ARTICLES

Law, Fact, and Discretion in the Federal Courts: An Empirical Study  
Robert Anderson IV

Bargained Justice: Plea-Bargaining’s Innocence Problem and the Brady Safety-Valve  
Lucian E. Dervan

Recantations Reconsidered: A New Framework for Righting Wrongful Convictions  
Adam Heder & Michael Goldsmith

Remedying the Misuse of Nature  
Sanne H. Knudsen

Due Process Denied: The Forgotten Constitutional Limits on Choice of Law in the Enforcement of Employee Covenants Not to Compete  
David A. Linehan

The Penguin and the Cartel: Rethinking Antitrust and Innovation Policy for the Age of Commercial Open Source  
Stephen M. Maurer

Class Action Defendants’ New Lochnerism  
Mark Moller

A Minimalist Approach to Same-Sex Divorce: Respecting States that Permit Same-Sex Marriages and States that Refuse to Recognize Them  
Robert E. Rains

The Common Law of Disability Discrimination  
Mark C. Weber

NOTES

E-Verify During a Period of Economic Recovery and High Unemployment  
Emily Patten

Incentive for Innovation or Invitation to Inhumanity?: A Human Rights Analysis of Gene Patenting and the Case of Myriad Genetics  
Laurie E. Abbott
SUBMISSIONS: The Utah Law Review receives submissions in March and April. Submissions should include a cover letter, providing the title of the work and the 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 2011–2012

BROWN, TENEILLE R. (2009), Associate Professor of Law and Adjunct Professor of Internal Medicine. 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.

""
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.
Riehs, rita t. (1986), Director, S.J. Quinney Law Library and Professor of Law. B.A., 1972, Univ. of Minnesota; J.D., 1975, Univ. of Idaho; M.L.L., 1978, Univ. of Washington.
Rosky, clifford j. (2008), Associate Professor of Law. B.A., 1996, Amherst College; J.D., 2001, Yale Univ.
TETER, MICHAEL (2011), Associate Professor of Law. B.A., 1999, Pomona College; J.D., 2002, 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.
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.
# UTAH LAW REVIEW

## Volume 2012  Number 1

### Board of Editors 2011–2012

**Editor-in-Chief**  
JAVIER M. MIXCO

**Managing Editor**  
J. MASON KJAR

**Executive Footnote Editor**  
THOMAS GOODWIN

**Note & Comment Editors**  
KIMBERLY CHILD  
MATT DODD  
BRADFORD LIDDELL  
SARA MONTOYA

**Senior Staff**  
JORDAN TOONE

**Executive Articles Editor**  
ERIKA SKOGARD

**Symposium Editor**  
MINDI HANSEN

**Associate Editors**  
REBECCA HELD  
ALAN HOUSTON  
MEGHAN SHERIDAN  
KASSIDY WALLIN

**Staff Members**  
JARED FORSGREN

**Articles Editors**  
MAUREEN TOOMEY ARMOS  
RACHAEL BOETTCHE  
LAUREN HOSLER  
MATTHEW KAUFMANN  
ADAM PACE  
CHRISTOPHER PETERSEN  
KRISTIN SCHERRRUBLE  
MARC THERRIEN

**Editor-in-Chief**  
JAVIER M. MIXCO

**Executive Articles Editor**  
ERIKA SKOGARD

**Symposium Editor**  
MINDI HANSEN

**Associate Editors**  
REBECCA HELD  
ALAN HOUSTON  
MEGHAN SHERIDAN  
KASSIDY WALLIN

**Staff Members**  
JARED FORSGREN

**Executive Online Editor**  
BARRY STRATFORD

**Articles Editors**  
MAUREEN TOOMEY ARMOS  
RACHAEL BOETTCHE  
LAUREN HOSLER  
MATTHEW KAUFMANN  
ADAM PACE  
CHRISTOPHER PETERSEN  
KRISTIN SCHERRRUBLE  
MARC THERRIEN

**Editorial Assistant**  
ANGELA TURNBOW

### Advisory Committee

**Advisory Committee**  
PATRICIA M. LEITH  
CLAYTON J. PARR  
AMY J. WILDERMUTH  
MICHAEL D. ZIMMERMAN

**Editorial Assistant**  
ANGELA TURNBOW
# TABLE OF CONTENTS

## ARTICLES

<table>
<thead>
<tr>
<th>Title</th>
<th>Author(s)</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Law, Fact, and Discretion in the Federal Courts: An Empirical Study</td>
<td>Robert Anderson IV</td>
<td>1</td>
</tr>
<tr>
<td>Bargained Justice: Plea-Bargaining’s Innocence Problem and the <em>Brady</em> Safety-Valve</td>
<td>Lucian E. Dervan</td>
<td>51</td>
</tr>
<tr>
<td>Recantations Reconsidered: A New Framework for Righting Wrongful Convictions</td>
<td>Adam Heder &amp; Michael Goldsmith</td>
<td>99</td>
</tr>
<tr>
<td>Remedying the Misuse of Nature</td>
<td>Sanne H. Knudsen</td>
<td>141</td>
</tr>
<tr>
<td>Due Process Denied: The Forgotten Constitutional Limits on Choice of Law in the Enforcement of Employee Covenants Not to Compete</td>
<td>David A. Linehan</td>
<td>209</td>
</tr>
<tr>
<td>The Penguin and the Cartel: Rethinking Antitrust and Innovation Policy for the Age of Commercial Open Source</td>
<td>Stephen M. Maurer</td>
<td>269</td>
</tr>
<tr>
<td>Class Action Defendants’ New Lochnerism</td>
<td>Mark Moller</td>
<td>319</td>
</tr>
<tr>
<td>A Minimalist Approach to Same-Sex Divorce: Respecting States that Permit Same-Sex Marriages and States that Refuse to Recognize Them</td>
<td>Robert E. Rains</td>
<td>393</td>
</tr>
<tr>
<td>The Common Law of Disability Discrimination</td>
<td>Mark C. Weber</td>
<td>429</td>
</tr>
</tbody>
</table>
NOTES

E-Verify During a Period of Economic Recovery
and High Unemployment

Emily Patten

475

Incentive for Innovation or Invitation to Inhumanity?:
A Human Rights Analysis of Gene Patenting
and the Case of Myriad Genetics

Laurie E. Abbott

497
LAW, FACT, AND DISCRETION IN THE FEDERAL COURTS: AN EMPIRICAL STUDY

Robert Anderson IV*

INTRODUCTION

The standards of review that appellate courts use to review trial court decisions occupy an enigmatic position in legal practice and scholarly commentary. The conventional wisdom among appellate advocates and judges is that standards of review are key determinants—perhaps the key determinant—in the success of an appeal. Indeed, one prominent federal appellate judge wrote that the standard of review “more often than not determines the outcome” of an appeal and another described the standard of review as “everything.” Flowing against this mainstream view, however, has long been an undercurrent of academic skepticism questioning whether standards of review actually constrain appellate judges. Even among appellate advocates, standards of review are sometimes viewed as “mere legalese” or “meaningless post hoc rationalizations” of appellate judicial decisions. Thus, the divergent streams of thinking mean that even though the

* © 2012 Robert Anderson IV. Associate Professor of Law, Pepperdine University School of Law. J.D., NYU (2000); Ph.D., Stanford (2008). I would like to thank Emily Graves and Ryan Griffee for excellent research assistance.

1 The term “standards of review” is regularly used to refer to at least four different types of review: (1) appellate review of trial court decisions; (2) appellate review of jury decisions; (3) judicial review of administrative agency action; and (4) judicial review of government action under the Constitution (e.g., “strict scrutiny” or “rational basis” review). See 1 STEVEN A. CHILDRESS & MARTHA S. DAVIS, FEDERAL STANDARDS OF REVIEW § 1.03 (4th ed. 2010). This Article deals primarily with the first category—appellate review of trial court decision-making—or what the Supreme Court sometimes calls “court/court review.” Dickinson v. Zurko, 527 U.S. 150, 153 (1999).

2 See, e.g., HARRY T. EDWARDS & LINDA A. ELLIOTT, FEDERAL COURTS STANDARDS OF REVIEW, at vii (2007) (stating that standards of review “are critically important in determining the parameters of appellate review”).


4 EDWARDS & ELLIOTT, supra note 2, at vii (quoting Judge Deanell R. Tacha of the Tenth Circuit Court of Appeals).

5 John Stick, Can Nihilism Be Pragmatic?, 100 HARV. L. REV. 332, 362 n.130 (1986) (identifying standards of review as “notorious for being open to judicial manipulation” along with standing and procedural due process).

6 EDWARDS & ELLIOTT, supra note 2, at vii (“[W]e have been surprised by the number of attorneys who treat the standards of review as mere ‘legalese.’”).

7 CHILDRESS & DAVIS, supra note 1, at 1–14 (explaining it is “tempting” to view standards of review in this way, but arguing that they do “appear to have real value”).
The standard of review is the single most commonly invoked legal doctrine in appellate opinions, it is also probably among the least understood.

The doctrinal perspective on standards of review is relatively easy to describe, at least in its broad outlines. The federal courts are divided into trial and appellate levels, each with a differentiated role in the adjudicatory process, and the standard of review largely delineates, at least in theory, the line between those roles. The standard of review divides judicial decisions into three principal categories—fact, law, and discretion—each allocated to the trial or appellate level. Trial courts have the primary responsibility for fact-finding and discretionary rulings on evidence and procedure. As a result, factual determinations and discretionary rulings are reviewed deferentially by the appellate courts and rarely reversed. The appellate courts, in contrast, have primary responsibility for interpretation and development of the law. As a result, legal conclusions of the trial court are freely (nondeferentially) reviewed by the appellate courts, and are more frequently reversed. In each case, standards of review delineate the boundaries between trial and appellate roles, determining the degree of deference appellate courts are required to give to trial courts. This textbook description of the adjudicatory process might be called the “legal model” of judicial decision-making.

---

8 The standard of review is also typically among the first issues discussed in an appellate opinion. See Wald, supra note 3, at 1391 (“Courts talk about the standard of review up front in most opinions . . . .”).

9 Although the discussion in this Article is oriented primarily towards the federal courts of the United States, the division of roles between trial and appellate courts appears in most judicial systems around the world, although often with different divisions of roles. See Henry J. Abraham, The Judicial Process 1 (1998).

10 Martin B. Louis, Allocating Adjudicative Decision Making Authority Between the Trial and Appellate Levels: A Unified View of the Scope of Review, the Judge/Jury Question, and Procedural Discretion, 64 N.C. L. Rev. 993, 997 (1986) (arguing that the standard of review is “the principal means by which adjudicative decisional power and responsibility are divided between the trial and appellate levels”).


12 See Edward H. Cooper, Civil Rule 52(a): Rationing and Rationalizing the Resources of Appellate Review, 63 Notre Dame L. Rev. 645, 657 (1988) (“[T]rial courts are primarily responsible for sifting the evidence and finding the facts . . . .”).

13 See infra Part I.A.

14 See infra notes 50–51 and accompanying text.

15 See Cooper, supra note 12, at 657 (“[A]ppellate courts are primarily responsible for developing the law.”).

16 See infra notes 42–46 and accompanying text.

17 Of course, there is no single “legal model” of appellate decision-making. The term is used primarily by attitudinal scholars to refer to judicial decision-making constrained by legal principles. Jeffrey A. Segal & Harold J. Spaeth, The Supreme Court and the Attitudinal Model Revisited 45–53 (2002).
There are reasons to doubt, however, whether the legal model’s prescriptive ideal of trial and appellate roles matches the descriptive reality of appellate adjudication. The linchpin that holds the legal model together is the idea of deference—that an appellate court will voluntarily “set[] aside its own judgment” in favor of the judgment of the trial court on factual and discretionary issues. In other words, the idea of deference requires that an appellate court affirm at least some trial court decisions even when the appellate might have made a different decision had it originally decided the matter itself. This assumption becomes problematic as a practical matter, however, when the ideology of the trial court and the appellate court diverge. When the appellate court disagrees with a finding of fact or discretionary ruling made by a trial court it might be tempted to accord less deference to that decision than the legal model would prescribe. And even if the appellate court does defer to the trial court, that very deference might perversely create an incentive for the trial court to manipulate its own decision-making process, favoring rulings based on factual or discretionary grounds to insulate itself from reversal. The skeptic would argue, therefore, that the dual temptation to deviate from the norm of deference could undermine the legal model, distorting fact-finding and legal interpretation at both the trial and appellate levels.

The theoretical concern that the formal legal model of law, fact, and discretion might not adequately capture the reality of the adjudicative process has been an important topic in legal theory for many decades. At least as far back as the legal realism movement, scholars expressed skepticism about the constraining force of legal doctrine and the accuracy of factual determinations. In 1930, Leon Green suggested that the fuzzy line between law and fact could invite manipulation by trial courts. In 1949, Jerome Frank outlined the problematic interaction of facts and legal rules in a perspective known as “fact skepticism,” arguing that “trial-court fact-finding is the soft spot in the administration of justice.” Green and Frank both noted that findings of fact, like conclusions of law, are malleable and subjective at the trial court level. At the appellate level, however, findings of fact are theoretically “fixed” and insulated from review. The appellate courts, for their part, were seen as manipulating the very definitions of law, fact, and discretion to recapture control over factual or discretionary questions insulated

---

19 *Id.* at 1074 (describing a “provisional definition of deference” as one decision-maker’s willingness to follow another decision-maker’s “determination, despite the fact that it might have reached a different conclusion had it reasoned independently”).
20 Legal realism was an early twentieth century legal philosophical movement that, in broad terms, explored the extent to which legal reasoning actually mattered in the arguments that lawyers made and the final decisions of judges. Frederick Schauer, *Thinking Like a Lawyer: A New Introduction to Legal Reasoning* 124–47 (2009).
21 See *infra* note 71 and accompanying text.
22 See, e.g., Jerome Frank, *Courts on Trial* 74 (1949).
Thus, for decades, scholars have argued that the malleability of facts at trial combined with their rigidity on appeal could create an important vulnerability in a hierarchical court system—encouraging manipulation of deferential standards of review by both trial and appellate courts.

The realists and their contemporaries did not identify clearly, however, what empirical implications flowed from this vulnerability, the details of which have been sketched by attitudinal judicial politics scholarship and positive political theory literature in the last decade. In particular, one line of research in this literature has developed a theory that courts and administrative agencies try to further their ideological and policy goals through selecting “decision instruments” that make their decisions more difficult to overturn. The basic argument is that to the extent that reviewing courts or other political bodies find reversing certain types of decisions more costly than others, lower level strategic actors can protect their decisions from reversal by basing decisions on grounds that are more costly to overturn. If some types of decisions by a trial court—for example, factual determinations and legal rulings—are more costly for the appellate court to review, the trial court might have incentives to base decisions on grounds that are insulated from review, such as factual determinations. These contributions from positive political theory build on the skepticism of the legal realists to answer in part the realists’ open question—the extent to which the malleability of law and fact and the strategic use of standards of review influence judicial decision-making.

This Article builds on the foundation of the “decision instrument” literature to develop a theory of strategic interaction based on asymmetric information between the federal trial and appellate courts. This Article’s theory suggests that although clear error and abuse of discretion are both deferential standards of review, they will have very different effects on interaction between trial and appellate courts. Specifically, the theory argues that clear error review, which applies to district court fact-finding, is associated with an informational advantage to the trial court.

---


24 The “attitudinal” model is a political science approach to studying courts that “sees judicial decision-making as determined by the attitudes or preferences of individual judges, whose votes in particular cases reflect their sincere policy preferences largely unconstrained by legal precedent.” Pauline T. Kim, Lower Court Discretion, 82 N.Y.U. L. Rev. 383, 384–86 (2007). Positive political theory approaches are similar in their focus on judges’ policy objectives, but positive political theory approaches “emphasize strategic interactions among judges and between judges and other political actors,” where judges “tak[e] account of the likely response of other actors and the institutional context in which they operate.” Id. at 384–85.


26 See infra notes 42–56 and accompanying text.
that will exacerbate ideological conflict between the trial and appellate courts. In contrast, the theory predicts that abuse of discretion review of procedural and evidentiary rulings is associated with relative symmetry of information that will lead to district court compliance and an absence of overt ideological conflict.

The Article then gathers empirical evidence to examine whether the “legal model,” the “decision instrument” theory, or the “informational” approach developed in this Article best explains the role of ideology in standards of review. The empirical analysis section uses newly developed data to peer inside the “black box” of judicial fact-finding and discretionary decisions using an analysis of appellate outcomes. The analysis extends the decision instrument literature by examining how two different deferential standards of review—clear error and abuse of discretion—affect the probability of reversal of district court decisions, and how those deferential review standards differ from one another and from de novo review. The answers to these basic questions lay the necessary foundation for understanding how deferential review might be used strategically. The analysis then considers how conflicting ideology of the district and appellate courts interacts with standards of review, specifically examining whether judges strategically invoke deferential review. Finally, the empirical analysis tests whether and how the effect of ideological divergence between the courts varies under conditions of informational asymmetries between the trial and appellate courts.

The results support the conclusion that standards of review matter, but not in the way that either the conventional legal model of deference or the positive political theory would predict. The effect of deferential review appears to be more complex, in part because, as the theory developed in this Article suggests, the two principal deferential review standards (clear error and abuse of discretion) have qualitatively different effects. The results are consistent with the legal model in that deferential standards of review appear to considerably decrease the probability of outright reversal, and the analysis reveals no evidence that judges manipulate standards of review. On the other hand, the results suggest the counterintuitive conclusion, predicted by the informational theory, that findings of fact are associated with more manifested ideological disagreement than discretionary rulings or conclusions of law. The Article concludes that trial judges rely on asymmetric information inherent in fact-finding to achieve their goals when the clear error standard already applies, rather than invoking that standard strategically to insulate their decisions from review. Appellate judges, in turn, use ideological cues to discipline these attempts to capitalize on the information

---

27 See infra notes 98–99 and accompanying text.
28 See infra notes 203–204 and accompanying text.
29 See infra Part III.B.
30 See infra Part III.C.
31 See infra Part IV.
32 See infra Part III.C.
33 See infra Part IV.F.1.
asymmetry, making fact-finding more ideologically contentious than legal conclusions.34

In Part I, the Article briefly describes the doctrinal framework of standards of review and sketches a theory of how law, fact, and discretion interact in a judicial hierarchy. The theory suggests three fundamental, but unanswered research questions for empirical testing: (1) whether standards of review affect the probability of reversal of the district court, (2) whether judges select ideologically among deferential and nondeferential standards of review, and (3) how the effect of ideology varies among the standards of review. Part II describes the data and measurement strategy used in this Article, including original data on standards of review developed for this analysis. Part III presents the results of the empirical analysis using the new dataset of standards of review to examine these research questions. The results suggest a complex relationship among standards of review, ideology, and information that does not fully correspond with either the legal model or the positive political theory approaches to the judiciary. Part IV interprets the results and discusses implications for policy decisions about the courts of appeals.

