicts of interest or lack of independence. The duty of loyalty is also violated when directors base their decision on a motive other than a good-faith belief that it will serve the best interests of the corporation.

[Thus,] the business judgment form of judicial review encompasses three elements: a threshold review of the objective financial interests of the board whose decision is under attack (i.e., independence), a review of the board’s subjective motivation (i.e., good faith), and an objective review of the process by which it reached the decision under review (i.e., due care).

The fact that the directors who made the challenged decision were independent and disinterested does not guarantee that they have satisfied their duty of loyalty. It merely means that their impartiality was not compromised by financial self-interest. Other circumstances may cause a director to base a decision on reasons other than a good-faith belief that it will serve the corporation’s best interests. Vice Chancellor Strine recently observed that the concept of good faith functions as a “constant reminder (1) that a fiduciary may act disloyally for a variety of reasons other than personal pecuniary interest; and (2) that, regardless of his motive, a director who consciously disregards his duties to the corporation and its stockholders may suffer a personal judgment for monetary damages for any harm he causes.”

In Citron v. Fairchild Camera and Instrument Corp., the court observed:

While the absence of significant financial adverse interest creates a presumption of good faith, or a prima facie showing of it, the question of bona fides may not be finally determined on that basis alone. Rather, that question calls for an ad hoc determination of the board’s motives in this

---

173 Stone v. Ritter, 911 A.2d 362, 370 (Del. 2006). “[T]he fiduciary duty of loyalty is not limited to cases involving a financial or other cognizable fiduciary conflict of interest. It also encompasses cases where the fiduciary fails to act in good faith.” Id.

174 Id.

175 In re RJR Nabisco, Inc. S’holders Litig., No. 10389, 1989 WL 7036, at *13 (Del. Ch. Jan. 31, 1989); see also Disney V, 906 A.2d 27, 52 (Del. 2006) (stating the presumptions of the business judgment rule “can be rebutted if the plaintiff shows that the directors breached their fiduciary duty of care or of loyalty or acted in bad faith [and that,] [i]f that is shown, the burden then shifts to the director defendants to demonstrate that the challenged act or transaction was entirely fair to the corporation and its shareholders”); Grobow v. Perot, 539 A.2d 180, 187 (Del. 1988) (recognizing the role of good faith in the business judgment rule); Polk v. Good, 507 A.2d 531, 536 (Del. 1986); Smith v. Van Gorkom, 488 A.2d 858, 873 (Del. 1985); Aronson v. Lewis, 473 A.2d 805, 812 (Del. 1984).

176 Nagy v. Bistricer, 770 A.2d 43, 48–49 n.2 (Del. Ch. 2000); see also Guttman, 823 A.2d at 506 n.34 (“The reason for the disloyalty (the faithlessness) is irrelevant, the underlying motive (be it venal, familial, collegial, or nihilistic) for conscious action not in the corporation’s best interest does not make it faithful, as opposed to faithless.”).
particular instance. As in other contexts, however, this inquiry into a subjective state of mind necessarily will require inferences to be drawn from overt conduct—the quality of the decision made being one notable possible source of such an inference.\textsuperscript{177}

\section*{D. Establishing an Improper Motive}

Courts have examined the motives of independent and disinterested directors under two circumstances. The first involves those who approve corporate actions that also benefit persons with whom the directors have an affiliation. In this circumstance the decision may have been motivated by a desire to help the directors’ friends or colleagues.\textsuperscript{178} The prototypical example of this line of cases involves a board decision to take corporate action to defeat a hostile takeover. The second circumstance in which the court will question the motives of independent directors is where the probable outcome of the decision bares no rational relation to the corporation’s best interests.\textsuperscript{179}

\subsection*{1. Incidental Benefit to the Directors’ Affiliates}

Defensive action that protects a corporation and its stockholders from an unsolicited takeover bid also has the effect of protecting the directors’ positions on the board and senior managers’ jobs. That raises a question as to whether the decision to block the takeover was motivated by a good-faith belief that the takeover threatened the best interests of the corporation’s stockholders, or whether it was motivated by a desire to save the jobs.