I. DOCTRINE, DEFERENCE AND INFORMATION

A. Legal Doctrine

The organization of the federal judiciary is premised on the division of labor between trial and appellate courts, the boundaries of which are delineated by the standards of review. The standard of review divides trial court decisions into three broad categories, each of which is subject to a different standard: (1) conclusions of law, (2) fact-finding, and (3) discretionary rulings.35 Trial court conclusions of law, or what Hart and Sacks called the “law declaration” function,36 specify what consequences the legal doctrine attaches to various factual situations, including the factual situation as found by the trial court.37 Trial court fact-finding is the process

\[34\] See infra Part IV.F.2.

\[35\] See Pierce v. Underwood, 487 U.S. 552, 559 (1988). See also Professor Rosenberg’s memorable description: “[A]ll appellate Gaul is divided into three parts for review purposes: questions of fact, of law and of discretion.” Maurice Rosenberg, Judicial Discretion of the Trial Court, Viewed From Above, 22 SYRACUSE L. REV. 635, 645–46 (1971). In later work, Professor Rosenberg added mixed questions of law and fact for a total of four categories. Rosenberg, supra note 11, at 32. Edwards and Elliott identify four principal standards of review—de novo, clearly erroneous, abuse of discretion, and plain error. Edwards & Elliott, supra note 2, at 3.


\[37\] The devotee of Hart and Sacks will note the absence of the third category of “law application” where courts apply the law to the facts so as to render a judgment or decision that binds the parties to a particular legal dispute. Id. at 351. This law application function
of determining the historical facts relevant to the case, 38 or what has been called “a case-specific inquiry into what happened here.” 39 Discretionary rulings typically include trial court decisions on detailed questions of trial supervision, procedure, and evidence. 40 Each of these basic components of the adjudicative process—law, fact, and discretion—is subject to its own level of appellate scrutiny according to a “standard of review.” 41

The decisions in the first category—questions of law—are reviewed “de novo” by the appellate court, meaning that the appellate court is not required or expected to give any deference to the trial court. 42 Instead, the appellate court exercises its own judgment on the legal questions presented, exercising a form of review sometimes referred to as “free, independent, or even plenary review.” 43 In such cases, “[t]he appellate courts merely ask themselves whether they agree with the trial judge’s resolution of the legal issue. If not, they reverse him quick as a flash . . . .” 44 The appellate court is entitled to—and should—take into account the legal analysis of the trial court, 45 but the appellate court has the right and even the duty to exercise its own independent judgment. 46

In contrast to legal conclusions, factual determinations by the trial court are reviewed deferentially. 47 Appellate review of factual findings by the district court is limited to whether those findings are “clearly erroneous,” 48 a very high bar to

---

38 This is what Hart and Sacks call the “fact identification” function. HART & SACKS, supra note 36, at 350.
41 CHILDRESS & DAVIS, supra note 1, at 1–3. In addition to the degree of deference, the standard of review can also determine the materials the appellate court will look at in the course of its review, and even how hard the appellate court will look for error. Id.
42 See BLACK’S LAW DICTIONARY 924 (9th ed. 2009).
43 CHILDRESS & DAVIS, supra note 1, at 2–89.
44 Rosenberg, supra note 35, at 646.
46 Id. at 238 (“When de novo review is compelled, no form of appellate deference is acceptable.”).
47 There are exceptions to the deferential review of factual findings, of which the “constitutional fact” one is the most notable. See, e.g., Monaghan, supra note 39, at 231. This doctrine grew out of the “jurisdictional fact” doctrine of earlier cases. See Christie, supra note 37, at 26–31.
48 The standard of review for bench trial findings of fact in civil cases and criminal cases as to nonguilt questions is “clear error.” CHILDRESS & DAVIS, supra note 1, § 7.01, at 7–3. The standard of review for jury findings of fact in both criminal and civil cases are variously described as “substantial evidence” or “reasonableness.” Id. §§ 3.04, 7.01, at 7–3.
reverse a trial judge’s factual findings. The clear error standard does not require reversal only if there is no evidence to support the finding. Instead, as the canonical formulation of the clearly erroneous standard suggests, a finding of fact is clearly erroneous “when although there is evidence to support it, the reviewing court on the entire evidence is left with the definite and firm conviction that a mistake has been committed.” The key point is that the standard of review constrains the appellate court to defer to the trial court’s findings of fact, meaning that the appeal is primarily an appeal of legal issues, not a new trial of the whole case. Indeed, the conventional wisdom on review of findings of fact is very simple: “appellate courts do not engage in factual evaluation,” so that the clear error review is “usually cited to justify refusal to interfere with the fact findings made in the trial court.”

The third and final major standard of review—abuse of discretion—also requires deference to the trial court. The abuse of discretion standard applies to a wide range of procedural and evidentiary decisions that are entrusted to the discretion of the trial judge. Consistent with the broad range of decisions that are considered discretionary, there is no one formulation of the abuse of discretion standard. Instead, a highly contextual inquiry varies depending on the setting. The concept of discretion in this context “implies the power to choose within a range of acceptable options.” The trial judge can be reversed when he or she goes beyond that acceptable range, but the appellate court must defer to a decision within the acceptable range, even if the appellate court would have made a

The standard of review for judicial findings of fact in nonguilt issues is the same. Id. § 7.01, at 7–3.

49 United States v. U.S. Gypsum Co., 333 U.S. 364, 395 (1948). The “substantial evidence” standard applied to jury findings of fact is in theory even more deferential, although in practice many courts tend to equate the two standards. CHILDRESS & DAVIS, supra note 1, § 2.07, at 2–43 to 2–46.


53 For example, rulings about the probative versus the prejudicial effect of evidence under Rule 403 of the Federal Rules of Evidence is subject to abuse of discretion review, see Old Chief v. United States, 519 U.S. 172, 183 n.7 (1997), as is the decision to grant or deny permanent injunctive relief, see Weinberger v. Romero-Barcelo, 456 U.S. 305, 320 (1982), and preliminary injunctions, see Ashcroft v. ACLU, 542 U.S. 656, 664 (2004), and the decision to admit or exclude expert testimony, see Joiner, 522 U.S. at 138–39.

54 See EDWARDS & ELLIOTT, supra note 2, at 67 (“[T]he variety of matters committed to the discretion of district judges means that the standard is necessarily variable. It implies no single level of scrutiny by the appellate courts.”). See also CHILDRESS & DAVIS, supra note 1, § 4.01, at 4–12 (“[T]here is no such thing as one abuse of discretion standard.”).

55 CHILDRESS & DAVIS, supra note 1, § 4.01, at 4–3.
different decision. The precise formulation of the standard will vary from case to case, but in all cases it is “deference that is the hallmark of abuse-of-discretion review.”

B. Hypotheses

The legal doctrine of standards of review is clear: appellate courts are to defer to trial court findings of fact and discretionary rulings, but not to conclusions of law. In light of the skeptical undercurrent about standards of review, however, the question remains as to what causal effect, if any, the doctrinal command of deference actually has on the appellate court’s review. As a preliminary matter, there is little doubt that appellate courts regularly defer to trial court decisions regardless of the standard of review, if for no other reason than the sheer volume of appellate disputes that make plenary review of all decisions impossible. Factual disputes in particular are especially complicated, and therefore would naturally receive some deference, whether or not the standard of review commanded such deference. But the conventional account of appellate process makes a much stronger claim—that appellate judges defer to trial court fact-finding and discretionary rulings not merely out of necessity, but also out of obedience. In other words, the conventional legal model claims it is the formal legal standard of review, not merely constraints on the appellate court’s resources, that causes the appellate judge to review the trial court more deferentially.

The conventional account is vulnerable, however, to at least three types of temptations that could undermine the prescriptive rule of deference. First, appellate judges might invoke deferential standards of review in their opinions and yet freely reverse decisions with which they disagree. Second, even if appellate judges faithfully defer when they apply deferential standards of review, appellate judges might strategically invoke deferential standards of review when they agree with the lower court’s decision, and nondeferential standards when they disagree with the lower court’s decision. Third, even if appellate judges faithfully defer according to the deferential standards and do not strategically invoke them, the very fact of deference may create an incentive for the trial courts to manipulate their decisions toward deferentially reviewed grounds of decision. It is possible that judges do not succumb to the three temptations and do not strategically choose between deferential and nondeferential standards. Indeed, deferential review may be a

---

56 Joiner, 522 U.S. at 143.

57 Indeed, the “legal model” may be inaccurate not only in that judges do not defer when they are required to, as explored in this Article, but in that judges do defer when they are not required to, whether out of sheer necessity or for other reasons. There is some evidence for this behavior, which is referred to as the “affirmance effect.” See Chris Guthrie & Tracey E. George, The Futility of Appeal: Disciplinary Insights into the ’Affirmance Effect’ on the United States Courts of Appeals, 32 Fla. St. U. L. Rev. 357, 358 (2005); see also Kevin M. Clermont & Theodore Eisenberg, Litigation Realities, 88 Cornell L. Rev. 119, 150 (2002) (showing the “affirmance rate to be about 80%”).

58 See Clermont & Eisenberg, supra note 57, at 150–52.
rational organizational institution that leaves room for ideological disagreement in
the judicial hierarchy. The next three sections theorize the temptation possibilities
and take preliminary steps toward understanding which account best describes the
effect of deferential review.

1. **Do Standards of Review Affect Reversal Rates?**

The first research question developed in this Subpart is whether deferential
review is associated with a lower probability of reversal compared to de novo
review. The general understanding of clear error review and abuse of discretion
review is that reversal is very unlikely, tending to shift power from the appellate
court to the trial court.59 Yet the question of what effect, if any, standards of review
have on the outcome of individual cases has not been subject to extensive
empirical analysis. The conventional wisdom must be tested against outcomes,
because even if appellate panels invoke deferential standards of review when legal
doctrine suggests they should, the judges may simply fail to actually defer to the
trial court.60 In other words, the appellate court might recite the standard of review
but not feel meaningful constraint on its decision to affirm or reverse. This is
probably the most fundamental question to ask about standards of review, because
if the standard of review is unrelated to the probability of reversal, then both the
legal model and the “decision instrument” model are based on false assumptions.61

Why would the standard of review be viewed as a constraint the appellate
panel might want to ignore? The most obvious reason would suggest that
ideological disagreement between the trial and appellate courts would provide a
strong temptation to ignore deferential standards of review.62 The first research
question investigated, therefore, is whether standards of review or differences in
ideology between the district court and appellate court better explain appellate
outcomes. This involves comparing outcomes along two dimensions. First, as the

---

59 Louis, *supra* note 10, at 1045 (“Designating a procedural determination as
discretionary results in a limited scope of review on appeal, few reversals, a reduced
number of appeals, and, therefore, trial level hegemony over the question.”).

60 Of course, the appellate court might then be subject to reversal by the circuit en
banc or by the Supreme Court, which does occur from time to time. See, e.g., Gall v.
United States, 552 U.S. 38, 56 (2007) (“[A]lthough the Court of Appeals correctly stated
that the appropriate standard of review was abuse of discretion, it engaged in an analysis
that more closely resembled *de novo* review of the facts presented and determined that, in
its view, the degree of variance was not warranted.”).

61 The legal model would be wrong because it commands deference to the trial court,
which should decrease the reversal rate. The strategic model would be wrong because
judges cannot strategically invoke standards of review if they make no difference to the
outcome.

62 See, e.g., Guthrie & George, *supra* note 57, at 364–71 (“The Court is more likely to
grant certiorari to review lower court decisions that are ideologically inconsistent with the
Court’s current majority because the Supreme Court acts ideologically, and thus it is more
likely to grant certiorari to reverse than to affirm the lower court.”).
standard of review changes does the rate of reversal or affirmance change? Second, as the ideology of the median member of the appellate panel diverges from the ideology of the district court judge, how does the probability of reversal change? The first hypothesized factor—the standard of review—although relatively straightforward and intuitive, does not appear to have been tested empirically in the court/court context until very recently. The second factor—the effect of ideology—does not appear to ever have been compared with standards of review.

2. Do Judges Strategically Choose Standards of Review?

The second research question is whether trial courts, appellate courts, or both strategically manipulate deferential review to achieve desired outcomes. If appellate courts do defer when deferential standards apply, as even some skeptical observers have conceded they likely do, trial courts might have an incentive to behave strategically. Specifically, a trial court might attempt to base a vulnerable decision on factual or discretionary grounds when the appellate court is likely to disagree with the trial court’s decision, and on legal grounds when the appellate court is likely to agree with the trial court’s conclusion. Similarly, appellate panels might characterize favored trial court decisions as “factual” or “discretionary,” invoking deferential review, while characterizing disfavored decisions as “legal,” invoking de novo review. The necessary conditions for this type of manipulation are that: (1) deferential review affects the reverse/affirm decision and (2) deferential review is, at least in part, the object of choice by the trial court, the

63 See Corey Rayburn Yung, Judged by the Company You Keep: An Empirical Study of the Ideologies of Judges on the United States Courts of Appeals, 51 B.C. L. REV. 1133, 1136 (2010); see also Joshua B. Fischman & Max M. Schanzenbach, Do Standards of Review Matter? The Case of Federal Criminal Sentencing, 40 J. LEGAL STUD. 405, 406 (2011) (“Although standards of review are often the subject of intense political and judicial debate, there is little empirical evidence that these standards have an effect on either district or circuit judges.”). Another important study examined the standards of review in the context of review of administrative decisions, but did not test the effect of ideology on court/court review. See Frank B. Cross, Decision Making in the U.S. Court of Appeals 53 (2007).

64 See, e.g., Yung, supra note 63, at 1136 (“Only recently have studies begun to develop comprehensive measures of ideology based upon actual judicial performance.”).

65 Wright, supra note 23, at 770 (“The courts which have disregarded Rule 52 in substituting their judgment for that of the trial court could accomplish the same purpose while complying with the rule merely by announcing that the finding with which they disagree is ‘clearly erroneous.’ But I think we can safely assume that appellate judges do make a conscientious attempt to confine their review to that authorized by law, and that, so far as human frailties permit, they do not regard a finding as clearly erroneous merely because it differs from the finding they might themselves have made.”).

66 See, e.g., Henry J. Friendly, Indiscretion about Discretion, 31 EMORY L.J. 747, 776 (1982) (“A favorite method used by appellate courts to avoid the discretion rule in temporary injunction cases is to find that the district court proceeded on an erroneous view of the law . . . .”).
appellate court, or both. The first condition—that standards of review matter—was described above. The second condition—the manipulability of deferential review—is described below.

Do trial judges, appellate judges, or both have the flexibility to move freely between deferential review and nondeferential review, as some of the strategic instrument literature would suggest? The argument draws considerable support from the vast literature on the elusiveness of a distinction between law and fact. Indeed, the idea that judges think strategically about law and fact has been a part of the legal literature at least since the legal realist movement, and is still vibrant today. In a quotation from over 80 years ago, the legal realist Leon Green incisively remarked:

No two terms of legal science have rendered better service than “law” and “fact.” They are basic assumptions; irreducible minimums and the most comprehensive maximums at the same instant. They readily accommodate themselves to any meaning we desire to give them. In them and their kind a science of law finds its strength and durability. They are the creations of centuries. What judge has not found refuge in them? The man who could succeed in defining them would be a public enemy.

The basic idea of the quotation is clear; the distinction between the terms “law” and “fact” is malleable at best, and the malleability is not entirely unwelcome to the judges who wield the terms. But Green goes farther, arguing that judicial actors—primarily trial judges in this context—can strategically insulate...
their decisions from scrutiny and “take refuge” by choosing to ground the decision in law or in fact.  

The malleability of the law-fact distinction works both ways, however, creating strategic opportunities for both the reviewer and the reviewed. Appellate judges might tend to characterize decisions they wish to reverse as conclusions of law, and decisions they wish to affirm as findings of fact, selectively using de novo and deferential review standards as “ideological weapon[s].” After all, as Judge Wald has pointed out, appellate courts do not simply “apply ‘the law’ to the facts,” which would assume a clean distinction between the two, but they also “have to decide what the ‘standard of review’ is,” meaning that the distinction is endogenous to the choice process. The “mixed question of law and fact” is the most obvious context in which the appellate court has an explicit opportunity to choose the standard of review. This is where many scholars observe the “struggle for power” between trial and appellate courts. Over the years, authors have argued that appellate courts have in fact manipulated the standard of review, often by moving more and more decisions from deferential review into the de novo category.

The standard of review, therefore, is potentially a strategic choice for the appellate court and there is the possibility that courts manipulate standards of review just as Green’s observed decades ago. This perspective has been invigorated in recent decades by positive political theory accounts of judicial decision-making. Here, the emphasis is typically on how the lower court or an administrative agency can strategically choose grounds for the decision that makes review by the higher-level actor more costly. This might include, for example, the choice of an agency to make policy by rulemaking or adjudication, or an appellate court’s choice between critiquing an agency’s statutory interpretation and critiquing an agency’s reasoning process. In each case, the first choice (rulemaking and statutory interpretation) involves legal questions that can be

72 Indeed, trial judges can even manipulate the type of facts they rely on, shifting the emphasis from “ultimate facts” to “historical facts.” See Louis, supra note 10, at 1015 (“Trial judges can still slant or stack their historical fact-findings to shield their ultimate fact-findings from reversal.”).

73 See Adamson, supra note 70, at 1028–32.

74 Wald, supra note 3, at 1391.

75 Louis, supra note 10, at 1003 (“[T]heir inevitable struggle for power must center on the process of law application and in particular on the scope of review of mixed law/fact questions.”).

76 See, e.g., Martin Shapiro, Appeal, 14 LAW & SOC’Y REV. 629, 648 (1980) (“Appellate courts have worked so hard at eroding the boundary between facts and law that today almost any issue can be characterized as a question of law or a mixed question of law and fact, and in either case appropriate for appellate consideration.”).

77 See infra Part IV.

78 See infra Part IV.

79 Tiller & Spiller, supra note 25, at 349–51.

reviewed at low cost and the second choice (adjudication and reasoning process) involves fact-intensive inquiries that can be reviewed only at high cost. A few empirical studies have investigated these theories and whether their model predictions are supported.81

The work described above has dealt primarily with judicial-administrative interaction, but a more recent article in this line of research directly addresses the court/court strategic issues that this Article attempts to untangle.82 One recent piece, for example, specifically examines the district court’s attempt to strategically characterize decisions as findings of fact or conclusions of law.83 It tests the hypothesis that district court judges will be more likely to ground their sentencing decisions as “departures” (law-based) rather than “adjustments” (fact-based) when there is alignment with the reviewing circuit court.84 In contrast, the theory argues that lower courts will be limited to the fact-based adjustments when there is ideological divergence between the sentencing court and the appellate court, because fact-based adjustments receive deferential review while law-based departures do not.85 If this is true and appellate judges routinely choose the standard of review based on their attitude toward the case before them, this might bias the results in the first two hypotheses.

To test whether deferential review is strategically invoked, this Article examines the relationship between ideology and standards of review at the district court and appellate court level. The first part of the hypothesis is that appellate courts would apply more deferential standards in cases where they are ideologically close to the lower court, and more searching standards where they are ideologically farther from the lower courts. The second part of the hypothesis is that district courts will “take refuge” in deferential review standards when the district court is ideologically farther from the center of the relevant appellate circuit. If indeed the distinction between law and fact or between law and discretion is elusive, traditional legal theory does not offer a solution for dealing with this strategic behavior.

3. Do Judges Strategically Choose Within Standards of Review?

In addition to the possibility that judges might ignore deferential review or strategically select among standards of review, judges might also behave ideologically within standards of review. In other words, judges might behave more ideologically under deferential review than under de novo review or vice versa. There are at least two types of reasons why judicial behavior might vary

81 See, e.g., id. at 64–67 (applying empirical methodologies to a study of judicial review of administrative action).
83 Id.
84 Id. at 25–26.
85 See id. at 33.
under different standards of review. The first type of reason is that the command of deferential or nondeferential review itself might affect ideological behavior. One might hypothesize, for example, that appellate judges behave less ideologically when they are instructed to review the trial court deferentially. The second type of reason is that the context in which the relevant standard of review is applied might affect ideological behavior. One might hypothesize, for example, that clear error or abuse of discretion review might prove less ideological not because of the command of deference, but because of the narrow, limited nature of rulings on these issues that primarily affect the case before the court. To understand how the command and the context might affect decision-making within each standard of review, it is valuable to examine the purposes and applicability of deferential review.

The basic test for whether deferential review is appropriate is remarkably functional—focusing on two questions: (1) whether the district court is “better positioned” to decide the issue and (2) whether “probing appellate scrutiny” will “contribute to the clarity of legal doctrine.” These two functional considerations flow from the two primary purposes of the appellate courts, error correction and development of a consistent, coherent body of law. The two inquiries are sometimes described collectively in terms of whether the issue is “fact-intensive,” “fact-dependent” or “fact-bound,” but the inquiries are analytically separate, as the Supreme Court itself has recognized.

The first type of situation where deference is appropriate is where the facts that are relevant to the decision are more accessible to the trial judge than to the appellate panel. This is the rationale for the deferential “clearly erroneous”
standard on review of facts themselves, but also applies to mixed questions of law and fact where “the district court may have insights not conveyed by the record,” meaning that some facts “may be known only to the district court.”\textsuperscript{91} The trial judge spends much more time interacting with the evidence, witnesses, and details of the case, and of course the trial judge alone has access to demeanor evidence, whereas the appellate court only has the “cold printed record.”\textsuperscript{92} Indeed, in a very real sense, the trial judge is himself or herself a “witness[ ] to what transpires in the courtroom.”\textsuperscript{93} As a result, facts may be completely lost on appeal,\textsuperscript{94} for as the Supreme Court has stated, “nervousness cannot be shown from a cold transcript.”\textsuperscript{95} In these situations, the decision is “fact-bound” in a sense, but the reason for deference is really that the district court has information the appellate court cannot acquire without a tremendous investment of resources, if at all.