Where a majority of the directors who approve the defensive action are independent outsiders, the argument that they were motivated to save their jobs is not enough to deny them the protection of the business judgment rule. Their interest in retaining their positions on the board is not deemed sufficient to


\textsuperscript{178} See, e.g., Strassburger v. Earley, 752 A.2d 557, 581–82 (Del. Ch. 2000) (observing that evidence showing a board decision was based on a desire to aid senior managers at the expense of the minority stockholders would be sufficient to show the decision was not made in good faith).

\textsuperscript{179} See, e.g., Thomas v. Kempner, No. 4138, 1973 WL 460, at *6 (Del. Ch. March 22, 1973) (noting insistence of directors in continuing to deal only with a company bidding a lower offer than with some new higher bidders, thereby ignoring the principle of competitive bidding); Robinson v. Pittsburgh Oil Ref. Corp., 126 A. 46, 48 (Del. Ch. 1924) (“[T]here is no appearance of a suspicion-arousing motive or interest peculiar to itself which would raise doubts concerning the faithfulness of the majority of the directors of the defendant corporation to their obligations to it in negotiating the sale in question.”).
compromise their independence. Thus, the business judgment rule would entitle them to the presumption that they acted in accordance with their duty of loyalty. The corporation’s senior managers, on the other hand, have a material stake in the outcome of a takeover. It threatens their positions as corporate officers, the primary source of their income. The question then becomes: was the independent directors’ decision to block a hostile takeover motivated by a perceived need to protect the corporation’s stockholders from threats posed by the takeover or by their desire to preserve their colleagues’ jobs and a desire to preserve their positions as directors? The issue is one of good faith.

Because both motives (defense of the corporation and retention of senior managers’ jobs and board seats) could reasonably be inferred from the decision to block the takeover, the Delaware Supreme Court, in the famous Unocal decision held that the directors must rebut the possible inference that they were motivated by an improper purpose. They do this by proving that the defense was intended to serve a valid corporate purpose (protection of stockholders from a structurally or substantively coercive tender offer) and was a reasonable way to accomplish that purpose. If a valid corporate purpose was present, the courts presume that it was the primary motive and accord the board’s decision the protections of the business judgment rule.

This mode of analysis, which the courts call “enhanced scrutiny,” is also applied when the motive behind an independent, disinterested board’s decision to sell the company appears to be influenced by considerations other than the best interests of the corporation or its shareholders. For example, in the Topps Co.
Shareholders Litigation the Court of Chancery was called on to review the decision-making process by which the board decided to sell the company to a favored bidder who was willing to continue to employ the company’s senior managers after it acquired control. The court enjoined the proposed transaction with the favored bidder because the board’s unwillingness to cooperate with the other bidder, who would have paid a higher price, may have been influenced by the favored bidder’s promise to preserve the jobs of senior corporate managers.

In the Netsmart Technologies, Inc. Shareholders Litigation, the court was again called on to review the process by which the board decided to sell the company. The plaintiffs argued that the board’s good faith was placed in issue by facts suggesting the board’s decision to seek bids exclusively from financial buyers and to exclude strategic buyers from the pool of potential purchasers might have been motivated by the fact that the financial buyers would be likely to retain existing management to run the company, whereas the strategic buyers would be likely to fire existing management and use their own managers to run the combined company.

Earlier cases followed the same approach. Because careful and disinterested directors are entitled to a presumption that they acted in good faith, these cases required the plaintiff to show that, viewed in light of facts known to the board at the time the decision was made, it was probable that the decision would benefit some interest other than the corporation’s best interests. The clearest example is where an apparently disinterested board makes a judgment that is essentially inexplicable except on the basis of an otherwise unproven inappropriate motive—such as personal favoritism or antipathy. Such a case might arise, for example, where an apparently disinterested board rejects a higher bid in favor of a lower one, on the same terms.

offered terms that would moot claims that would be asserted against board members by holders of certain corporate bonds); In re Topps Co. S’holders Litig., 926 A.2d 58, 92 (Del. Ch. 2007) (finding the board breached its fiduciary duty when it used terms in the acquisition agreement with the favored bidder, who offered to retain existing management, to block a pending offer of greater value); In re Netsmart Techs., Inc. S’holders Litig., 924 A.2d 171, 176, 198 (Del. Ch. 2007) (hinting, but declining to make a finding of fact, that board did not include strategic buyers among those to whom the company was offered because a strategic buyer would be less likely to retain existing management than would the financial buyers).