The second inquiry, which examines whether de novo appellate examination will contribute to legal doctrine, is the one that focuses on whether the issue is fact-sensitive, fact-intensive, or fact-dependent. The rule is highly dependent on the precise facts of the case, whether or not those facts are peculiarly within the special command of the trial judge. The question is one of judicial resources, but here the appellate court is deciding whether formulating detailed rules on a particular issue exceeds the benefit from those rules.\textsuperscript{96} In such cases, deference is indicated when the issue is not appropriate for appellate decision because “impracticability of formulating a rule of decision” when the issue involves “multifarious, fleeting, special, narrow facts that utterly resist generalization . . . .”\textsuperscript{97} The appellate court may well have access to the same information as the trial court at relatively low

\textsuperscript{91} Pierce, 487 U.S. at 560.

\textsuperscript{92} The earliest example of this heavily used term that I have been able to locate is the concurring opinion of Judge McFarland in \textit{People v. Phelan}, 56 P. 424, 431–32 (1899) (“I desire to say, however, that if the testimony were to be considered simply as it stands in the cold, printed record, without any reference to the appearance, manner of testifying, etc., of the witnesses, I would not feel sure in holding that it warranted a conviction. But the credibility of the witnesses rests with the jury, and, as it depends upon many things which cannot be reproduced here, I do not feel that this court would be justified in setting aside the verdict on the ground of want of evidence.”).

\textsuperscript{93} Christie, \textit{supra} note 37, at 46 (citing \textit{JEROME FRANK, LAW AND THE MODERN MIND} 109–10 (1930)).

\textsuperscript{94} Id. at 50 (“No record on appeal, however complete, can capture the richness of the trial situation.”).

\textsuperscript{95} Snyder v. Louisiana, 552 U.S. 472, 479 (2008) (quoting Louisiana v. Snyder, 942 So. 2d 484, 496 (2006)).


cost, but no general rule could be formulated that would helpfully resolve future cases.

The first type of argument for deferential review involves the high cost of the appellate court becoming informed of the facts, requiring a mechanism for the trial court to develop and communicate factual information to the appellate court. The paradigmatic example of this situation is the clear error review of judicial findings of fact. The second type of argument involves the high cost of communicating detailed rules from the appellate court to the trial court in fact-dependent contexts. The paradigmatic example of this situation is the abuse of discretion review of trial court rulings, such as evidence and procedure. In both cases, deference to the trial court is appropriate, but only in the first case does the trial court have private information.

The distinct nature of these two reasons for deferential review leads to a third hypothesis about standards of review. If the district court has private information about the case it can behave in an ideological manner without perfect monitoring by the appellate court. That is, the appellate court cannot completely discipline ideological behavior by the trial court, because even if the case is appealed, the appellate court cannot invest the resources necessary to critically evaluate findings of fact. Because the appellate court cannot directly observe the underlying facts of the case, the appellate court may feel compelled to rely on signals, such as the trial court’s ideology, to infer the “true” facts. The appellate court reviewing a decision of a district court judge known to have a different ideological viewpoint might invest more resources in reviewing the facts underlying the decision.

For abuse of discretion review of evidentiary and procedural decisions, however, the trial court is not likely to have a considerable informational advantage relative to the appellate court. The appellate court can generally inform itself of all relevant information about the trial court’s procedural and evidentiary rulings at relatively low cost. The district court knows this, which means that the district court would rationally hew much closer to the appellate panel’s likely preferences (i.e., the ideological makeup of the whole circuit from which panels are drawn). In other words, the trial court knows that it has an informational advantage vis a vis the appellate court in findings of fact, but does not have a considerable advantage in discretionary rulings. If this theory is true, then judges

---

98 Louis, supra note 10, at 1040 (“[The] appellate court is also likely to be more expert and reliable in matters of procedure than is a single trial judge or agency whose raison d’être is the acquisition of substantive expertise. The number of pages in the record relevant to procedural rulings also should ordinarily be fewer than those relevant to determinations going to the merits. Hence, the time required for free review of procedural questions will not ordinarily be as great as for substantive ones.”).

99 These are generalizations, of course, that do not hold in every case. Some clear error contexts involve less private information, as when the factual findings depend primarily on documentary evidence, and some involve more private information. These differences vary the level of deference that is appropriate. See, e.g., Cooper, supra note 12, at 653–54 (oral testimony, credibility issues, and longer, more complex trials may call for more deference, even within clear error review, so the standard is variable).
should behave differently under the two deferential standards of review, with ideological differences most pronounced in appellate review of fact-finding decisions of the trial court.

The difference between the clear error review of findings of fact and the abuse of discretion review of procedural and evidentiary decisions therefore produces a third research question: do appellate court decisions that involve deferential review but no asymmetric information or that involve asymmetric information but no deferential review show less ideological effect than those that involve both? This hypothesis involves comparing deferential standards of review that involve little private information (review under the abuse of discretion standard) with standards of review that involve both deferential review and substantial amounts of private information (review of fact-finding for clear error).

C. Interaction Between Hypotheses

These three research questions are not separate, discrete predictions about the judicial system, but interrelated propositions that address the role of appellate review in the context of ideology and asymmetric information. The first hypothesis tests whether deferential review increases the probability of affirmance. The second hypothesis tests whether an affirmative question to the first hypothesis induces judges to choose the standards of review based on ideological divergence between the trial and appellate courts. The third hypothesis tests whether deferential standards of review themselves are associated with ideological disagreement because of asymmetric information between the trial courts and the appellate courts. The three research questions together are necessary for a complete understanding of how deferential review works in a hierarchical court system with divergent ideological preferences.

II. THE DATA

Testing the hypotheses developed above required collecting data on how findings of fact and discretionary decisions are treated on appeal and the extent to which ideological divergence between trial courts and appellate courts influences that treatment. The federal judicial system provides a happy coincidence for conducting this type of empirical research on appeals. The three-judge panels in the court of appeals are assigned randomly to cases, providing a “natural experiment” for assessing the influence of ideology on various empirical questions.100 Thus, there is little risk that there will be a significant selection bias with respect to which judges within a circuit hear a particular case. To the extent ideology appears to predict the probability of reversing lower court decisions under the various standards of review, we may have greater confidence in that result than in the results of most observational studies. Investigating the three research

---

questions outlined above requires data on the standards of review applied in appellate cases. This analysis starts in the same place as many other analyses of appellate decision-making—the most well-known and comprehensive available database on the United States Court of Appeals developed by Donald Songer and his colleagues.

The Songer database consists of a random sample of United States Court of Appeals cases between the years 1925 and 1996. This analysis incorporates all cases in the Court of Appeals database, and the update, to produce a database covering the years 1985 through 2000. These years were chosen to provide a relatively balanced mix of Democratic and Republican appointees to the bench. The data for years 1985 to 1996 were taken from the original Songer database, and the data for years 1997 through 2000 were taken from the update.

The Court of Appeals database contains many variables that are useful to this empirical study but does not contain data on standards of review. To address this problem, original data was collected for this Article about the most common standards of review applied to federal district courts—de novo, clear error, and abuse of discretion. For each case, the data provide information on (1) whether the appellate court invoked one or both of the deferential standards of review—“clear error” or “abuse of discretion,” and (2) whether the appellate court found the relevant standard for reversal was met (that is, the district court committed clear error), was not met (the district court did not commit clear error) or did not clearly state whether the relevant standard for reversal was met.

The theory is that the “clear error” standard will serve as a proxy for judicial fact-finding and therefore asymmetric information of the trial court. The coding strategy makes use of the fact that federal appellate courts generally review trial court findings of fact under the “clearly erroneous” standard. This standard applies in civil cases because of Federal Rule of Civil Procedure 52(a), and in many criminal cases because of statutory enactments and case law. Thus, we can use the

---

101 See supra Part I.B.
103 See id. The cases are actually from a stratified sample by circuit rather than simple random samples.
105 Using original data collected by the author, the coding was conducted by identifying each instance in which the terms “clear error,” “clearly erroed,” “clearly erroneous,” or similar unambiguous references to the clearly erroneous standard occurred. The paragraph was then read and coded as finding clear error or finding no clear error only if the court’s conclusion was unambiguous. If an appellate court reviewed multiple findings and found clear error as to some and no clear error as to others, the case was coded as having found clear error.
appellate court’s invocation of the clearly erroneous standard as an indication of whether the court is reviewing fact-finding of the lower court. This strategy should capture most instances of appellate review of findings of fact, at least when properly reviewed pursuant to a deferential standard of review. This is because parties are required to identify the standard of review in their briefs, and courts usually explicitly address the standard of review in their opinions. Indeed, in at least one case the Supreme Court has criticized an appellate court for not explicitly identifying the clearly erroneous standard when it reviewed findings of fact. Thus, the “clearly erroneous” standard is likely a reliable indicator of whether the appellate court is reviewing the lower court’s factual findings.

In addition to whether standards of review and ideology affect appellate outcomes, one of the key questions in the theory developed in the previous section is how asymmetric information and appellate deference to trial court ideology interact. To disentangle the effects of incomplete information and deference we need a point of comparison with the clear error standard—an example of deferential appellate review where information is relatively symmetrical between the trial and appellate court. Conveniently, the “abuse of discretion” review applied to various evidentiary and procedural rulings fits this description. Abuse of discretion is a deferential standard of review, but one that takes place in the context of relatively symmetric information between the trial and appellate levels. Thus, by comparing cases reviewed under the clearly erroneous standard of review with those reviewed under the abuse of discretion standard of review, we can examine the effect of varying asymmetric information while holding deference relatively constant. A variable was added to determine if the appellate courts’ applied the “abuse of discretion” standard, which is deferential like the “clearly erroneous” standard. Any differences between the ideological patterns in “clear error” review from “abuse of discretion” review would help to disentangle the effects of a deferential standard from those of asymmetric information.

To determine the effect of ideological differences on standards of review, the analysis requires measures of judicial ideology for the district courts, the appellate panels, and the circuits. Although political scientists have attempted to trace
ideological effects in judicial decision-making for many years, the art and science of measuring ideology for judges in the lower federal courts is in its infancy, and to say that such measures are controversial would be an understatement. Because this Article requires a consistent measure that can be applied to district courts and appellate courts, it follows recent work on the lower courts by measuring ideological preferences of judges with Giles-Hettinger-Peppers scores (“GHP scores”). The GHP scores are ideological scores for judges that range from -1 (most liberal) to 1 (most conservative) based on ideological estimates for the political actors involved in selecting the judges—that is, the President and the Senators from the judge’s state. The scores are often referred to as common-space NOMINATE scores because they are designed to put judges, representatives, senators, and the president on a common scale, and are based on NOMINATE scores developed by Poole and Rosenthal.

The scores are developed for the political actors (President and senators) based on roll-call voting in Congress, and then these scores are in turn used as estimates of judicial ideology. To develop a score for a particular judge, GHP uses the score for the appointing President or the Senator(s) with courtesy over the appointment, if any. If senatorial courtesy was not operative in the judge’s appointment, the GHP score is the common space score of the President who appointed the judge. If senatorial courtesy was operative, the GHP score is the

---


114 E.g., CROSS, supra note 63, at 19 (discussing historic method of using “the party of the president who appointed the judge as a guide to the judge’s own ideology”).


116 Id. at 631.


118 POOLE & ROSENTHAL, supra note 117, at 23–24.

119 Giles, Hettinger & Peppers, supra note 115, at 631.

120 “Senatorial courtesy is an unwritten rule of the Senate that gives a senator of the same party as the president the ability to block a nominee to a district court judgeship in the senator’s state.” David S. Law, Appointing Federal Judges: The President, the Senate, and the Prisoner’s Dilemma, 26 CARDOZO L.REV. 479, 493–494 (2005).

121 See id.
common space score of the senator with courtesy (or the mean of the two Senators if both had courtesy). 122

The GHP scores are noisy but allow a continuous measure of ideological preferences that are designed to be comparable over time. In addition, the innovation of GHP over common space scores of the appointing president used in other studies 123 appears to increase the accuracy of the scores, 124 at least with respect to appellate judges. Although another study has suggested that for district judges presidential ideology is dominant over senatorial ideology, 125 a consistent measure of preferences between the two court levels is highly desirable for conducting research on hierarchical interaction. Accordingly, the GHP scores are used throughout this Article.

The measure of ideology for the district court judge is simply his or her GHP score. Following existing literature, the measure of ideology for the appellate panel is the median of the panel’s GHP scores. 126 The key ideological variable throughout most of the analyses is the distance between the ideology of the district court and the ideology of the appellate court (hereinafter “IdeologyDiff”), which is simply the absolute value of the difference of the two scores. 127 Some summary statistics for the judges’ scores in each circuit and the IdeologyDiff score are set forth in Table 1, below. Higher GHP scores represent more conservative ideological values, lower GHP scores represent more liberal values, and larger values of IdeologyDiff represent greater ideological distance between the trial court and the appellate panel.

122 Id.
126 See, e.g., CROSS, supra note 63, at 32–33 (arguing that the panel median ideology best reflects the ideology of the panel from a theoretical and empirical perspective).
127 This approach is substantially the same as that taken in a previous study of reversals using the same dataset (without standards of review). See VIRGINIA A. HETTINGER, STEFANIE A. LINDQUIST, AND WENDY L. MARTINEK, JUDGING ON A COLLEGIATE COURT 91 (2006). The authors in that study used the ideology of the opinion writer rather than of the panel for purposes of consistency with their other results, but noted that the alternative measures made little substantive difference. See id. at 135.
Table 1. Mean GHP Scores.

<table>
<thead>
<tr>
<th>District Court</th>
<th>Appellate Panel Median</th>
<th>IdeologyDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Circuit</td>
<td>-0.072</td>
<td>0.419</td>
</tr>
<tr>
<td>Second Circuit</td>
<td>-0.117</td>
<td>0.253</td>
</tr>
<tr>
<td>Third Circuit</td>
<td>-0.052</td>
<td>0.333</td>
</tr>
<tr>
<td>Fourth Circuit</td>
<td>0.163</td>
<td>0.358</td>
</tr>
<tr>
<td>Fifth Circuit</td>
<td>0.213</td>
<td>0.304</td>
</tr>
<tr>
<td>Sixth Circuit</td>
<td>0.116</td>
<td>0.412</td>
</tr>
<tr>
<td>Seventh Circuit</td>
<td>0.08</td>
<td>0.365</td>
</tr>
<tr>
<td>Eighth Circuit</td>
<td>-0.02</td>
<td>0.358</td>
</tr>
<tr>
<td>Ninth Circuit</td>
<td>0.09</td>
<td>0.402</td>
</tr>
<tr>
<td>Tenth Circuit</td>
<td>0.044</td>
<td>0.392</td>
</tr>
<tr>
<td>Eleventh Circuit</td>
<td>0.143</td>
<td>0.262</td>
</tr>
<tr>
<td>DC Circuit</td>
<td>-0.058</td>
<td>0.527</td>
</tr>
</tbody>
</table>

A quick look at the Table shows the relative dominance of Republican appointees during the period (most of the entries are positive). As would be expected, the appellate panels are generally more moderate and less variable than the district court judges, which results from taking the median of the panel ideologies. The circuits show a fair amount of variability in their ideological tendencies, but the figures roughly parallel expectations.

III. EMPIRICAL EVIDENCE ON STANDARDS OF REVIEW

This Part analyzes the research questions developed above in light of the data described in Part II. The results presented in this Part break down into three subparts—one for each of the hypotheses presented in Part I-B. Part III-A considers the question of whether judges appear to be constrained by the standards of review or whether they appear to choose ideologically despite standards of review. Part III-B considers the question of whether judges ideologically choose among standards of review. Part III-C considers the question of whether judges choose ideologically within standards of review—in other words, does the effect of ideology play out differently between deferential review and de novo review on the one hand, or between conditions of asymmetric information and conditions of symmetric information, on the other.

A. Do Standards of Review Affect Reversal Rates?

The first research question was whether deferential standards of review correlate with decreased reversals and increased affirmances relative to de novo

---

128 See supra Part I.B.
review. The legal model would suggest such a relationship, but skeptics would question the constraining force of the command of deference. Table II below presents a simple cross tabulation of the proportions of reversal and affirmance by the standard of review applied.129

<table>
<thead>
<tr>
<th></th>
<th>Affirmed</th>
<th>Affirmed in Part and Reversed/Vacated in Part</th>
<th>Reversed or Vacated</th>
<th>Number of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clearly Erroneous Only</td>
<td>0.616</td>
<td>0.165</td>
<td>0.220</td>
<td>643</td>
</tr>
<tr>
<td>Abuse of Discretion Only</td>
<td>0.617</td>
<td>0.188</td>
<td>0.195</td>
<td>1027</td>
</tr>
<tr>
<td>Both</td>
<td>0.583</td>
<td>0.274</td>
<td>0.142</td>
<td>372</td>
</tr>
<tr>
<td>Neither (De Novo)</td>
<td>0.584</td>
<td>0.104</td>
<td>0.313</td>
<td>3903</td>
</tr>
</tbody>
</table>

Table II does not show the clear relationship between the affirm/reverse decision expected by the legal model, but instead shows some surprisingly nuanced relationships among the standards of review and the outcome of the appeal. In particular, whether one sees standards of review as mattering depends critically on whether one is presented with the rate of “affirmance” or the rate of “reversal.”130 The rate of affirmance does not vary much across the different standards of review, suggesting that deferential standards of review make little difference in appellate outcomes. The rate varies from about 58 percent in nondeferential review cases to about 62 percent in the deferential review cases, a relatively modest difference, but in the expected direction. However, in cases where both clearly erroneous and abuse of discretion standards apply, the rate of affirmance is actually lower than the cases where neither deferential standard of review applies.

The situation is reversed, however, if one looks at the rate of reversal rather than the rate of affirmance, where standards of review appear to play a significant role. The reversal rate varies from about 31 percent when no deferential standard applies to 22 percent in the clearly erroneous standard cases, 20 percent in the abuse of discretion standard cases, and 14 percent when both deferential standards applied. Table II leaves the puzzling conclusion that both the rate of affirmance and the rate of reversal are minimized when both deferential standards of review apply. These basic figures make it clear that there is something between affirmance and reversal that is closely tied to deferential standards of review.

129 The analysis in this section is similar to the “outcomes analysis” of standards of review that others have performed for administrative decisions. Paul R. Verkuil, An Outcomes Analysis of Scope of Review Standards, 44 WM. & MARY L. REV. 679, 682 (2002); see also CROSS, supra note 63, at 53.

130 For this purpose, cases “vacated” are treated as reversals.
The missing piece in the puzzle is that not all appellate decisions “reverse” or “affirm” trial court decisions in toto; instead, a certain percentage of decisions affirm the trial court in part and reverse or vacate in part. This percentage ranges from a relatively insignificant 10 percent of the cases when no deferential standard applies to a fairly substantial 27 percent of the cases when both deferential standards apply, a significant proportionate increase. The principal effect of standards of review appears to shift cases from the “reversed” category to the “affirm in part and reverse in part” category. This means that any conclusion as to whether deferential standards of review affect the outcomes of appeals depends greatly on whether one counts decisions affirmed in part and reversed in part as “affirmances,” “reversals,”131 or whether one simply omits them altogether. Indeed, omitting these “mixed” outcomes is exactly what studies of the appellate courts tend to do.132 In many cases the omission makes sense, but it seems clear that these mixed outcome decisions have an integral relationship with standards of review. As a result, any discussion of “reversal rates” or “affirmance rates” is potentially misleading without taking into account this effect.

The data therefore appear to support the proposition that deferential standards of review make reversals less likely relative to other outcomes. To the extent that mixed outcomes are interpreted as affirmances rather than reversals, the results support the “legal model” that standards of review reduce the reversal rate. The conclusion does not necessarily mean that the standards of review actually constrain the ideological decision-making of judges nor does it mean necessarily that judges are not actually constrained much more than the results suggest. This is because a variety of selection effects could produce the same results in the absence of constraint, or work to attenuate the results in the presence of a strong constraint. The alternative interpretation is that deferential standards of review are more likely to be mentioned when a case is more complex and the complexity of the case significantly increases the probability of a mixed outcome.133

The affirmance rate depends not only on the appellate court’s decision, but also on the strategic anticipation of that decision by trial courts and litigants.134 Civil litigants probably do not flip coins to decide whether to appeal particular issues, meaning that the cases and issues that are appealed probably depend on

131 Cross coded as “affirmed” the categories of “affirmed,” “petition denied,” or “appeal dismissed” as listed in the Songer database, Songer, supra note 102, and “reversed” as all other categories, Frank B. Cross, Decisionmaking in the U.S. Circuit Courts of Appeals, 91 CALIF. L. REV. 1457, 1500–01 (2003).
132 See Stefanie A. Lindquist, Wendy L. Martinek & Virginia A. Hettinger, Splitting the Difference: Modeling Appellate Court Decisions with Mixed Outcomes, 41 LAW & SOC’Y REV. 429, 430 (2007) (“[M]ost studies of appellate court outcomes routinely exclude from analysis cases where the outcome cannot be clearly classified as in favor of one party or the other or cases that have ‘ambiguous results.’”).
133 Id. at 446.
134 See Schanzenbach & Fischman, supra note 63, at 408–09.
self-selection based on the characteristics of the cases themselves. But criminal appeals provide a useful way to test the influence of selection effects. Criminal defendants in general have strong incentives to appeal, and because of public financing of criminal appeals, indigent criminal defendants in particular have no incentive not to appeal. As a result, criminal defendants tend to appeal using a kitchen sink approach. They are much more likely to appeal every possible issue no matter how small the probability of success. Thus, criminal defendants will often raise clear error on appeal even if there is no evidence of factual error, a relationship that we will use in interpreting the results below.

Second, judges might strategically select standards of review, biasing the results observed in this section. If judges tended to apply deferential standards of review to decisions they agree with and nondeferential standards of review to decisions they disagree with, then deferential standards of review would show a higher affirmance rate even if they did not affect appellate review. This possibility—the second of the three research questions—is taken up in the next section.

B. Do Judges Strategically Choose Among Standards of Review?

The strategic theory of standards of review predicts at least two types of strategic behavior, as described in Part I. First, to the extent that appellate courts are obedient to deferential standards of review, trial courts will potentially have an incentive to use factual, evidentiary, or procedural rulings to insulate their decisions from the increased scrutiny applied under de novo review. Second, appellate courts apply deferential standards of review to decisions they agree with and nondeferential standards to decisions they do not agree with. As Frank Cross pointed out in a similar analysis of standards of review of administrative agencies:

"The judge may have some discretion in choosing a review standard and may select the standard that best suits the judge’s preferred ideological outcome. If so, the association might not demonstrate the constraining power of the review standard but only the judge’s use of a deferential standard when his or her extralegal preferences are for deference."
Either one of these possibilities could mean that the results above do not reflect deference, but rather strategic maneuvering by the trial or appellate courts.

Testing the possibility of strategic choice among standards of review requires an examination of how judicial preferences and standards of review interact. There are two possible selection effects: (1) the trial court may choose to ground its decisions on rulings that receive deferential review, which it should do when it expects an ideologically dissimilar appellate court; and (2) the appellate court may strategically choose standards of review, preferring deferential review when the trial court is ideologically similar and de novo review when the trial court is ideologically dissimilar. The two effects would be difficult to untangle, but the random assignment of panels creates a means of distinguishing strategic trial court action from strategic appellate court action. This is because the district court does not know the identity of the appellate panel when it makes its decision, so it must make any strategic decisions based on the composition of the circuit.\textsuperscript{139} The appellate court, on the other hand, knows the identity of the trial judge when it makes its decision, so it can calculate without guesswork. Thus, to the extent trial judges manipulate standards of review, the manipulation should correlate with the district judge’s distance from the ideological center of the circuit.\textsuperscript{140} To the extent that the appellate court manipulates standards of review, that manipulation should correlate with the district judge’s distance from the ideological center of the appellate panel.\textsuperscript{140} Table III presents ideological distances for each of these standard of review categories below.

<table>
<thead>
<tr>
<th>Table III. Ideological Distances by Standard of Review.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Ideological Distance from Appellate Panel</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Clearly Erroneous Only</td>
</tr>
<tr>
<td>Abuse of Discretion Only</td>
</tr>
<tr>
<td>Both</td>
</tr>
<tr>
<td>Neither (De Novo)</td>
</tr>
</tbody>
</table>

\textsuperscript{139} Guthrie & George, \textit{supra} note 57, at 368–74 (explaining that a strategic district court would take the circuit’s preferences into account in decision-making).

\textsuperscript{140} To the extent that the appellate panel anticipates and responds to the possibility of en banc reversal by the whole circuit, the two effects would be difficult to separate. There is, in fact, evidence that it is more likely that the en banc circuits discipline the panels in the rehearing when the appellate panel reverses the district court. Tracey E. George, \textit{The Dynamics and Determinants of the Decision to Grant En Banc Review}, 74 WASH. L. REV. 213, 267 (1999).

\textsuperscript{141} The circuit median ideology was computed using the median ideology of the judges in each circuit in the cases in the dataset. The circuit median ideology was allowed to adjust over time, fixed within each 100 volumes of the Federal Reporter.
The data do not reveal any indication that the ideological distance between the district court and the appellate court affects the probability of invoking deferential review. The relevant comparison is between the standards of review—that is, comparing within each column.\footnote{The average ideological distance from the district court to the circuit court median is smaller than the average distance to the appellate panels because the circuit median is not as variable as the panel median.} If district judges were acting strategically, one would expect deferential review to correlate with larger distances from the circuit median, but in fact the numbers are nearly identical. If appellate panels were acting strategically, one would expect deferential review to correlate with smaller distances from the panel median, but in fact the opposite is true. The same is true even when the other variable is held constant, suggesting no support for the predicted relationship of ideological distance and the probability of applying deferential standards of review. This result is encouraging because if, indeed, this Table showed a systematic relationship between standard of review and ideology, we might worry that the relationships in the other analyses have been miss-specified or at least plagued by serious selection bias.

C. Ideology within Standards of Review

The results so far suggest some support for the legal model, and no support for the strategic invocation of standards of review. The main research question remains to be tested however—how standards of review and ideology interact in a judicial hierarchy. Specifically, how much do the standards of review—de novo, abuse of discretion, and clear error—explain the outcomes of appeals relative to ideology, and how does the effect of ideology differ within each standard of review?

To investigate these patterns in a more systematic way this Part uses multinomial logistic regression analysis\footnote{In general, regression analysis is a tool that “examines the relationship between a quantitative dependent variable \(Y\) and one or more quantitative independent variables.”\cite{Fox}} to determine how selected variables correlate with the outcomes of appellate cases. The dependent variable\footnote{The dependent variable (also called the “outcome” or “response” variable) is the event one is attempting to explain in a regression analysis.} is the outcome of the appeal, which can take one of three values: (1) affirmed, (2) reversed, or (3) affirmed in part and reversed in part.\footnote{The outcomes are treated as nominal, rather than ordinal, leading to the use of multinomial logistic regression rather than ordered probit. A priori, it would seem natural to think of these categories as “ordered,” with “affirm in part and reverse in part” an\footnote{Multinomial logistic regression is an extension of logistic regression (also known as logit) that is used when the dependent variable has more than two nominal categories.\cite{Long}}} The key independent

\begin{itemize}
  \item \footnote{John Fox, Applied Regression Analysis, Linear Models, and Related Methods 16 (1997). Logistic regression (also called logit) is a specialized version of regression analysis used when the dependent variable is binary. See J. Scott Long, Regression Models for Categorical and Limited Dependent Variables 34–35 (1997). Multinomial logistic regression is an extension of logistic regression (also known as logit) that is used when the dependent variable has more than two nominal categories. Id. at 148–51.} The average ideological distance from the district court to the circuit court median is smaller than the average distance to the appellate panels because the circuit median is not as variable as the panel median.
\end{itemize}
variables\textsuperscript{146} of interest are (1) the ideological distance between the appellate court and the district court, denoted IdeologicalDiff, (2) whether the abuse of discretion standard of review applied, (3) whether clear error applied, and (4) the interaction of IdeologicalDiff with the standard of review. Table III below sets for the results for these key variables of interest.

The interpretation of a multinomial logistic model is similar to that of a regular logistic regression model, except that because there are three categories of outcomes here, there are three columns for each model. The first column under each model shows how the variables affect the probability of an outcome of “Reverse” relative to an outcome of “Affirm” for that model (with no control variables for case type). The second column under each model shows how the variables affect the probability of “Mixed” (for example, affirm in part and reverse in part) versus “Affirm” for that model. The third column for each model shows how the variables affect the probability of “Reverse” versus “Mixed” for that model. The top number in each box is the coefficient for the relevant variable that relates the variable to the outcome for that column, and the bottom number in each box is the p-value for that coefficient. Positive coefficients indicate that the relevant outcome (Reverse or Mixed) becomes more likely as that variable increases and negative coefficients indicate that the relevant outcome (Reverse or Mixed) becomes less likely as that variable increases.

intermediate outcome between “Reverse” and “Affirm” thereby suggesting a natural ordering. Indeed, that is exactly how another analysis of the same data analyzed the outcomes. See HETTINGER, LINDQUIST & MARTINEK, supra note 127, at 97 (“[W]e think of affirming or reversing in full as representing opposite ends of a spectrum, with affirming in part/reversing in part constituting the middle ground between them.”). The data on standards of review presented above, however, show that the “Mixed” outcome of affirm in part and reverse in part is qualitatively different from the two others, not an intermediate category between the two.

\textsuperscript{146} The independent variables (also called “predictor” variables) are the variables used to explain the dependent variable.
Table III shows strong support for some aspects of the legal model of standards of review, but echoes the initial conclusion in Subpart A that the relationships are more complex than the legal model suggests. On the one hand, both deferential standards of review clearly correlate with a decreased probability of an outcome of Reverse relative to Affirm or Mixed, a finding consistent with the legal model. But the standards of review act differently from one another in their effect on the Mixed outcome. The abuse of discretion standard of review is strongly associated with a much higher probability of a Mixed outcome relative to either Affirm or Reverse. The clear error standard, in contrast, is strongly associated with a higher probability of Mixed relative to Reverse, but only with a slightly higher probability of Mixed relative to Affirm. Thus, the clear error standard appears to greatly reduce the probability of Reverse, shifting the likely outcome toward both Affirm and Mixed. The abuse of discretion standard appears to greatly increase the probability of Mixed, shifting the likely outcome not only away from Reverse, but also away from Affirm. Thus, the two deferential standards of review appear to behave differently, suggesting that either the context or the command of deference or both are different between the two.

The next set of variables incorporates the role of ideology by considering how IdeologyDiff relates to the other variables. Recall that IdeologyDiff is the estimated degree of ideological disagreement between the trial and appellate courts in the case. The IdeologyDiff variable standing by itself has very little and unclear impact in Model 1. This means that in cases where no deferential standard of review applies (de novo), as the ideological distance between the trial court and the

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Reverse versus Affirm</th>
<th>Mixed versus Affirm</th>
<th>Reverse versus Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.605*** (&lt;0.001)</td>
<td>-1.576*** (&lt;0.001)</td>
<td>0.971*** (&lt;0.001)</td>
</tr>
<tr>
<td>Clear Error</td>
<td>-0.643*** (&lt;0.001)</td>
<td>0.067 (0.680)</td>
<td>-0.709*** (&lt;0.001)</td>
</tr>
<tr>
<td>Abuse of Discretion</td>
<td>-0.276* (0.046)</td>
<td>0.751*** (&lt;.001)</td>
<td>-1.027*** (&lt;0.001)</td>
</tr>
<tr>
<td>IdeologyDiff</td>
<td>0.005 (0.970)</td>
<td>-0.278 (0.167)</td>
<td>0.283 (0.188)</td>
</tr>
<tr>
<td>IdeologyDiff x Clear Error</td>
<td>0.775* (0.020)</td>
<td>0.989** (0.005)</td>
<td>-0.215 (0.607)</td>
</tr>
<tr>
<td>IdeologyDiff x Abuse of Discretion</td>
<td>-0.723* (0.019)</td>
<td>-0.768* (0.019)</td>
<td>0.045 (0.910)</td>
</tr>
</tbody>
</table>

P-values in parentheses.
*** Denotes p < .001
** Denotes p < .01
* Denotes p < .05
appellate court increases, the probability of the appellate court reversing, vacating, or reversing in part does not change much. In a sense, this finding is surprising given the emphasis of ideology in judicial politics literature, and indeed some studies find that ideology is a strong predictor of appellate court outcomes. At the same time, however, political science scholars have largely failed to find the large ideological effects that are present in the Supreme Court at the appellate court level. Indeed, a recent study using this same data as this Article found no ideological influence over the reverse/affirm decision, a result that ran contrary to what the study’s authors had predicted.

The prior study did not, however, look at ideology in the context of deferential standards of review, and that is where Table III suggests a principal source of the ideological conflict lies. When clear error review applies, differences in ideology between the trial court and the appellate courts makes Reverse and Mixed more likely relative to Affirm. When abuse of discretion applies, differences in ideology between the trial court and the appellate court actually make Affirm more likely relative to Reverse or Mixed. Thus, the ideological difference between the trial court and appellate court only appears to reduce the probability of an Affirm decision when much more when the deferential clear error standard applies than when the de novo standard applies, a highly counterintuitive conclusion.

To dig deeper into this surprising result, Model 2 breaks down the effect of ideology by case type, adding a set of control variables for three broad case types in the database: (1) criminal, (2) civil rights and civil liberties, and (3) other.

---

147 See, e.g., Richard Revesz, Environmental Regulation, Ideology, and the D.C. Circuit, 83 VA. L. REV. 1717, 1717–21, 1735 (1997) (finding ideological voting by party of appellate judges in environmental cases in the District of Columbia Circuit); Cross, supra note 131, at 1462–90, 1515 (finding support for both a “political model” and a “legal model” of decision-making on the Court of Appeals, with the “legal model” providing more explanatory power).

148 Frank Cross has hypothesized that the difference results in part from the fact that the Supreme Court mostly reviews “close cases,” in which the law is relatively evenly balanced on each side of a dispute.” Cross, supra note 63, at 17. He explains that “[t]he circuit courts, by contrast, hear many more cases and cannot pick and choose the cases they decide,” meaning the “much broader set of cases is unlikely to be so evenly balanced on the law, so circuit court judges may not be forced to consider ideology, as Supreme Court justices are.” Id.

149 See Hettinger, Lindquist & Martinek, supra note 127, at 98.

150 The Songer database codes for civil rights cases but does not have a separate single variable for civil liberties cases. See Donald R. Songer, The United States Courts of Appeals Data Base Documentation for Phase 1, JUD. RES. INITIATIVE, 77–78, http://www.cas.sc.edu/poli/juri/cta96_codebook.pdf (last visited Apr. 14, 2012). Civil liberties cases were coded as those falling into the categories of First Amendment, Due Process, and Privacy. See id. at 77–89. Note that civil rights is the dominant type of case in this category, accounting for 935 cases versus 295 for all civil liberties cases combined.
There are two reasons for controlling for these specific areas of law. First, there is
the potential that the area of law, rather than the clear error standard, is driving the
ideological effects found in Table III. If different areas of law apply the standards
of review with different frequencies, and different areas of law have different
levels of ideological disagreement, then ideology may appear to be activated by the
standard of review when really the different areas of law drive the result. Indeed,
existing work on ideology in the appellate courts tends to rely rather extensively on
using only civil liberties cases, something that has been criticized by legal
scholars.152 The reason is probably that civil rights/civil liberties cases tend to
show stronger ideological effects than most other categories with respect to the
case outcome, both in appellate courts and in district courts.153 So it is necessary to
target for civil rights and civil liberties cases to ensure that the effects observed in
Table III do not disappear.

The reasons for controlling for criminal cases are probably even more
compelling than for civil rights and liberties cases. As with civil rights and civil
liberties cases, criminal cases may tend to activate ideological disagreement more
than the average noncriminal case.154 But unlike civil rights and civil liberties
cases, criminal cases tend to be affirmed much more frequently than noncriminal
cases,155 raising the possibility of bias in the effect of standards of review. There
are a variety of reasons for each of these facts, but one of the most important ones
in the current context is that the incentives for appeal are quite different in criminal
cases.156 Criminal defendants are much more likely to appeal than defendants in
other types of cases, and are more likely to appeal than the prosecution in criminal
cases themselves.157 These unique features of criminal cases have the potential to
bias the results if not controlled for in the model.

Therefore, Model 2 includes variables for whether the case was a criminal
case, a civil rights/civil liberties case, and also includes interaction effects of each

151 The “other” category includes the remaining three case types in the database, (1)
labor relations, (2) economic activity and regulation, and (3) miscellaneous. See id. at 85–
87. The second category amounts for the vast majority of cases in the “other” group. Id.
152 Barry Friedman, Taking Law Seriously, 4 PERSP. ON POL. 261, 271 (2006) (noting
the overreliance on civil liberties cases in political science scholarship).
153 C.K. ROWLAND & ROBERT A. CARP, POLITICS AND JUDGMENT IN FEDERAL
DISTRICT COURTS 37 (1996) (explaining that in the district courts, “partisan effects are
concentrated primarily in the realm of civil rights and liberties, and that partisan
polarization in this realm is increasing”).
154 See id. at 39 (noting that district court judges “seem[] to have split along ‘political’
lines more often on criminal justice and on civil rights matters than they did with other
types of cases”).
155 See, e.g., Guthrie & George, supra note 57, at 362–63 (comparing the affirmance
rate of criminal cases to cases overall).
156 See supra notes 134–137 and accompanying text.
157 See John Scalia, Bureau of Justice Statistics, Special Report, Federal Justice
available at http://bjs.ojp.usdoj.gov/content/pub/pdf/fca99.pdf (reporting that in federal
criminal appeals “95% of appeals were filed by the defendant; 5% by the government”).
of these case types with IdeologyDiff. This interaction variable allows us to check whether the ideological effects of standards of review found above are merely the result of different standards of review in different areas of law. These control variables are not extensive, but there are reasons to believe that extensive control variables are not necessary, and additional control variables were tested and did not change the results presented in Table III or IV in any significant respect.

The main reason is that random assignment of appellate judges to panels tends to mitigate many of the selection effects that we would otherwise want to control for. Thus, even with relatively few controls, any ideological effects seem plausible as causal variables.

The principal alternative model included control variables for the circuit in which the appeal was heard. One reason for controlling by circuit is that different circuits have different case law about when to apply the various standards, as well as about what level of deference the standards themselves require. Second, the circuits may differ systematically in terms of the types of cases that reach the courts of appeals, whether because of social factors, see SONGER, supra note 112, at 49–50, or institutional factors, see id. at 53–54. Indeed, differences in ideology among circuits continue even when regional effects are controlled. See id. at 125–28. Third, the circuits vary ideologically from one another, and there are varying mixes of ideological actors within circuits. The results when the circuits were controlled were qualitatively the same as those presented.
Table IV. Results from Multinomial Logit Model 2.

<table>
<thead>
<tr>
<th></th>
<th>Reverse versus Affirm</th>
<th>Mixed versus Affirm</th>
<th>Reverse versus Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>** Intercept</td>
<td>-0.264** (0.002)</td>
<td>-1.369*** (&lt;0.001)</td>
<td>1.105*** (&lt;0.001)</td>
</tr>
<tr>
<td>Clear Error</td>
<td>-0.482** (0.003)</td>
<td>0.269 (0.108)</td>
<td>-0.752*** (&lt;0.001)</td>
</tr>
<tr>
<td>Abuse of Discretion</td>
<td>-0.251 (0.073)</td>
<td>0.786*** (&lt;0.001)</td>
<td>-1.037*** (&lt;0.001)</td>
</tr>
<tr>
<td>IdeologyDiff</td>
<td>-0.452* (0.015)</td>
<td>-0.555* (0.036)</td>
<td>0.104 (0.716)</td>
</tr>
<tr>
<td>IdeologyDiff x Clear Error</td>
<td>0.739* (0.031)</td>
<td>0.867* (0.017)</td>
<td>-0.128 (0.766)</td>
</tr>
<tr>
<td>IdeologyDiff x Abuse of Discretion</td>
<td>-0.720* (0.020)</td>
<td>-0.803* (0.015)</td>
<td>0.083 (0.835)</td>
</tr>
<tr>
<td>CivilRightsAndLiberties</td>
<td>-0.380* (0.008)</td>
<td>0.156 (0.340)</td>
<td>-0.536** (0.004)</td>
</tr>
<tr>
<td>CivilRightsAndLiberties x IdeologyDiff</td>
<td>0.994*** (&lt;0.001)</td>
<td>0.341 (0.379)</td>
<td>0.653 (0.120)</td>
</tr>
<tr>
<td>Criminal</td>
<td>-0.816*** (&lt;0.001)</td>
<td>-0.844*** (&lt;0.001)</td>
<td>0.028 (0.878)</td>
</tr>
<tr>
<td>Criminal x IdeologyDiff</td>
<td>0.653* (0.019)</td>
<td>0.694 (0.057)</td>
<td>-0.041 (0.921)</td>
</tr>
</tbody>
</table>

P values in parentheses.
*** Denotes p < .001
** Denotes p < .01
* Denotes p < .05

The results in Table IV show that the findings from the original are robust against controls for the area of law. Only one coefficient (IdeologyDiff for Reverse versus Affirm) changed its sign between the two models, and the reason can be discerned by looking at Table IV. Table IV disaggregates the effects of ideology by case type and shows that ideology has different effects in different areas of law. The criminal and civil rights/civil liberties cases show a fairly strong ideological effect in the expected direction. In the criminal category, ideological distance between the trial and appellate courts makes Reverse or Mixed much more likely relative to Affirm. In the civil rights/civil liberties category, ideological distance between the trial and appellate courts makes Reverse much more likely relative to Affirm, but has less of an effect on Mixed versus Affirm. The ideological nature of civil rights decisions in particular is one that the literature has repeatedly
demonstrated. The existing literature on criminal cases does not show as strong an ideological effect, but the results here clearly show not only that there is such an ideological effect, but also that this effect is distinct from the effect of clear error.