186 In re Topps, 926 A.2d at 63.
187 See id. at 91.
188 In re Netsmart Techs., Inc. S’holders Litig., 924 A.2d 171, 175–76 (Del. Ch. 2007).
189 Id. at 198.
Another example is provided by *Robinson v. Pittsburg Oil Refining Corp.*, where the court observed “there would be no conceivable justification” for a board decision to accept the lower of two similar bids for the purchase of certain corporate assets.\(^{191}\) Also, in *Thomas v. Kempner*, the court characterized the board’s decision to negotiate a possible sale of the corporation’s assets exclusively with the bidder who offered a price lower than other bids to be a “fundamental error of business judgment” that warranted injunctive relief.\(^{192}\)

The same approach can be generalized to other circumstances where directors have social or other connections with persons who stand to benefit from a corporate decision. In these circumstances, where the relationships do not involve the kind of influence that would call the directors’ independence into question, the directors should be required to show that their decision was motivated by a reasonable belief that it would serve a valid corporate purpose. Because these directors are still presumed to be independent (as that term is traditionally understood), if they can show that they reasonably believed their decision would serve a proper corporate purpose, their independence should entitle them to the presumption that the corporate purpose was the primary purpose and that any benefit to their friends and colleagues was merely incidental.

### 2. Irrational Business Judgments

The traditional business judgment rule allows courts to review the substantive merits of the challenged transactions approved by independent directors for the limited purpose of determining whether the decision was irrational,\(^{193}\) “egregious,”\(^{194}\) “beyond reason,”\(^{195}\) or a “gross abuse of discretion.”\(^{196}\) The question is whether the decision, viewed in light of facts as they were known to the directors at the time the decision was made, was so unlikely to benefit the

---

\(^{191}\) 126 A. 46, 49 (Del. Ch. 1924) (noting that if the only difference between two bids is the cash price, then acceptance of a lower bid would be indicative of fraud).

\(^{192}\) No. 4138, 1973 WL 460, at *6 (Del. Ch. Mar. 22, 1973); *see also* Bennett v. Breuil Petroleum Corp., 99 A.2d 236, 239 (Del. Ch. 1953) (noting plaintiff claimed defendants caused the corporation to issue new shares, not for a proper corporate purpose, but for the purpose of impairing plaintiff’s interest in the corporation “and to force him out of the corporation upon the management’s own terms”).

\(^{193}\) Sinclair Oil Corp. v. Levien, 280 A.2d 717, 720 (Del. 1971) (“A board of directors enjoys a presumption of sound business judgment, and its decisions will not be disturbed if they can be attributed to any rational business purpose.”).


\(^{195}\) *Id*.

corporation that it could not have been rationally based on a good-faith belief that it would serve the corporation’s best interests.

As a practical matter, the directors’ belief in the rightness of their decision-making conduct and the course of action only becomes an issue when decisions turn out badly. When a decision turns out well, its rightness is confirmed by events. That directors believed in good faith they were making the right decision becomes important when events prove that the directors’ belief was wrong. But the fact that a decision is proved by events to have been a bad one is not enough to call the board’s motivation into question. Decisions by independent and disinterested directors are upheld if the decisions are not irrational. If a fiduciary makes the wrong decision with the good-faith belief that it was right, then the fiduciary should be blameless. In that sense, “good faith” is a defense.

In the RJR Nabisco decision, the court explained that this limited review of the rationality of a challenged decision made by independent directors allows the court to determine whether the transaction is so “egregious” or reflects such a gross “abuse of discretion” that it supports the inference that its approval was not based on a good-faith belief that it would serve the best interests of the corporation. And in J.P. Stevens & Co., Shareholders Litigation, the court said it could “review the substance of a business decision made by an apparently well motivated board for the limited purpose of assessing whether that decision is so far beyond the bounds of reasonable judgment that it seems essentially inexplicable on any ground other than bad faith.”