In contrast, in the “other” cases, ideological distance between the trial and appellate courts is actually associated with a statistically significant increase in the probability of Affirm relative to Reverse and Mixed, counter to what one might expect. Thus, in the more controversial contexts of criminal law and civil rights/civil liberties, ideological distance between the trial and appellate courts produces the expected result—lower probability of Affirm and higher probability of reversing or vacating, at least in part. In contrast, in the “other” cases—primarily economic issues and diversity cases—ideological distance between the trial and appellate courts is actually associated with a higher probability of Affirm. This is why the overall IdeologyDiff effect in Model 1 was close to zero, because the ideological and nonideological case types balance each other out.

The results so far show that the clear error standard of review is associated with more ideological behavior between the trial and appellate courts than abuse of discretion or de novo review. Moreover, given the random assignment to panels, the interpretation of ideology as having a causal effect seems quite plausible. There are not many persuasive alternative explanations for why a president or senator’s common-space ideology score would predict reversals of findings of fact. The results do not necessarily indicate, however, that ideology predicts when the appellate panel will actually reverse the findings of fact. Instead, the results indicate simply that the appellate panel is more likely to reverse the judgment in such cases. It could be, for example, that the appellate panel simply “finds” some other way to reverse the judgment when it dislikes the findings of fact. The context of clear error review may create opportunities for ideological disagreement to translate into reversals on other grounds. Similarly, the abuse of discretion can attenuate ideological disagreement on the reverse-affirm decision without actually implying that cases are affirmed on that ground. The conclusions

---

160 For example, in a prominent recent study by Cass Sunstein and his colleagues all civil rights related categories showed an ideological effect of the party of the appointing president. See ELLMAN, SAWICKI, SCHKADE & SUNSTEIN, supra note 100, at 17–46, 54–57.

161 See, e.g., id. at 47–54 (showing no significant ideological effect in criminal appeals).

162 Although the other study looking at the same data controlled for civil rights/civil liberties claims, see HETTINGER, LINDQUIST & MARTINEK, supra note 127, at 94. The study did not examine the interaction of civil rights/civil liberties cases with ideology, which may explain why they did not find this ideological effect.

163 It is conceivable that within a circuit there is some regional effect that bleeds into both the common-space scores and the propensity to reverse findings of fact. This interpretation, however, would require some explanation why regional animosity would not affect the reverse/affirm decision or the abuse of discretion decision in similar ways, which it does not.
so far have been based on coding appellate outcomes and standards of review without actually engaging with the court’s decisions on those standards of review.

The information in Table V and VI, therefore, provides corroborating evidence for the relationships described above based on the additional original data developed as part of this project. As described in Part II above, each case invoking clear error review was read to determine whether the appellate decision unambiguously stated the lower court fact-finding was clearly erroneous, was not clearly erroneous, or did not unambiguously state one way or the other. 164 Each case invoking abuse of discretion review was read to determine whether the court unambiguously stated the lower court abused its discretion, unambiguously stated the lower court did not abuse its discretion, or did not unambiguously state one way or the other. 165 Table V presents the relationship of ideology to clear error review of findings of fact and Table VI presents the relationship of ideology to abuse of discretion review.

---

164 See supra note 105 and accompanying text.

165 Recall that the cases were coded as ruling a finding was “clearly erroneous” only if the court explicitly said the finding was “clearly erroneous” or another similarly unambiguous statement. The same applied to whether the finding was “not clearly erroneous” and whether or not a ruling was an “abuse of discretion.” See supra notes 105–111 and accompanying text. If the same case found one ruling to be clear error or abuse of discretion and found another ruling to be not clear error or not an abuse of discretion, the case was coded as finding clear error or abuse of discretion, respectively.
The numbers here closely reflect the patterns observed in the multinomial logistic regressions presented above, providing strong corroboration of the results.\(^\text{166}\) The table for clear error is almost the mirror-image reversal of the table for abuse of discretion, just as the results in Tables III and IV suggested. In Table V, the distance between the appeals court and the district court is larger when the appellate court held the trial court’s findings of fact clearly erroneous than when the appellate court held the trial court’s findings of fact are not clearly erroneous. In Table VI, the distance between the appellate court and the district court is larger when the appellate court held that the trial court did not abuse its discretion than when the appellate court held that the trial court did abuse its discretion. Moreover, cases in which the clearly erroneous standard was invoked with no unambiguous ruling do not appear to differ from cases in which the standard was not invoked. And cases in which the abuse of discretion standard was invoked with no unambiguous ruling do not appear to differ in any systematic respect from cases in which the standard was not invoked. Thus, clear error review and abuse of discretion review appear to have almost exactly opposite implications for the effect of ideology.

\(^{166}\) The results here are not independent of those in the previous Table, as this analysis is on a subset of the larger dataset. Thus, these results are not intended as a replication, but as corroboration that the “clear error” and “abuse of discretion” standards are driving the reversals in the larger analysis.
A cautionary note about the sample sizes should be added here. The appellate court ruled that the trial court committed clear error in only eighty cases out of over 6000 total cases—amounting to about 1.3 percent of the cases. Thus, although the figures in Tables V and VI approach or exceed standard measures of statistical significance, the total impact of these rulings may seem small. From an empirical perspective, however, the data allow us to note that reversals of findings of fact are not as unusual as the 1.3 percent figure might suggest. That is because many cases that reach the court of appeals do not involve any significant fact-finding by the trial judge, such as appeals of motions to dismiss, or in some jury trials. In fact, the clear error standard is only invoked in 1027 cases, or approximately 17 percent of the total. Of those, only 621 cases contained a definitive ruling on the clearly erroneous standard one way or the other. So in those cases where the issue of clear error was actually addressed and decided by the appellate court, clear error was found in approximately 14 percent of the cases.167 Thus, although explicit rulings of clear error constitute a small part of the appellate docket, as suggested below168 that may result from suboptimal strategy on the part of appellate counsel.

IV. EXPLAINING DEFERENCE

The results presented in Part III pose fundamental challenges to traditional ways of thinking about ideology and standards of review in the federal courts. The decision that one might expect to be the most clearly ideological—the decision of the appellate court to affirm or reverse the lower court’s legal conclusions under de novo review—is only ideological in the criminal, civil rights, and civil liberties contexts. Even these highly charged cases, however, have less ideological effect than the decision that one might expect to be the least ideological, clear error review of findings of fact. Surprisingly, cases invoking the clear error standard are more ideological than criminal cases and arguably more ideological than civil rights and civil liberties cases.169 Moreover, the two deferential standards of review studied—clear error and abuse of discretion—have almost exactly opposite implications for the effect of ideology on appellate outcomes. The next Part draws on the findings in Part III, piecing together the relationships of each of the elements of adjudication—law, fact, and discretion—with judicial ideology. The analysis compares each category and concludes that information asymmetry between the trial and appellate levels, not deference alone, accounts for the differences in appellate treatment of law, fact, and discretion.

167 Of course, these cases may be composed disproportionately of the stronger claims of clear error, but a similar figure appears in criminal cases, where the claims are likely weaker. See infra Part IV.F.2.

168 See infra Part IV.B.

169 Whether clear error is more ideological than civil rights/civil liberties depends on how one weights the Reverse/Affirm decision versus the Mixed/Affirm decision. Civil rights and liberties are slightly more ideological than clear error in the Reverse/Affirm decision, but clear error is considerably more ideological than civil rights and liberties in the Mixed/Affirm decision.
A. Law

The ideological difference between the trial court and the appellate court did not have a significant effect on reversal when all de novo cases were analyzed together. This result is consistent with the findings of similar studies, where the authors have been surprised by the results.170 As Table IV illustrates, however, the ideological effects become clear when the analysis drills down into different case categories and standards of review. Although surprising from the perspective of the standard attitudinal model in political science, the result is quite consistent with the informational approach taken in this Article. In the de novo cases, the appellate court can monitor the district court’s legal conclusions at low cost. The district court knows this, so in most cases a strategic district court would attempt to hew to the appellate court’s preferences.171 Hence, the prior studies have failed to find any ideological effect on the decision to reverse the trial court.172 As the results presented in Table IV show, however, the ideological effects are there; they are just found in the clear error standard of review, the criminal cases, and the civil rights/civil liberties cases.173 The results for the clear error cases are consistent with the informational theory developed in this Article, but what accounts for the differences between criminal, civil rights, and civil liberties cases and the other cases? The most obvious explanation is that these areas of law are simply more controversial,174 or perhaps raise issues of principle on which district judges are unwilling to compromise to avoid reversal. Another theory, however, would actually explain these cases on the same ground as the ideological effect in the context of clear error review. One of the more extensive studies of the federal district courts argued that ideology manifests itself at the district court level primarily in “new, ambiguous fact situations.”175

---

170 See, e.g., HETTINGER, LINDQUIST & MARTINEK, supra note 127, at 117 (describing the finding of no ideological effect as “remarkable” because “[t]he scholarship devoted to appellate reversal strongly indicates that the match (or mismatch) of preferences between reviewing judges and those under review matters a great deal.”).

171 See Guthrie & George, supra note 57, at 369–74 (explaining that a strategic district court would take the circuit’s preferences into account in decision-making).

172 See HETTINGER, LINDQUIST & MARTINEK, supra note 127, at 117.

173 See id. at 118 (suggesting that ideological considerations might not manifest themselves “in the decision of a panel to reverse a district court except under specialized conditions”). The Hettinger, Lindquist, and Martinek study, although it did not separately estimate ideological effects in civil rights/civil liberties cases, did note that ideological effects had been found in such cases in other studies. See id. (citing Haire et al., supra note 123, at 162).

174 See HETTINGER, LINDQUIST, AND MARTINEK, supra note 127, at 118 (“[I]t may be the case that ideological considerations do not manifest themselves in the decision of a panel to reverse a district court except under specialized conditions, namely when the case involved is highly charged in terms of ideological content.”).

175 ROWLAND & CARP, supra note 153, at 39.
The study notes that ideological case categories tend to “provide the judge with an inordinate amount of fact-finding discretion and, concomitantly, relatively little fear of being overruled by an appellate court.”\textsuperscript{176} The authors explained that because such cases tend to “involve primarily factual determinations that are seldom overturned by courts of appeals,”\textsuperscript{177} the trial court finds its ideological power maximized in fact finding, where conventional wisdom says that findings of fact are immune from review. This certainly seems like a possible explanation for the ideology of civil rights and civil liberties appeals,\textsuperscript{178} but the hypothesis is much more directly tested with the clear error rule itself, which leads naturally to the next section.

\textbf{B. Fact}

The result that reversals of findings of fact are more ideologically motivated than reversals of conclusions of law or discretionary decisions is counterintuitive and potentially important for understanding how appeals work. This finding contradicts an important piece of conventional wisdom in legal theory, that appellate courts almost never reverse findings of fact, and that the reversals that do occur are driven not by ideology but largely by manifest trial court errors. As an eminent legal text on appellate courts put it,

\begin{quote}
If your case arises in a settled area of the law and you lost at trial because the court found that the historical facts favored the other side, you have almost no chance on appeal. Such pure findings of fact generally can be attacked successfully only where they lack any rational connection to the record—where they seem to have materialized out of thin air—or occasionally when the vast weight of the evidence persuades the appellate court that a finding is surely wrong.\textsuperscript{179}
\end{quote}

\textsuperscript{176} Id. at 39–40.

\textsuperscript{177} Id. at 41. The authors cite here a quotation from a personal interview with U.S. District Judge Constance Motley, who observed

\begin{quote}
[t]he power [of the trial judge] is maximized when the outcome hinges on questions of fact. The judge can rely on those facts that persuade him and explain his decision in light of those facts. . . . If there is not an evidentiary basis for the judge’s interpretation of the facts, this interpretation, unlike interpretations of the law, is virtually immune from appellate review.
\end{quote}

\textsuperscript{178} Indeed, the effect of ideology under the clear error standard is smaller in Table IV (where criminal and civil rights cases are controlled) than in Table III (where they are not), suggesting that a fair amount of the ideology associated with the clear error standard comes from the criminal or civil rights cases.

\textsuperscript{179} DANIEL J. MEADOR, MAURICE ROSENBERG & PAUL D. CARRINGTON, APPELLATE COURTS: STRUCTURES, FUNCTIONS, PROCESSES, AND PERSONNEL 225 (1994).
This excerpt is illustrative of the literature on federal appeals, in that even those scholars who think deeply about standards of review in the federal courts have not identified evidence of ideological reversal of findings of fact.

Table V suggests why even a careful reading of many cases would not reveal the ideological nature of fact-finding without undertaking a quantitative analysis like the one in this Article. Explicit rejections of findings of fact occur with some regularity, but are still unusual. After all, in this sample of roughly 6,000 cases, only eighty instances of explicit rejections of trial court findings of fact appeared, accounting for only about 1 percent of the cases sampled. Even if an analyst had read those 6,000 cases, he or she would be unlikely to see the connections between clear error review and appellate outcomes. Moreover, it appears that in many cases the appellate court is not transparently engaging the trial court’s findings of fact, but “finding another way” to reverse. Only through the empirical analysis of a large number of cases, correlating the outcomes to the standard of review, does one see that the second assumption about fact-finding—that reversals of findings of fact are for objective, even clerical errors—appears to be wrong.\(^{180}\) Indeed, reversals of findings of fact reviewed under the clearly erroneous standard appear to be more ideologically driven than any other category or standard of review. Yet the strategic invocation of deference explanation—that trial or appellate judges strategically choose among law and fact—does not appear to have any support in the empirical analysis.

The empirical evidence is particularly striking, however, when we take into account the crudeness of the measures of ideology and fact-finding that are used in this analysis. The GHP scores are noisy measures of ideology, relying as they do on a single-dimensional measure of the President or senators involved in the judge’s appointment.\(^{181}\) Accordingly, we would expect that many of the individual judges’ ideologies do not correspond very closely to the scores that GHP Scores assign to them.\(^{182}\) The measures of appellate treatment of fact-finding, moreover, are rather noisy as well. In many cases, the appellate opinions do not explicitly say that lower court findings of fact are clearly erroneous or not clearly erroneous.\(^{183}\) As the summary statistics in the tables above suggest, there are many cases in which the appellate court invokes a standard of review without ever clearly stating whether the relevant decision stands or falls under that standard of review. Thus, there are undoubtedly many cases in which the courts rejected or affirmed factual

\(^{180}\) One counterargument might be that reversals of findings of fact are indeed for clerical, objective errors, but that appellate courts overlook clerical errors when trial courts are ideologically close. This possibility, however, seems less plausible than the conclusion drawn here, and is no less an ideological explanation than that in the text.

\(^{181}\) See Giles, Hettinger & Peppers, supra note 115, at 631.

\(^{182}\) If we had used this same technique on the Supreme Court, for example, it would predict that Justices Souter and Stevens would be conservative, a prediction that does not correspond closely with the individual justices. This example is particularly telling because Supreme Court justices are ideologically vetted more carefully than lower court judges.

\(^{183}\) See supra Tables V–VI.
findings that are not coded as such. So both the measures of ideology and the measures of appellate treatment of fact-finding are noisy. And yet the fact that these measures are noisy, and that the hypothesized effect is still statistically significant, strengthens the claim that ideology is a very powerful influence over findings of fact.184

All of this suggests that fact-finding is a key variable the district court can (and does) use to attempt to obtain its preferred case outcomes. As argued above, courts have two key variables they can manipulate to produce outcomes in individual cases: the findings of fact and the conclusions of law.185 The conclusions of law pronounced by the lower courts are observable and reversible by appellate courts at relatively low cost, and therefore deviations from the appellate court preferences are easily detected and reversed. The underlying facts of individual cases, however, are observable by the appellate court only at relatively high cost, and are difficult to reverse.186 This informational advantage means that the appellate court has strong reasons for deferring to the trial court in many cases—what some scholars have called “epistemic deference.”187 The informational advantage also means, however, that the trial court may attempt to exploit its private information about the case facts when the preferences of the trial court diverge from those of the appellate court. Thus, the findings of fact rendered by trial courts and reviewed by appellate courts may in fact be the trial court’s primary tool for achieving its ideological preferences over case outcomes.

In some cases, however, the trial court will go too far with its findings of fact, and the appellate court will reverse them. These reversals are particularly likely to occur when the trial court and the appellate panel diverge ideologically. The trial court acts first, so it does not know the decision that the appellate court will render or even the identity of the members of the appellate panel. Thus, the trial court must act on the basis of the expected ideology of the appellate panel. In contrast, the appellate court has the benefit of knowing not only the identity of the district judge, but also observing the decision itself. But the appellate court lacks a critical

184 Moreover, there are reasons to believe that the evidence for effect of ideology on reversals of findings of fact is even more compelling than presented here. Recall that the measure of appellate panel ideology used here was the median GHP score of the panel. In a fair number of cases where the appellate court reversed for clear error one judge dissented. In such cases we know the dissenting judge was not the median voter, so there is an argument that such scores should be included. If this adjustment is taken into account, the IdeologyDiff variable becomes much more significant. It was not used in the Tables in order to err on the conservative side, but it is reassuring that these adjustments increase the significance level of the IdeologyDiff variable.

185 See supra notes 82–89, 146–159 and accompanying text.

186 See supra notes 90–95 and accompanying text.

187 Horwitz, supra note 18, at 1085 (“epistemic deference” means that “courts defer to other institutions when they believe that those institutions know more than the courts do about some set of issues, such that it makes sense to allow the views of the knowledgeable authority to substitute for the courts’ own judgment”). Authors distinguish from “legal authority” as a ground for deference. Id.
piece of information in the fact-finding context—the private information about the facts that are not revealed by the record. Thus, the appellate court must rely on the perceived ideology of the district court to infer the private information of the district court, disproportionately reversing ideologically dissimilar district courts when clear error review applies—exactly what the empirical evidence presented above suggests. Thus, trial court judges are probably more likely to take liberties with the facts than with trial procedure or legal conclusions, and appellate judges are correspondingly more likely to reverse findings of fact on ideological grounds than trial procedure or legal conclusions. This Article argues that the reason is because clear error cases show a strong ideological effect and the abuse of discretion and de novo cases do not.

C. Discretion

The abuse of discretion standard provides an interesting point of comparison with clear error. Both are deferential standards of review, both are associated with a decrease in the probability of a “Reverse” outcome at the appellate level, and trial court rulings under both are roughly comparable in terms of the proportion of cases in which they are upheld and overturned. Yet the clear error standard provokes the most ideological disagreement of any of the explanatory variables and abuse of discretion shows no evidence of ideological behavior. The fact that both standards are deferential both in theory and in empirical effect shows that deference alone is not the cause of the ideological divide between trial and appellate courts, but rather something else. This Article argues that the “something else” is the asymmetric information enjoyed by the district court in the fact-finding arena but not in the discretionary arena.

What then is the role of deference in abuse of discretion review? The key factor courts have identified as counseling deferential review in these cases is the 

---

188 There may have been a time when courts were more likely to reverse findings of fact than for abuse of discretion, but the data suggests that this is no longer true. See Louis, supra note 10, at 999–1000 (“Traditionally, appellate courts were much less likely to find an abuse of discretion in the procedural or evidentiary context than to reverse a trial level finding of ultimate fact going to the merits. That is no longer true, or at least as true. Appellate courts today more readily find abuses of procedural discretion and, in some substantive areas like personal injury, less readily overturn findings of ultimate fact than they did a half century ago. Overall, appellate review of the two discretionary functions is coalescing, and what differences remain perhaps are more statistical and atmospheric than doctrinal.”).

189 Indeed, the abuse of discretion standard appears to show evidence of “anti-ideological” behavior, in the sense that under that standard an appellate court that is ideologically farther from the trial court is more likely to affirm the trial court’s decision.

190 Abuse of discretion can involve private information of the trial court as well. See Rosenberg, supra note 35, at 663–65 (explaining that sometimes such a ruling is “based on facts or circumstances that are critical to decision and that the record imperfectly conveys”). But the theory developed in Part II argues that private information is much more prevalent and consequential in the clear error scenario.
“impracticability of formulating a rule of decision” because they involve “facts that utterly resist generalization” or because of their “novelty.” In cases where “courts’ ability to articulate legal rules in sufficient detail or with sufficient clarity that those rules will adequately bind future courts” is not present, the rationale for de novo review is weakened, and deferential review strengthened. Although some have theorized that judges tend to classify more important issues as questions of law in order to retain control over the issues, and less important issues as matters of discretion, this is not exactly what this Article is arguing. Rather, the argument is that in abuse of discretion cases the cost of formulating detailed legal rules exceeds the benefit, so judges rationally engage in a sort of abstention from appellate review. In this respect, fact-finding and discretion are similar; appellate review of these categories is unlikely to yield any generally applicable rules of law. But deference to fact-finding of the district court has a role in shaping the district court’s incentives, a theme that is developed in the next section.

D. Theorizing Institutionalized Deference

The comparison of law, fact, and discretion makes it possible to examine the underlying purposes of deferential review with a theoretically and empirically driven perspective. The most common articulated reason for deferential review in the factual context is one rooted in institutional competence—the idea that by their situation or their experience the trial judges find facts better than appellate judges. The support for this argument is typically based on the judge’s presence in the courtroom and interaction with the facts, including the opportunity to observe the demeanor of witnesses, nonverbal communication, and other information that is not evident in the printed record. A second argument is that the trial judge develops a certain expertise in fact-finding that comes with experience. These two factors, taken together, are thought to lead to the

191 Id. at 662.
192 Chad M. Oldfather, Universal De Novo Review, 77 GEO. WASH. L. REV. 308, 326 (2009) (arguing that universal de novo review seems to lack a single coherent justification for the practice).
193 Louis, supra note 10, at 1045–46 (“The designation [of discretionary review], however, is applied primarily to the less important questions. The more important procedural questions tend to be classified as questions of law.”).
194 See Oldfather, supra note 50, at 444–45 (“Notions of relative institutional competence form the grounds on which justifications of appellate deference to trial-level fact finding are almost universally formulated.”); see also Ellen E. Sward, Appellate Review of Judicial Fact-Finding, 40 U. KAN. L. REV. 1, 23 (1991) (“The clearly erroneous standard is simply part of the mechanism for allocating fact-finding authority to the court viewed as (usually) most competent to determine the facts.”).
195 Oldfather, supra note 50, at 445–47.
196 Id. at 447–48; see also Anderson v. City of Bessemer, 470 U.S. 564, 574 (1985) (“The trial judge’s major role is the determination of fact, and with experience in fulfilling that role comes expertise.”).
conclusion that “appellate courts are not very good at fact-finding,” or sometimes that the appellate court should yield to the “unchallenged superiority of the [trial] court’s fact-finding ability.” In reality, however, there is a good efficiency argument for deferential review even if the trial judge is not a “superior” fact-finder. So long as the trial judge has adequately performed his or her fact-finding task, any appellate effort will be duplicative, and duplication of efforts is inefficient.

The institutional competence argument, if it is accurate, is one reason why the district judge might have superior information about the facts than the appellate judge. The argument is incomplete, however, in that it ignores two important factors that are necessary to make it work. First, if the preferences of the trial judge and the appellate court are different, the trial judge may have incentives to reveal his or her private information selectively (or not at all), making deference potentially dangerous. Second, the standard institutional competence rationale ignores another important effect of deference: the effect of incentives on the trial court’s behavior. The trial judge may have diminished incentives to expend the resources to develop information about the facts if the appellate court does not defer. Thus, the conventional rationale says that the appellate court defers to the trial judge because the trial judge is better informed, but this Article adds that the trial judge is better informed in part because of the appellate court’s deference. As a result, if asymmetric information is an important advantage of trial judges in producing their favored outcomes, then eliminating that advantage will reduce the trial judge’s incentive to carefully, thoroughly investigate factual matters.

---

197 Oldfather, supra note 50, at 439 (arguing that this “conventional wisdom is misguided” because appellate courts have some advantages over trial courts in fact-finding).


199 See Sward, supra note 194, at 24 (arguing that efficiency in the form of saving “the parties and society a large expenditure of legal and judicial resources. . . . is probably the best rationale for limited review of trial court factfinding”).

200 Anderson, 470 U.S. at 574–75 (“The rationale for deference to the original finder of fact is not limited to the superiority of the trial judge's position to make determinations of credibility. . . . Duplication of the trial judge's efforts in the court of appeals would very likely contribute only negligibly to the accuracy of fact determination at a huge cost in diversion of judicial resources.”).

201 Some scholars have argued that the purported institutional advantages of the trial court in fact-finding are overstated. See, e.g., Oldfather, supra note 50, at 444–66.

202 An interesting question is whether this equilibrium behavior actually constitutes “deference.” One careful analysis of deference has argued that “deference involves a decisionmaker following a determination made by some other individual or institution that it might not otherwise have reached had it decided the same question independently.” Horwitz, supra note 18, at 1072. That definition seems technically satisfied by the behavior described here, if perhaps not within the spirit of the definition.

203 Of course, not all judges would behave this way. Some judges are not particularly ideological, and others would feel a sense of duty to investigate the facts regardless of whether his or her rulings will receive deference. But some judges would behave this way,
The institutional rationale of deference, therefore, can be divided into top-down communication and bottom-up communication. In clear error review, the problem is communicating information about facts from the bottom upward to the appellate court. The trial court may have incentives to shirk its efforts to develop information if the appellate court does not defer, but the appellate court has trepidation about deferring when the trial court has ideologically different preferences. In abuse of discretion review, the problem is communicating preferences about legal rules from the top downward to the trial courts. The cost of formulating detailed rules about procedural and evidentiary matters exceeds the benefit, so deference here is a form of rational abstention. The analysis in this Article uncovers a link between ideological conflict and deference. But it is not the deference itself that creates the ideological conflict in fact-intensive environments. It is the relative immunity to appellate review that creates the incentives for a trial court to develop private information about the facts, which in turn leads to ideological conflict. In the abuse of discretion arena, the appellate courts give the trial court deference, but there is no private information for the trial courts to exploit.

The positive political theory literature in general has been hampered in this regard by failing to take seriously one of the principal ostensible purposes of deferential standards of review—the asymmetric information between the trial and appellate courts. This is a specific example of a broader critique of attitudinal and positive political theory approaches to the judiciary—that they fail to take law and legal institutions seriously. See, e.g., Kim, supra note 24, at 386–87.

Instead of testing the constraining effects of legal institutions, the empirical scholarship is generally dismissive of them. This is ironic because at the same time, the empirical scholarship seems to accept that fact-finding is an objective, exogenous “given” in legal analysis. Moreover, standards of review deal with one of the most central legal institutions in courts—the division of labor in finding facts and legal interpretation—and implicate the core methodological strengths of positive political theory—institutions, ideology, and information. The purpose of this Article is to show how the positive political theory tradition can so we must think of the incentives of the “bad judge” who, like Holmes’s “bad man,” acts in his or her own self interest. See Oliver Wendell Holmes, The Path of the Law, 10 HARV. L. REV. 457, 459 (1897) (“If you want to know the law and nothing else, you must look at it as a bad man, who cares only for the material consequences that such knowledge enables him to predict, not as a good one, who finds his reasons for conduct, whether inside the law or outside of it, in the vaguer sanctions of conscience.”).

This is a specific example of a broader critique of attitudinal and positive political theory approaches to the judiciary—that they fail to take law and legal institutions seriously. See, e.g., Kim, supra note 24, at 386–87.

There are very few empirical pieces attempting to use standards of review as a datum, at least outside the administrative law context. One recent exception is Corey Yung. See generally Yung, supra note 63 (introducing a new technique for measuring the ideology of judges based up on judicial behavior in the U.S. courts of appeals).

Friedman, supra note 152, at 272.
shed light on the institutional mechanisms of deference that enable credible communication of information through the judicial hierarchy.

E. Data Issues

The data in this analysis comes from a sample of published appellate decisions, which raises the potential issue of selection bias. Most legal disputes never go to court at all, most of those that go to court are settled, and most of those that are not settled are never appealed, leading to a potentially significant selection effect of relying on appellate decisions. Furthermore, even among these appellate decisions, this study makes use of only “published” opinions even though many are “unpublished,” creating another possible selection effect.207 The unpublished decisions may differ in significant respects from published decisions, for example, by having a higher affirmance rate.208

The population from which one is trying to draw inferences is the key issue in selection bias. The fact that few disputes make it all the way to a federal appellate court, and that those few disputes are not representative of the underlying population of disputes means that the data in this study may not be representative of the underlying population of legal disputes. The goal of this study, however, is not drawing inferences about the universe of disputes, but about standards of review on appeal.209 Once we confine the scope of the inquiry in this study to appellate opinions, it is not clear how selection bias might undermine the results. Specifically, it is not clear why ideologically divergent courts would be more likely to publish decisions when they reverse findings of fact than nonideological courts. On the other hand, it does seem possible that litigants are more likely to appeal the findings of fact when the district court judge is ideologically divergent from the center of the circuit. This hypothesis must await further research.

F. Normative Implications and Directions for Future Work

1. Choosing Standards of Review

One of the most obvious normative questions this study could shed light on is the question of what standard of review should apply in different appellate contexts. A consensus seems to have formed among the commentators that a “functional” approach is better than static doctrinal categories for allocating review standards,210 but the commentators’ functional approaches differ greatly. One

207 See id. at 271.
208 See Guthrie & George, supra note 57, at 361–62 (comparing the affirmance rate for published opinions to the affirmance rates for opinions overall, and deducing that unpublished opinions have a “much higher” affirmance rate).
209 A problem with using unpublished decisions is that many are so perfunctory that they cannot be used for meaningful analysis. See, e.g., Wald, supra note 3, at 1373–74 (referring to unpublished opinions as “in effect a class of legal ‘untouchables’”).
210 See, e.g., Miller v. Fenton, 474 U.S. 104, 104 (1985) (holding that Congress did
article suggested the law-fact dichotomy, be replaced with “direct observation” (deferential review) versus “reflection on the evidence” (nondeferential review).\textsuperscript{211} Another approach asks the appellate court to “openly grapple with the question of institutional competence in the context of assessing factual matters” to decide the standard of review.\textsuperscript{212} What these functional approaches have in common is that they reject static categories of “law” and “fact” as providing the appropriate dividing lines between deferential and nondeferential review.

The analysis in this Article would suggest a different functional allocation of responsibility between the trial and appellate courts. The degree of deference in the standard of review should take into account the incentives for fact-finding that the standard of review will create for the district court. The logic of this argument could potentially point to problems with reform proposals focusing on standards of review. For example, some scholars have argued that the institutional competence rationale for deference to the trial court in fact-finding is exaggerated.\textsuperscript{213} That assertion leads to the conclusion that courts should be more willing to “reexamine facts” where factual accuracy is particularly important, as in the criminal context.\textsuperscript{214} The informational approach developed in this Article suggests the possibility that increased scrutiny at the appellate level could decrease the district judge's incentive to invest in factual investigation. Overall, that would mean lower quality fact-finding at the trial level, increased appellate scrutiny, and diversion of resources from the court system. Without examining the incentives of the district court, one might overlook this potential perverse effect of a reform effort.

Moreover, whether the appellate court can simply impose a standard of review and achieve the desired results depends crucially on causal arguments about the effect of the standard of review. Take, for example, the debate that has surrounded \textit{United States v. Booker}\textsuperscript{215} regarding the standard of review in criminal sentencing decisions.\textsuperscript{216} The findings in this Article could shed light on the debate over the extent to which the standard of review adopted in such cases will actually affect district court decision-making. Indeed, the effect of clear error on exacerbating ideological disagreement was strongest in the criminal context, suggesting a role for this analysis in that debate.\textsuperscript{217} But without knowing whether
the standards of review themselves cause deference, as opposed to the context in which the standards are applied causing the deference, one cannot decide between the alternatives on an informed basis. The analysis in this Article suggests the strong possibility that the context of deferential review—asymmetric information—provides the explanation for why deference is rational, not necessarily the command of deference. Resolving this question must await future research.218

2. Implications for Appellate Advocacy

Finally, the results of this study may suggest some implications for the practice and strategy of appellate litigation. The conventional wisdom argues that the facts are “fixed” on appeal and that there is little point in challenging them. In most cases, appellate advocates will focus on other issues perceived to have a higher probability of success.219 The analysis in this Article suggests that there may be more opportunities to prevail on appeal through factual argumentation than is commonly believed. If indeed the factual context can trigger reversal when the trial court is ideologically distant from the appellate panel, appellants might benefit from an argument of clear error. The “outcomes analysis” conducted in this Article suggests that merely looking at the proportion of cases in which appellate courts rule that fact-finding is “clearly erroneous” may not give a full picture as to how often factual findings actually trigger reversal. There is some evidence, it would seem, that reversals of findings of fact are not as unusual as legal practitioners believe. It seems possible that some losing litigants could have actually won challenges to findings of fact if they had made those challenges.220

Indeed, even if one does look at cases in which the appellate court explicitly found clear error in factual findings, one still might counsel more appeals on this ground. As discussed above,221 in cases where the issue of clear error was actually addressed and decided by the appellate court, clear error was found in approximately 14 percent of the cases. Of course, cases in which the clear error issue was decided are probably biased towards those in which there actually was evidence of district court factual error because appellate advocates probably avoid

218 Some excellent work along these lines has already been done by Schanzenbach and Fischman, who use a “natural experiment” of varying review standards to attempt to make causal arguments about whether “standards of review matter.” See Schanzenbach & Fischman, supra note 63, at 413.

219 See, e.g., Meador et al., supra note 179, at 284–85 (encouraging appellate advocates to attempt to recharacterize factual issues as conclusions of law because of the difficulty of challenging trial court findings of fact).

220 There are good reasons why attorneys in the noncriminal context would avoid raising clear error of findings of fact as an issue on appeal. Attorneys who are repeat players would probably suffer a reputational cost from regularly asking appellate panels to engage the trial court’s findings of fact, especially when there is no objective evidence of error. In criminal cases, these constraints are less binding, and judges know it.

221 See supra note 167 and accompanying text.
these arguments when they are not strong. But even in criminal cases, where convictions are regularly appealed automatically and indiscriminately, the rate of reversal of findings of fact was 9 percent. Thus, even in a case category where extraneous issues are not weeded out by appellate counsel, there is approximately a 10 percent chance of winning on the issue—a chance that could be worthwhile.

V. CONCLUSIONS

The standard of review is invoked in almost every appellate opinion, but the analysis in this Article suggests that scholars and legal practitioners have little clear understanding of how these doctrines work in practice. This Article takes some preliminary steps toward sketching a theory about the role of standards of review in the federal courts, and the possibility of deference in the judiciary more generally. The results support the proposition that standards of review “actually matter” as a leading doctrinal treatise asserts. But standards of review do not matter in the simple, conventional sense of increasing the affirmance rate and decreasing the reversal rate. In fact, deferential standards of review tend to promote “mixed” results and have an ideological dynamic of their own.

The empirical evidence in this Article suggests that findings of fact are indeed ideological in nature—more ideological than almost any other issue category, including the perennially explosive areas of civil rights/civil liberties. In spite of the ideologically charged nature of fact review, the analysis also suggests that ideology does not significantly affect the choice of standards of review themselves—the results show no evidence that trial or appellate judges manipulate the standard of review to achieve ideological goals. Thus, the evidence supports the conclusion that both the “legal model” and the political science literature on judicial politics are missing an important part of how plays out in a judicial hierarchy.

222 Most criminal convictions are appealed, and in some jurisdictions, up to 90 percent of criminal defendants appeal. See Oldfather, supra note 50, at 489 n.225. In such appeals, it should be clear that there is very limited self-selection in the decision to appeal, making these cases potentially better measures of the likelihood of reversal of findings of fact.

223 The problem is compounded (or perhaps evidenced) by the fact that standards of review are given “short shrift” in law teaching, and usually edited out of casebooks and relegated to the interstices of the traditional law school curriculum. See Edwards & Elliott, supra note 2, at vii.

224 Childress & Davis, supra note 1, at 1–2.
BARGAINED JUSTICE: PLEA-BARGAINING’S INNOCENCE PROBLEM
AND THE BRADY SAFETY-VALVE

Lucian E. Dervan∗

INTRODUCTION

On January 17, 1959, James Zeno walked deliberately up three flights of
stairs and knocked on the door of Beatrice Lynumn’s Chicago apartment.2 Zeno,
who had been arrested by Chicago police for possession of narcotics earlier that
day, was acting as an informant in return for a promise of leniency.3 Lynumn
opened the door and Zeno walked inside.4 After a brief interaction, Zeno left the
apartment with a package of marijuana under his arm.5

A few minutes later, Lynumn was in custody.6 Despite the strong evidence
provided by Zeno, she vehemently professed her innocence as police filed into her

∗ © 2012 Lucian E. Dervan. Assistant Professor of Law, Southern Illinois University
School of Law, and former member of the King & Spalding LLP Special Matters and
Government Investigations Team. Special thanks to Professors Marc Miller, Ellen Podgor,
Christopher Slobogin, Andrew Leipold, Russell Covey, Jenny Roberts, Nancy Levit,
Vanessa Edkins, Paul McGreal, Peter Alexander, William Schroeder, Christopher Behan,
Nathan Sales, Luke Milligan, and Lea Johnston for their valuable comments regarding
earlier versions of this article and to my research assistants Angela Boley and Elizabeth
Boratto. Thanks also to Southern Illinois University School of Law for providing the
summer research grant that funded this research and to the Southeastern Association of
Law Schools (SEALS) for the opportunity to present a draft of this paper at its 2010
conference.

3 See id. On January 17, 1959, three Chicago police officers arrested James Zeno for
unlawful possession of narcotics. They took him to a district police station. There they told
him that if he “would set somebody up for them, they would go light on him.” He agreed to
“cooperate” and telephoned [Lynumn], telling her that he was coming over to her
apartment. Id.
4 Id.
5 Id.
6 See id. at 529–30. Officer Sims testified as follows:

He called Beatrice and said he had left his glasses in the apartment; she
opened the door and as she came out into the hall, I was standing in the common
hall, in the vestibule part with the door partly closed. As she walked down the
hallway toward Zeno, I opened the door and stepped into the hallway. I told her
she was under arrest and I grabbed her by her hands, both hands. At this point, I
told her that she had been set up, that she had just made a sale and I showed her
the package.

See id. at 529 n.1.
apartment. At that moment, the issue of most concern to Lynumn was not her guilt or innocence, however, but the fate of her children, ages three and four. The concern was particularly acute given that the children’s father had passed away and Lynumn was raising them on her own. Perhaps seeing an opportunity, the police began their interrogation right there in the apartment.

Then he [the police officer] started telling me I could get 10 years and the children could be taken away, and after I got out they would be taken away and strangers would have them, and if I could cooperate he would see they weren’t; and he would recommend leniency and I had better do what they told me if I wanted to see my kids again.

According to Lynumn, the offer of leniency and the belief that she would be able to remain with her children created an overwhelming motivation to cooperate. As a result, she confessed to the police in her apartment. Lynumn later stated, “[t]he only reason I had for admitting [guilt] to the police was the hope of saving myself from going to jail and being taken away from my children.”