197 Desert Equities, Inc. v. Morgan Stanley Leveraged Equity Fund, II, L.P., 624 A.2d 1199, 1208 (Del. 1993) (noting the definition of fraud is distinct from that of bad faith). The court defined “fraud” as “[a]n intentional perversion of truth for the purpose of inducing another in reliance upon it to part with some valuable thing belonging to him or to surrender a legal right.” Id. at 1208 n.16 (quoting BLACK’S LAW DICTIONARY 72 (5th ed. 1983)). Alternatively, the court defined “bad faith” as “not simply bad judgment or negligence, but rather . . . conscious doing of a wrong because of dishonest purpose or moral obliquity; it is different from the negative idea of negligence in that it contemplates a state of mind affirmatively operating with furtive design or ill will.” Id. (quoting BLACK’S LAW DICTIONARY 337 (5th ed. 1983)).

198 In re RJR Nabisco, Chancellor Allen said:

As I conceptualize the matter, such limited substantive review as the rule contemplates (i.e., is the judgment under review ‘egregious’ or ‘irrational’ or ‘so beyond reason,’ etc.) really is a way of inferring bad faith. I am driven to this view because I can understand no legitimate basis whatsoever to impose damages (or enter an injunction) if truly disinterested directors have in fact acted in good faith and with due care on a question that falls within the directors’ power to manage the business and affairs of the corporation.

1989 WL 7036, at *13 n.13 (citations omitted).

199 In re J.P. Stevens & Co., S’holders Litig, 542 A.2d 770, 780–81 (Del. Ch. 1988). There, the court explained that “this ‘escape hatch’ language has been variously stated in the Delaware opinions: ‘egregious’ decisions are said to be beyond the protections of the
The burden of proof in this effort is on the plaintiff. If the plaintiff can show facts that support the reasonable inference that a decision by independent, disinterested directors is based on a motive other than the best interests of the corporation, the decision is actionable for breach of the duty of loyalty. In *RJR Nabisco*, the court explained:

> [T]he protections of the business judgment rule [are not] available to a fiduciary who could be shown to have caused a transaction to be effectuated (even one in which he had no financial interest) for a reason unrelated to a pursuit of the corporation’s best interests. Greed is not the only human emotion that can pull one from the path of propriety; so might hatred, lust, envy, revenge, or, as is here alleged, shame or pride. Indeed any human emotion may cause a director to place his own interests, preferences or appetites before the welfare of the corporation. But if he were to be shown to have done so, how can the protection of the business judgment rule be available to him? In such a case, is it not apparent that such a director would be required to demonstrate that the corporation had not been injured and to remedy any injury that appears to have been occasioned by such transaction? Thus, the question here is not whether plaintiffs have asserted a legally cognizable theory with respect to their claim of bad faith, but whether it seems reasonably likely that they will be able to prove their assertion through the circumstances to which they point.

**VII. PROPOSED MODE OF JUDICIAL REVIEW OF GOOD FAITH**

The duty of loyalty requires directors to base each decision on a good-faith belief that it will serve the best interests of the corporation. An analysis of the directors’ good faith necessarily involves their subjective motivation. Motivation can be inferred from facts surrounding the business decision and by the substance of the decision itself.

---


201 1989 WL 7036, at *15.

202 For pleading purposes, intent and state of mind may be averred generally because “any attempt to require specificity in pleading a condition of mind would be unworkable and undesirable.” 5A CHARLES ALAN WRIGHT & ARTHUR R. MILLER, FEDERAL PRACTICE AND PROCEDURE § 1301 (3d ed. 2004). A claim of bad faith hinges on a party’s state of mind. See id. Under Rule 9(b), state of mind may be pled generally. See id. § 1301. Therefore, Delaware does not require bad faith to be pled with particularity. See Desert Equities, Inc., 624 A.2d at 1207.
Under the business judgment rule, a plaintiff can challenge directors’ good faith by proving the directors’ ability to make an impartial assessment of the corporation’s best interests was compromised by a direct or indirect financial self-interest. In this circumstance, the burden shifts to the directors to establish that the approved transaction was entirely fair to the corporation. If the transaction is fair, the court infers the directors were able to subordinate their self-interest and base their decision on the best interests of the corporation. If the transaction is not fair, the court reaches the opposite inference—the directors were motivated by self-interest rather than the best interests of the corporation.