Lynumn later recanted her admission of guilt and proceeded to trial where, despite her renewed proclamations of innocence, the confession was entered into evidence. Rather than refute Lynumn’s account of the events, the police officers involved admitted that the events had transpired largely as the defendant had described. Based on the weight of her earlier admission of guilt and Zeno’s testimony, Lynumn was convicted and sentenced to not less than ten years in prison. Just as the officers had accurately predicted would occur during her

---

7 See id. at 530.
8 Id. at 531.
9 Id.
10 Id.
11 Id.
12 See id. at 531–32.
13 Id. at 532.
14 See id. at 529.
15 See id. at 532. According to one officer:

I asked her who the clothing belonged to. She said they were her children’s. I asked how many she had and she said 2. I asked her where they were or who took care of them. She said the children were over at the mother’s or mother-in-law. I asked her how did she take care of herself and she said she was on ADC. I told her that if we took her into the station and charged her with the offense, that the ADC would probably be cut off and also that she would probably lose custody of her children. That was not before I said if she cooperated, it would go light on her. It was during that conversation.

16 Id. at 533.
16 See id. at 529.
arrest, Lynumn was sentenced to a decade behind bars for challenging the state to prove her guilt in court rather than cooperating in her own prosecution.\footnote{See id. at 531 (“Then he started telling me I could get 10 years and the children could be taken away . . .”)}

In 1963, the United States Supreme Court took up the issue of the admissibility of Lynumn’s confession.\footnote{See id. at 528.} In a unanimous opinion, the Justices concluded that the confession was inadmissible as a coerced, rather than voluntary, statement.\footnote{Id. at 534.} The Court stated, “[i]t is thus abundantly clear that the petitioner’s oral confession was made only after the police had told her that state financial aid for her infant children would be cut off, and her children taken from her, if she did not ‘cooperate.’ These threats were made while she was encircled in her apartment by three police officers and a twice convicted felon who had purportedly ‘set her up.’”\footnote{Id. (quoting Blackburn v. Alabama, 361 U.S. 199, 208 (1960)).} Based on precedent establishing that confessions must be both voluntary and intelligent, the Court stated that where the defendant’s “will was overborne at the time [s]he confessed,” the confession cannot be deemed the product of a “rational intellect and a free will.”\footnote{Id. (“We think it clear that a confession made under such circumstances must be deemed not voluntary, but coerced.”).} The court concluded that it was a blatant violation of the United States Constitution for the police to threaten Lynumn with the reality of the situation if she failed to cooperate through confession, even though they were correct when they told her that she faced a possible sentence of ten years in prison which would result in her losing custody of her children.\footnote{See id. (“We think it clear that a confession made under such circumstances must be deemed not voluntary, but coerced.”).}

In 2004, just over forty years after the Lynumn case, another mother of two small children was brought before the criminal justice system, and, once again, she was threatened with ten years in prison unless she agreed to cooperate.\footnote{See Michelle S. Jacobs, Loyalty’s Reward—A Felony Conviction: Recent Prosecutions of High-Status Female Offenders, 33 FORDHAM URB. L.J. 843, 857 (2006); Mary Flood, Lea Fastow in Plea-Bargain Talks; Former Enron CFO’s Wife Could Get 5-Month Term but Deal Faces Hurdles, HOUS. CHRON., Nov. 7, 2003, at A1.} Also like Lynumn, the threat of ten years in prison meant that her children would be without both parents.\footnote{See Jacobs, supra note 23, at 857.} As might be expected, the defendant once again acquiesced to the demands of her accusers, yet this time the Supreme Court was silent.\footnote{See Mary Flood, Wife of Former Enron Chief Financial Officer Begins Prison Sentence in Houston, HOUS. CHRON., July 13, 2004, at A1.}

Lea Fastow served as Director and Assistant Treasurer of Corporate Finance at the now infamous Texas energy trading company Enron from 1991 until 1997.\footnote{Jacobs, supra note 23, at 856.} Her husband, Andrew Fastow, was also an integral part of the corporation and served as Chief Financial Officer from 1997 until the corporation’s collapse in
2001. Although Lea Fastow was a stay-at-home mother raising two small children in 2001, federal investigators determined that she had known of her husband’s fraudulent financial dealings and had even assisted him in perpetrating the frauds. In response, the government indicted Lea Fastow along with her husband, who had already been charged with ninety-eight counts of criminal conduct. Lea Fastow’s indictment contained six counts, including conspiracy to commit wire fraud and defraud the United States, money laundering conspiracy, aiding and abetting, and filing a false tax return.

Conviction on all six counts of the indictment would result in a prison term of up to ten years, but the government was more interested in persuading Lea Fastow to cooperate than in convicting her. As a result, the government offered her a deal. In return for confessing her guilt in court, the prosecution would only charge her with a single count of filing a false tax return and the government would recommend a sentence of one year of supervised release. The deal also included an agreement that Lea Fastow and her husband, who also intended to plead guilty in return for leniency, would not have to serve their prison sentences simultaneously, thus ensuring their children would always have one parent at

---

27 Id.
28 See id. at 856–57.

During the time in question, Andrew Fastow and Michael Kopper created several Special Purpose Entities (SPEs) to hold off-balance sheet treatment of assets held by Enron. Ms. Fastow assisted with concealing the fraudulent nature of two of the SPEs. In both cases, Ms. Fastow accepted “gifts” in her name and in the names of her children, knowing that the gifts were kickbacks. In another instance, the Fastows were attempting to hide the fact that Ms. Fastow’s father was used as an “independent” third party of RADR. When the Fastows realized that the father’s ownership would trigger a reporting requirement, they had him pull out of the deal. Ms. Fastow convinced her father to file a false tax return in an effort to continue hiding their involvement in the SPE.

Id.; see also Flood, supra note 23, at A1.


30 Id.


As the lead prosecutor in the case stated, “[t]he Fastows’ children can be taken into account in deciding when Andrew Fastow will begin serving his sentence. There is no reason for the government, when it can [avoid it], to have a husband and wife serve their sentences at the same time.”

For Lea Fastow, the reality of her situation removed any free will she might have had to weigh her options. With two small children at home and the prospect that she and her husband might both be in prison at the same time, the decision to accept the offer was made for her. As one family friend stated, “[i]t’s a matter of willing to risk less when it’s for her children than she would risk if it were just for herself.” As such, she succumbed to the pressure to confess her guilt and accepted the deal.

Eventually, the judge forced the government to revise its offer because he believed five months was too lenient. As a result, Lea Fastow pleaded guilty to a misdemeanor tax charge and was sentenced to one year in prison. Nevertheless, the agreement to confess her guilt in return for a promise of leniency lessened her sentence by nine years and ensured her children would not be without a parent. As promised, Andrew Fastow was not required to report to prison for his offenses until after Lea Fastow was released.

---

34 Jacobs, supra note 23, at 859 (“During the renegotiation of the second plea, it was widely reported that Ms. Fastow was interested in a plea that would allow her children to stay at home with one parent while the other was incarcerated, rather than running the risk that both parents would be incarcerated at the same time. The government apparently acquiesced to this request.”).


36 See Lynum v. Illinois, 372 U.S. 528, 534 (“[T]he question in each case is whether the defendant’s will was overborne at the time he confessed. If so, the confession cannot be deemed ‘the product of a rational intellect and a free will.’” (citations omitted) (quoting Blackburn v. Alabama, 361 U.S. 199, 208 (1960))).

37 See Greg Farrell & Jayne O’Donnell, Plea Deals Appear Close for Fastows; Could Mean Task Force on Verge of Filing More Charges, USA TODAY, Jan. 8, 2004, at B.01 (“One of the reasons that Lea Fastow wants to limit her jail time to five months is that she and her husband have two young children, and they’re trying to structure their pleas so they’re not both in jail at the same time.”).

38 Flood, supra note 23, at A1 (“A family friend previously said Lea Fastow is willing to consider pleading guilty and forgoing a chance to tell her side to a jury because it would be better for her two small children and could ensure they would not be without a parent at home.”).

39 See Mary Flood, Fastows to Plead Guilty Today; Feds Now Focus on Skilling, Lay, HOUS. CHRON., Jan. 14, 2004, at A1 (“The plea bargains for the Fastows, who said they wanted to be sure their two children are not left parentless, have been in limbo for more than a week.”).

40 See Farrell & O’Donnell, supra note 37, at B.01 (“U.S. Judge David Hittner told Lea Fastow Wednesday that he refused to be locked in to the five-month prison sentence that her lawyers had negotiated with prosecutors.”); Flood, supra note 25, at A1.

41 See Farrell & O’Donnell, supra note 37, at B.01; Flood, supra note 25, at A1.

42 See Flood, supra note 25, at A1.
In a 1942 Supreme Court case, the Justices wrote, “a plea of guilty coerced by a federal law enforcement officer is no more consistent with due process than a conviction supported by a coerced confession.” Yet Lea Fastow’s guilty plea, the product of the same threats the Supreme Court found coercive forty years before, was simply assumed by everyone involved to be constitutional. This type of assumption seems well supported, as Lea Fastow’s guilty plea represents just one of hundreds of thousands of similar deals entered into by prosecutors and defendants every year in the American criminal justice system. The first question this Article seeks to answer, therefore, is how and why the Supreme Court shifted from being a protector of coerced defendants to being a supporter of the plea bargaining machine. The answer involves a great compromise by the Justices in 1970, a compromise necessitated by strains on the criminal justice system resulting from the additional rights afforded to defendants during the Warren Court’s due process revolution and the crushing and ever-increasing number of criminal prosecutions in America. The great compromise was to allow prosecutors and defendants to bargain for justice in hopes this might alleviate pressures on the system while simultaneously benefitting clearly guilty defendants willing to reap an advantage in return for an admission of guilt.

Unfortunately, the plea bargaining system has not evolved as the Justices had hoped. Rather than a system in which guilty defendants bargain for a mutually beneficial outcome and innocent defendants wage forward toward trial in hopes of acquittal, the plea bargaining system has come to engulf almost everyone. Today, the incentives to bargain are powerful enough to force even an innocent defendant to falsely confess guilt in hopes of leniency and in fear of reprisal. The second question this Article seeks to answer, therefore, is whether, in 1970, the Supreme Court was willing to permit the incentives offered for bargaining to become so powerful as to induce even innocent defendants to plead guilty.

---

45 See infra Part II (examining the Supreme Court’s great plea bargaining compromise of 1970).
46 See infra Part II.B (examining the reasons for the Supreme Court’s decision in United State v. Brady).
47 See infra Part I.B.
48 See infra Part III (examining plea bargaining’s innocence problem).
49 See infra Part III.A (examining the prevalence of innocent defendants pleading guilty).
50 See infra Part III.
51 See infra Part III.B (examining the Supreme Court’s safety-valve in United States v. Brady).
In answering this question, this Article proposes that the Supreme Court in *Brady* specifically contemplated the possibility that prosecutors would utilize the plea bargaining system in an impermissible manner and, as a result, created both a safety-valve and a litmus test for a failure of the safety-valve. The *Brady* safety-valve limits the amount of pressure that may be asserted against defendants by prohibiting prosecutors from offering incentives in return for guilty pleas that are so coercive as to overbear defendants’ abilities to act freely. The number of innocent defendants who are induced to falsely confess their guilt is a litmus test to determine when prosecutors have applied excessive pressure and the safety-valve has failed. The *Brady* Court stated that if the plea bargaining system began to operate in a manner resulting in a significant number of innocent defendants pleading guilty, the Court would reexamine the constitutionality of bargained justice. The significant innocence problem that plea bargaining has today indicates that the *Brady* safety-valve has failed and, as a result, the constitutionality of modern day plea bargaining is in great doubt.

In Part I, this Article examines the rise of plea bargaining in America and explores the forces behind this jurisprudential revolution. In Part II, this Article analyzes Supreme Court precedent regarding plea bargaining dating back to the 1800s. In so doing, this Article demonstrates that the Supreme Court was strongly opposed to bargained justice for over a century prior to its great compromise of 1970. Finally, in Part III, this Article examines a significant portion of the Supreme Court’s great compromise that has gone largely unnoticed by scholars, practitioners, and judges. The Supreme Court created a safety-valve in 1970 to ensure that the pressures placed on defendants to plead guilty would never reach impermissible levels. As this Part demonstrates, there is a significant innocence problem with plea bargaining today. This being the case, the *Brady* safety-valve may not be working and, if so, defendants, both innocent and guilty, are being offered unconstitutional incentives to confess.

52 See infra Part III.B.
53 See infra Part III.B.
54 See infra Part III.B.
55 See infra Part III.B.
56 See infra notes 257–261 (examining the literature regarding plea bargaining’s current innocence problem).
57 See infra Part I.
58 See infra Part II.
59 See infra Part II.
60 See infra Part III.
61 See infra Part III. For purposes of this article, the terms “innocent” and “innocence” are used to refer to any defendant who would not have been convicted at trial. Of course, this is not to say a defendant is necessarily free of moral culpability or responsibility for the charged conduct. Nevertheless, the terms “innocent” and “innocence” herein serve as terms to discuss those who are not criminally liable in the eyes of the law.
62 See infra Part III.
I. THE RISE OF THE PLEA BARGAINING MACHINE

Before delving deeply into Supreme Court plea bargaining precedent, it is important to understand the rise and operation of the modern day plea bargaining machine. Of particular significance is the fact that plea bargaining began its rise to dominance well before it had been approved as a form of justice by the Supreme Court in 1970. In fact, plea bargaining began to grow in popularity during periods of American history when Supreme Court precedent specifically found the practice of offering incentives for defendants to plead guilty unconstitutional.63

A. The Historical Rise of Plea Bargaining

The evolution of plea bargaining into a force that affects over 95 percent of defendants in the American criminal justice system took place mainly in the nineteenth and twentieth centuries.64 While the right to plead guilty dates back to English common law traditions, a new phenomenon began to appear in America shortly after the Civil War. It was during this period that state courts began witnessing an influx of appellate cases dealing with apparent “bargains” between defendants and prosecutors.65 With resounding frequency, these early experiments with bargained justice were rejected by the judiciary as demonstrated by the case excerpts below.66

The least surprise of influence causing [the defendant] to plead guilty when he had any defense at all should be sufficient cause to permit a change of the plea from guilty to not guilty.67

63 See infra Part II.B (discussing the Supreme Court’s rejection of bargained justice prior to 1970).
65 See Albert W. Alschuler, Plea Bargaining and Its History, 79 COLUM. L. REV. 1, 19 (1979) (“It was only after the Civil War that cases of plea bargaining began to appear in American appellate court reports.”).
66 See id. at 19–21.
67 Id. at 20.
No sort of pressure can be permitted to bring the party to forego any right or advantage however slight. The law will not suffer the least weight to be put in the scale against him.68

[W]hen there is reason to believe that the plea has been entered through inadvertence . . . and mainly from the hope that the punishment to which the accused would otherwise be exposed may thereby be mitigated, the Court should be indulgent in permitting the plea to be withdrawn.69

Despite these early defeats for plea bargaining, the idea of bargained justice did not die. On the contrary, though infrequent by today’s standards, plea bargaining continued to exist in the local and district court systems.70

By the turn of the century, plea bargaining was on the rise, but not because it served mutually beneficial considerations of prosecutors and defendants or because it advanced judicial economy.71 Plea bargaining began to thrive in the early twentieth century because judges and prosecutors began accepting bribes from defendants in return for “plea agreements” that guaranteed reduced sentences. According to Professor Albert Alschuler, “[t]he gap between these judicial denunciations of plea bargaining [in the late nineteenth century] and the practices of many urban courts at the turn of the century and thereafter was apparently extreme. In these courts, striking political corruption apparently contributed to a flourishing practice of plea bargaining.”72

While corruption introduced plea bargaining to the broader legal community, it was the rise in criminal cases before and during Prohibition that spurred its growth and made it a legal necessity.73 Between the early twentieth century and 1925, the number of cases in the federal system resulting in pleas of guilty rose sharply from 50 to 90 percent.74 In return for defendants’ assistance in moving a

68 Id.
69 Id. In an 1877 Wisconsin Supreme Court decision, the court stated that plea bargaining was “hardly, if at all, distinguishable in principle from a direct sale of justice.” Id. at 21 (quoting Wight v. Rindskopf, 43 Wis. 344, 354 (1877)).
70 See id. at 22.
71 See id. at 24.
72 See id.
73 See id. at 27; GEORGE FISHER, PLEA BARGAINING’S TRIUMPH: A HISTORY OF PLEA BARGAINING IN AMERICA 210–12 (2003).
74 [F]ederal prosecutions under the Prohibition Act terminated in 1930 had become nearly eight times as many as the total of all pending federal prosecutions in 1914. In a number of urban districts the enforcement agencies maintain that the only practicable way of meeting this situation with the existing machinery of federal courts . . . is for the United States Attorneys to make bargains with defendants or their counsel whereby defendants plead guilty to minor offenses and escape with light penalties . . . .

Alschuler, supra note 65, at 32.
flood of cases through an overwhelmed system, they were often permitted to plead guilty to lesser charges or given lighter sentences.\(^\text{75}\) As Prohibition was extinguished, the United States continued its drive to create new criminal laws, a phenomenon that only added to the courts’ growing case loads and the pressure to continue to use bargaining to move cases through the system.\(^\text{76}\)

By 1967, the American Bar Association (ABA) was proclaiming the benefits of plea bargaining, even though it had not, as of yet, been specifically approved by the Supreme Court.\(^\text{77}\) The ABA stated:

\[
\text{[A] high proportion of pleas of guilty and nolo contendere does benefit the system. Such pleas tend to limit the trial process to deciding real disputes and, consequently, to reduce the need for funds and personnel. If the number of judges, courtrooms, court personnel and counsel for prosecution and defense were to be increased substantially, the funds necessary for such increases might be diverted from elsewhere in the criminal justice process. Moreover, the limited use of the trial process for those cases in which the defendant has grounds for contesting the matter of guilty aids in preserving the meaningfulness of the presumption of innocence.}\(^\text{78}\)
\]

By the time the Supreme Court agreed that plea bargaining was an available form of justice in 1970, plea bargaining’s rise to dominance was already complete. The *Brady* decision was delivered in the shadows of a force that led 90 percent of

\(^{74}\) See Alschuler, *supra* note 65, at 27 (“By 1925, the percentage of convictions by guilty plea had reached almost 90 . . . .”).

\(^{75}\) See id. at 28 (“The rewards associated with pleas of guilty were manifested not only in the lesser offenses of which guilty-plea defendants were convicted but also in the lighter sentences that they received.”).

\(^{76}\) See id. at 33 (“Dean Pound observed that ‘of one hundred thousand persons arrested in Chicago in 1912, more than one half were held for violation of legal precepts which did not exist twenty-five years before.’”); see also Donald A. Dripps, *Overcriminalization, Discretion, Waiver: A Survey of Possible Exit Strategies*, 109 Penn St. L. Rev. 1155, 1156–61 (2005) (discussing the relationship between broadening legal rules and plea bargaining); William J. Stuntz, *The Pathological Politics of Criminal Law*, 100 Mich. L. Rev. 505, 519–20 (2001) (discussing the influence of broader laws on the rate of plea bargaining); Ronald F. Wright, *Trial Distortion and the End of Innocence in Federal Criminal Justice*, 154 U. Pa. L. Rev. 79, 129 (2005) (“Changes in federal sentencing practices during the 1980s and 1990s increased the certainty and size of the penalty for going to trial, and mightily influenced the guilty plea and acquittal rates during those times.”).

\(^{77}\) See AM. BAR ASS’N PROJECT ON STANDARDS FOR CRIMINAL JUSTICE, STANDARDS RELATING TO PLEAS OF GUILTY 2 (1968) [hereinafter ABA PROJECT ON STANDARDS FOR CRIMINAL JUSTICE].

\(^{78}\) Id.
criminal defendants in the 1960s to waive their right to trial and confess their guilt in court.  

**B. The Force Behind Plea Bargaining’s Rise**

As demonstrated in the previous subpart, plea bargaining rose to dominate the American criminal justice system in the nineteenth and twentieth centuries. But how did members of the legal system succeed in making plea bargaining attractive to defendants? Today, a general consensus has evolved within plea bargaining scholarship that plea bargaining became a dominant force as a result of prosecutors gaining increasing power and control in an ever more complex criminal justice system. As prosecutors’ powers to both operate within and manipulate the system grew, their ability to create incentives for defendants to plead guilty also escalated. The key element of this machine, of course, is prosecutorial discretion and the ability to select from various criminal statutes with significantly different sentences.