In the absence of direct or indirect financial self-interest, the directors enjoy the presumption that their decision was made in good faith. But a plaintiff may place their good faith at issue by showing the substance of the decision itself belies the presumption that the directors acted in good faith. This requires the court to examine the probable result to be achieved by the decision. If it is assumed that a person intends to achieve the likely consequences of his actions, it follows that the board’s motive for making a particular decision can be inferred from the probable outcome to be achieved by that decision. Thus, if the likely outcome of a board decision is to increase corporate value, a court can infer that it was motivated by a good-faith belief that it would serve the corporation’s best interests. Conversely, where the probable outcome of a decision, viewed in light of the facts known to the board at the time the decision was made, has no rational relation to the corporation’s best interests, a court could reasonably conclude that the decision was based on a purpose other than the best interests of the corporation. In other words, it was not based on a good-faith belief that it would serve the corporation’s best interests and thus constitutes a violation of the duty of loyalty. Such a decision would be deemed “irrational” under the traditional business judgment rule.

Where a decision is as likely to serve the interests of the corporation as it is some other constituency, the board’s motive is ambiguous. In this circumstance, the plaintiff must establish facts that give reason to believe the board’s decision was more likely motivated by a desire to serve the interests of the other constituency than the interests of the corporation. To do this, the plaintiff must show that, although free from conflicting financial self-interest, the directors’ ability to make an impartial assessment of the corporation’s best interests was compromised by a softer form of influence—institutional bias. To do this, a plaintiff would have to establish that the directors shared an affinity with persons who stood to benefit from the decision and that, viewed in light of the facts as they were known by the directors at the time the decision was made, the decision would be of greater benefit to those persons than to the corporation. In this circumstance, the independent directors should be allowed to prove that their decision was based on a reasonable belief that it would serve a valid corporate purpose. Given their independence and their proof of a valid corporate purpose, they would be entitled to the presumption that their primary purpose was a valid corporate one and that the benefit to the other constituency was merely incidental to that purpose. But if the directors were unable to show the decision was motivated by a valid corporate purpose, it would follow that it was motivated by a purpose other than
the best interests of the corporation, namely the interests of the other constituency. That would establish a violation of the duty of loyalty.

VIII. CONCLUSION

The approach to good faith advocated by this Article is consistent with the purpose of section 102(b)(7), traditional understandings of directors’ fiduciary duties, and established Delaware practice under the business judgment rule.

Section 102(b)(7) is intended to protect directors from personal liability for breach of the duty of care. But it is not intended to protect directors from liability for breach of the duty of loyalty or an act or omission that is not in good faith. The approach to good faith advocated by this article serves both purposes. A board decision not motivated by a good-faith belief that it will serve the corporation’s best interests is not consistent with the duty of loyalty. Such an act will expose the directors to personal liability. But a board decision motivated by a good-faith belief that it will serve the corporation will be protected.

Traditionally, directors’ fiduciary duties have been thought to consist of a duty of loyalty and a duty of care. The duty of loyalty defines the goal toward which the director’s performance of her statutory duties of decision making and monitoring would be aimed—the best interests of the corporation. The duty of care defines how the directors should perform their duties—with the degree of prudence that a reasonable person in a similar situation would employ. Good faith recognizes that directors cannot be expected to know with certainty that their decisions will benefit the corporation. Accordingly, the duty of loyalty requires that the directors base their decisions on a good-faith belief that it will serve the corporation’s best interests. A decision that is motivated by any purpose other than a good-faith belief that it will serve the best interests of the corporation violates the directors’ duty of loyalty.

Finally, the approach to judicial review of the good-faith component of the duty of loyalty recommended by this article is consistent with traditional practice under Delaware’s business judgment rule. The duty of loyalty requires that directors base their decision on a good-faith belief that it will serve the corporation. Such a belief can be called into question by showing the directors’ impartiality was compromised by financial self-interest. In this circumstance, the compromised director can demonstrate her decision was motivated by a proper corporate purpose (and not the conflicting self-interest) by proving that the decision was entirely fair to the corporation.

In the absence of financial self-interest, directors are entitled to the presumption that their decision was based on the best interests of the corporation. But a showing by a plaintiff that the decision was not rationally related to the corporation’s best interests is sufficient to establish that the decision was not motivated by a good-faith belief that it would benefit the corporation. And where the directors’ motives are ambiguous, the directors may establish their good faith by demonstrating that they reasonably perceived that their decision would serve a valid corporate purpose.