---

79 Diana Borteck, *Pleas for DNA Testing: Why Lawmakers Should Amend State Post-Conviction DNA Testing Statutes to Apply to Prisoners Who Pled Guilty*, 25 Cardozo L. Rev. 1429, 1439 n.43 (2004) (citing Corinna Barrett Lain, *Accuracy Where It Matters: Brady v. Maryland in the Plea Bargaining Context*, 80 Wash. U. L.Q. 1, 1 (2002)) (noting “that since the 1960s the plea bargaining rate has been around ninety percent”); see ABA Project on Standards for Criminal Justice, supra note 77, at 1–2 (“The plea of guilty is probably the most frequent method of conviction in all jurisdictions; in some localities as many as 95 per cent of the criminal cases are disposed of in this way.”). Today, guilty pleas account for over 95 percent of all federal cases. See U.S. Sentencing Comm’n, supra note 44, at fig.C.

80 For instance, through analysis of plea bargaining in Massachusetts, Professor George Fisher argues that as the criminal justice system became more sophisticated, prosecutors gained the power to use selective charge bargaining to offer reduced sentences for those willing to negotiate. See Fisher, supra note 73, at 210 (“[Sentencing Guidelines] invest prosecutors with the power, moderated only by the risk of loss at trial, to dictate many sentences simply by choosing one set of charges over another.”); Alschuler, supra note 65, at 40; Albert W. Alschuler, *Plea Bargaining and Its History*, 13 Law & Soc’y Rev. 211 (1979); Dervan, supra note 64, at 478.

While prosecutors possessed the ability to control charging decisions and sentencing recommendations throughout the nineteenth and twentieth centuries, their power to control the criminal justice system and offer defendants deals increased throughout the 1900s. For example, as the number of criminal statutes grew during the early twentieth century, prosecutors had more choices when charging defendants and more discretion when selecting reduced charges in return for guilty pleas. A further example is the implementation of the United States Sentencing Guidelines in the last decade of the twentieth century, a tool that greatly increased prosecutors’ control of the system and increased their ability to force defendants into plea agreements.

Before the advent of modern sentencing guidelines, both prosecutor and judge held some power to bargain without the other’s cooperation . . . . Today, sentencing guidelines have recast whole chunks of the criminal code in the mold of the old Massachusetts liquor laws. By assigning a fixed and narrow penalty range to almost every definable offense, sentencing guidelines often empower prosecutors to dictate a defendant’s sentence by manipulating the charges. Guidelines have unsettled the old balance of bargaining power among prosecutor, judge, and defendant by ensuring that the prosecutor, who always had the strongest interest in plea bargaining, now has almost unilateral power to deal.

prosecutors may often choose from more than one statutory offense.”); Geraldine S. Moohr, _Prosecutorial Power in an Adversarial System: Lessons from Current White Collar Cases and the Inquisitorial Model_, 8 Buff. Crim. L. Rev. 165, 177 (2004) (“The power of the prosecutor to charge is two-fold; the power to indict or not . . . and the power to decide what offenses to charge.”).

82 See Fisher, supra note 73, at 16–17.

83 See Alschuler, supra note 65, at 32 (“Dean Pound observed that ‘of one hundred thousand persons arrested in Chicago in 1912, more than one half were held for violation of legal precepts which did not exist twenty-five years before.’”).

84 See Fisher, supra note 73, at 17.

85 Fisher, supra note 73, at 17; see also Boyd, supra note 81, at 591–92 (“While the main focus on the Sentencing Guidelines appeared to be narrowing judicial discretion in sentencing, some critics argued that the Sentencing Guidelines merely shifted the federal judge’s discretionary power to federal prosecutors.”); Marc L. Miller, _Domination & Dissatisfaction: Prosecutors as Sentencers_, 56 Stan. L. Rev. 1211, 1252 (2004) (“The overwhelming and dominant fact of the federal sentencing system, beyond the Commission and the guidelines and mandatory penalties, is the virtually absolute power the system has given prosecutors over federal prosecution and sentencing.”).
While defendants also play an important role in the plea bargaining process, prosecutors’ control of charging decisions and their influence over sentencing are key elements that contributed to the system’s dominance.86

Prosecutors have been successful in using their increased powers to create incentives that attract defendants to plead guilty by structuring plea agreements where the sentence a defendant receives in return for pleading guilty is far lower than the sentence he or she risks with a loss at trial.87 In a 1981 article on plea bargaining, Professor Albert Alschuler wrote of this “sentencing differential” and stated, “[c]riminal defendants today plead guilty in overwhelming numbers primarily because they perceive that this action is likely to lead to more lenient treatment than would follow conviction at trial. A number of studies suggest this perception is justified.”88 Among these studies was an examination by David Brereton and Jonathan Casper that analyzed robbery and burglary defendants in three California jurisdictions.89 The results were striking and illustrated that defendants who exercised their constitutional right to a trial received significantly higher sentences than those who worked with prosecutors to reach an agreement.90

86 See Fisher, supra note 73, at 230 (“[P]lea bargaining grew so entrenched in the halls of power that today, though its patrons may divide its spoils in different ways, it can grow no more. For plea bargaining has won.”).


88 Albert W. Alschuler, The Changing Plea Bargaining Debate, 69 CALIF. L. REV. 652, 652–56 (1981). Alschuler goes on to state, “Although the empirical evidence is not of one piece, the best conclusion probably is that in a great many cases the sentence differential in America assumes shocking proportions.” Id. at 654–56; see also Jenia I. Turner, Judicial Participation in Plea Negotiations: A Comparative View, 54 AM. J. COMP. L. 199, 251 (2006) (“While practitioners disagree about the acceptability of a large sentence differential at the post-plea and post-trial sentence, they agree that such a differential is common.”).


90 See Brereton & Casper, supra note 89, at 55–59; see also Daniel Givelber, Punishing Protestations of Innocence: Denying Responsibility and Its Consequences, 37 AM. CRIM. L. REV. 1363, 1382 (2000) (“The differential in sentencing between those who plead and those convicted after trial reflects the judgment that defendants who insist upon a
Not limiting themselves to a mere observation of sentencing trends, the researchers also made an insightful statement regarding the impact of high differentials on the rates of plea bargaining.

The point of the preceding discussion is simple enough: when guilty plea rates are high, expect to find differential sentencing. We believe that recent arguments to the effect that differentials are largely illusory do not withstand serious scrutiny, even though this revisionist challenge has been valuable in forcing us to examine more closely what is too often taken to be self-evidently true.91

Significant differentials, Brereton and Casper argued, are a tool used to increase plea bargaining rates by increasing the incentives for negotiation.92

Thus, plea bargaining’s rise to dominance during the nineteenth and twentieth centuries resulted from prosecutors gaining increased power over the criminal justice system and, through such power, the ability to offer increasingly significant incentives to those willing to confess their guilt in court. Today, sentencing differentials have reached new heights and, as a result, the incentives for defendants to plead guilty are greater than at any previous point in the history of our criminal justice system.93 It should be no surprise, therefore, that every year more than 95 percent of defendants accept the government’s offers of leniency and plead guilty rather than risk the consequences of failure at trial.94

---

91 Brereton & Casper, supra note 89, at 69.
92 See id. at 45 (“It is this sentence differential (whether conceived of as a reward to guilty pleaders or as a punishment of those who waste the court’s time by ‘needless’ trials) which has traditionally been seen as the engine driving the plea-bargaining assembly line.”); see also Givelber, supra note 90, at 1382 (“The pragmatic justification for differential sentencing is simple and powerful: we want those charged with crimes to plead guilty, and differential sentencing provides an accused with a strong incentive to do just that.”).
93 See Dervan, supra note 64, at 478–88. Along with increasing sentencing differentials, increasingly severe sentences have also increased the motivation for defendants to accept offers of leniency. See id.
94 See U.S. SENTENCING COMM’N, supra note 44, at fig.C.
II. THE SUPREME COURT’S GREAT COMPROMISE REGARDING PLEA BARGAINING

A. The Court’s Early Rejection of Bargained Justice

While plea-bargaining quickly grew in prominence within the trenches of the criminal justice system during the nineteenth and twentieth centuries, the Supreme Court was much slower and more deliberate with its consideration of this new form of justice. In fact, for much of the time period during which plea-bargaining was rising from obscurity to prominence, the Supreme Court considered this bargained justice unconstitutional.  

This, of course, did not prevent judges, prosecutors, and defendants from continuing to bargain in the shadows, a phenomenon which eventually became a significant reason for the Supreme Court’s sudden embrace of plea-bargaining in 1970.

The starting place for a historical analysis of the jurisprudence of plea-bargaining is actually within the law of confession. This is because up until the twentieth century, guilty pleas were simply considered a form of confession that occurred inside a courtroom. Professor Albert Alschuler addressed the link between the law of confession and guilty pleas in his 1979 work regarding the history of plea-bargaining. 

While the legal phenomenon that we call a guilty plea has existed for more than eight centuries, the term ‘guilty plea’ came into common use only about one century ago. During the previous 700 years, what we call a guilty plea was simply called a ‘confession.’

The most noteworthy pre-American common law case regarding confession is the 1783 English case of *Rex v. Warickshall*. In this case, the English court stated that confessions are improper and inadmissible where they are induced by “promises of favor.” The court continued, “[W]hile the legal phenomenon that we call a guilty plea has existed for more than eight centuries, the term ‘guilty plea’ came into common use only about one century ago. During the previous 700 years, what we call a guilty plea was simply called a ‘confession.’”

The most noteworthy pre-American common law case regarding confession is the 1783 English case of *Rex v. Warickshall*. In this case, the English court stated that confessions are improper and inadmissible where they are induced by “promises of favor.” The court continued, “[A] confession forced from the mind by the flattery of hope, or by the torture of fear, comes in so questionable a shape

---

95 See infra text accompanying notes 124–197.
96 See infra text accompanying notes 136–138, 221–236.
97 See infra text accompanying notes 13–13, 221–236.
98 See Alschuler, supra note 65, at 12. Professor Alschuler’s article contains a gripping account of the link between the law of confession and the law of guilty pleas. Although I will examine many of the same cases, this Article will focus more squarely on the issue of voluntariness as that term has evolved throughout the nineteenth and twentieth centuries. See infra Part III (discerning the current meaning of the term “voluntary” through consideration of the history of plea bargaining case law as discussed in Part II.A.).
99 See Alschuler, supra note 65, at 12.
100 Id. at 12–13.
102 Id. at 234, 1 Leach at 263.
... that no credit ought to be given to it.” Thus, the concept that confessions and guilty pleas must be voluntary was born.

It was not until 1892 that the United States Supreme Court affirmed a guilty plea by a defendant in the matter of *Hallinger v. Davis*, wherein the Supreme Court clarified that the high standard for the admissibility of confessions utilized in England was equally applicable to a guilty plea in America. Hallinger was indicted by a grand jury in New Jersey for murder. Shortly thereafter, he entered a guilty plea to the charges. The lower court was so uncomfortable with the idea of a defendant waiving his or her right to trial, however, that it held the plea in abeyance and required the defendant to consult with counsel. Following the consultation, Hallinger again requested to plead guilty. After hearing testimony regarding the basis for the defendant’s plea, the court accepted the confession and sentenced Hallinger to death. In considering the case, the United States Supreme Court concluded that the defendant had “voluntarily” availed himself of the option to plead guilty and, while perhaps unwise, the decision did not violate any provisions of the Constitution.

---

103 Id. at 235, 1 Leach at 264.
104 See Alswhicher, supra note 65, at 12 (discussing Rex, 168 Eng. Rep. at 255, 1 Leach at 263–64). “It soon became clear that any confession ‘obtained by [a] direct or implied promise[,] however slight’ could not be received in evidence. Even the offer of a glass of gin was a ‘promise of leniency’ capable of coercing a confession.” Id. (citations omitted); see also Baldwin & McConville, supra note 64, at 287–91 (discussing the use of plea bargaining in England).
105 *Hallinger v. Davis*, 146 U.S. 314, 324 (1892).
106 See id. at 318–20. One of the earliest American cases discussing the need for guilty pleas to be voluntary was an 1804 Massachusetts Supreme Court case titled *Commonwealth v. Battis*, 1 Mass. 95, 95 (1804). Battis, an African American, was indicted in Massachusetts for the rape and murder of a thirteen-year-old Caucasian girl. Id. Upon being brought to court to answer the charges, Battis pleaded guilty. Id. Surprised by the defendant’s actions, the court remanded Battis to prison to reconsider his decision. Id. at 95–96. Later that day, Battis was brought back before the judge and again pleaded guilty. Id. at 96. After thorough questioning, it was determined that there had been no “tampering with him, either by promises, persuasions, or hopes of pardon, if he would plead guilty.” Id. The court accepted the plea and Battis was sentenced to death. Id. The *Battis* case provides a glimpse into criminal practice in the nineteenth century and demonstrates that courts were reluctant to permit defendants to plead guilty. See id. Further, when such guilty pleas were permitted, they were only allowed where it was clear the defendant had been offered no promises and had endured no threats in deciding to confess his or her crime. See id.
107 *Hallinger*, 146 U.S. at 315 (“Hallinger, the appellant, was on the 14th day of April, 1891, indicted by the grand jury of Hudson County for the murder of one Mary Hallinger.”).
108 Id.
109 Id.
110 Id.
111 Id. (“[H]e was condemned to be hanged.”).
112 Id. at 324.
Court concluded that pleas of guilty must be “voluntary,” a legal doctrine drawn directly from the English law of confession.113

Though the Supreme Court failed to define the term “voluntary” with any precision in Hallinger, it was not necessary. At the time, one needed only examine case law regarding the definition of the term “voluntary” in the context of confessions, because, as described above, guilty pleas were simply in-court confessions in the eyes of the law.114 One such case was the 1897 case of Bram v. United States.115 Bram was accused of murdering a ship’s captain and others.116 During his trial, the prosecution admitted a confession Bram made to police, though Bram protested that the confession had been extracted from him involuntarily.117 In ruling on the matter, the Supreme Court referred to the long history of case law establishing that one may not be compelled to testify against himself.118

But a confession, in order to be admissible, must be free and voluntary: that is, must not be extracted by any sort of threats of violence, nor obtained by any direct or implied promises, however slight, nor by the exertion of any improper influence... for the law cannot measure the force of the influence used, or decide upon its effect upon the mind of the prisoner....119

According to the Court, the “true test of admissibility is that the confession is made freely, voluntarily and without compulsion or inducement of any sort.”120 In the Bram case, the detective was found to have held out hope of leniency in return for Bram disclosing the identity of his accomplice.121 The Court concluded that

113 See id.; see also Alschuler, supra note 65, at 10 (noting that this was also the first U.S. case to uphold a guilty plea).
114 See supra notes 99–102 and accompanying text.
116 Id. at 537.
117 Id. at 538–39.
118 See id. at 545 (“There can be no doubt that long prior to our independence the doctrine that one accused of crime could not be compelled to testify against himself had reached its full development in common law...”); see also U.S. CONST. amend. V.
119 Bram, 168 U.S. at 542–43 (quoting 3 RUSSELL ON CRIMES 478 (6th ed.)).
120 Id. at 548 (quoting Wilson v. United States, 162 U.S. 613, 623 (1896)). Citing to Rex v. Thompson, (1783) 1 Leach 291, the Court stated:

In Rex v. Thompson... a declaration to a suspected person that unless he gave more satisfactory account of his connection with a stolen bank note his interrogator would take him before a magistrate, was held equivalent to stating that it would be better to confess, and to have operated to lead the prisoner to believe that he would not be taken before a magistrate if he confessed.

121 Id. at 551–52.
Bram’s statements were involuntary, because they were induced by a promise of benefits. Based on the rulings in Hallinger and Bram, one can deduce that, had a plea bargaining case risen to the Supreme Court in the late 1800s, the Court would have clearly opposed the idea, just as so many state appellate courts did in the post-Civil War period.

In 1878, the Supreme Court took up a series of cases termed the Whiskey Cases. The Court’s ruling in the Whiskey Cases dealt more with prosecutorial authority to offer deals than with the involuntary nature of an induced confession, but the cases offer confirmation of the Supreme Court’s antagonist view of bargained justice at a time when plea bargaining was just beginning its rise in the American criminal justice system. The Whiskey Cases involved prosecutions for violations of internal revenue laws regarding liquor sales. Perhaps in an effort to reward cooperation or clear a number of matters from his desk at one time, the prosecutor offered the defendants a deal in return for pleading guilty.

[The parties] entered into an agreement by which it was, among other things, agreed that if the said defendants would testify on behalf of the plaintiffs frankly and truthfully . . . , and should plead guilty to one count in an indictment . . . , and should withdraw their pleas in a certain condemnation case then pending against them . . . , no new proceedings should be commenced against said defendants on account of transactions then past.

At the time, it was not uncommon for defendants to receive immunity from the governor in return for testifying against an accomplice. The bargain here, however, went further in that the agreement did not emanate from the governor but directly from the prosecutor. Further, the agreement did not merely offer immunity in return for cooperation, but also required the defendants to plead guilty to a single charge of the indictment in return for the other two charges being dropped. Regarding this effort by the prosecution to induce the defendants to

---

122 Id.
123 See supra notes 64–70 and accompanying text (discussing the rise of plea bargaining in the post-Civil War period).
125 See supra notes 121–128 and accompanying text; see also Sanders v. State, 85 Ind. 318, 333 (1882) (holding that a guilty plea must be voluntary, and finding that where a defendant pleaded guilty for fear of mob violence, the plea was not voluntary).
126 See Whisky Cases, 99 U.S. at 594–95.
127 Id.
128 See Alschuler, supra note 65, at 15 (“[W]hen defendants were induced to testify against their accomplices, Anglo-American courts refused to convict them on the basis of their bargained confessions. The courts instead insisted that these defendants be given what modern lawyers would call transactional immunity.”).
129 See Whisky Cases, 99 U.S. at 598.
130 See id.
plead guilty, the Supreme Court stated, “[T]he district attorney had no authority to make the agreement alleged in the plea in bar,” and, as such, the bargain was unenforceable. 131 While the Whiskey Cases decision did not refer to the defendants’ guilty pleas as “involuntary,” the Court’s concern over this perception likely influenced its decision to prohibit such deal making.

As the Rex, Hallinger, Bram and Whiskey cases demonstrate, plea-bargaining was impermissible in the eyes of the Supreme Court during the late nineteenth century.132

However, while the Supreme Court and various state courts were succinctly rejecting attempts by prosecutors to offer inducements to defendants in return for guilty pleas, plea-bargaining continued to be utilized in the trenches. 133 It is not surprising that more cases did not reach the Supreme Court to challenge this growing phenomenon because defendants who were satisfied with their bargain had no reason to appeal their convictions. 134 Gradually, however, appellate cases began appearing in the federal system and, by the early twentieth century, the Court was once again addressing this ever-growing phenomenon.

In 1927, the Supreme Court was tasked with addressing the constitutionality of prosecutors offering incentives for defendants to plead guilty in the case of Kercheval v. United States.135 Kercheval pleaded guilty to a charge of using the mail to defraud and was sentenced to three years in prison.136 According to the defendant, he was induced to plead guilty because the prosecutor promised to

131 Id. at 606.
132 One legal annotation from the period summarized the law as follows:

We would conclude, from an examination of all the cases upon the subject, that where there is an inducement of any kind held out to the prisoner, by reason of which he enters the plea of guilty, it will . . . better comport with a sound judicial discretion to allow the plea to be withdrawn . . . , and especially so when counsel and friends represent to the accused that it has been the custom and practice of the court to assess a punishment less than the maximum upon such a plea . . . .

Alschuler, supra note 65, at 24 (quoting Hopkins, Withdrawal of Plea of Guilty, 11 CRIM. L. MAG. 479, 484 (1889)).

133 See supra notes 97–135 and accompanying text.

134 In 1968, the American Bar Association released a report that included discussion of the shadowy nature of plea bargaining up to that point in the American criminal system. ABA PROJECT ON STANDARDS FOR CRIMINAL JUSTICE, supra note 77, at vii (“The subject of this report—pleas of guilty, including the related matter of plea negotiations—was given a high priority by the Special Committee. One of the purposes of the Project was to turn a spotlight on what have been called ‘low visibility’ areas in the criminal process, those not likely to be scrutinized by the appellate courts and to be discussed in widely-printed opinions. This is such an area, and its importance is indicated by the fact that the great majority of all convictions are the result of pleas of guilty.”).

136 Id.
recommend a sentence of three months in jail.\textsuperscript{137} After being sentenced to three years in prison, Kercheval filed a petition seeking to enforce the prosecutor’s promise.\textsuperscript{138} The court declined, but did allow him to withdraw his guilty plea.\textsuperscript{139} At trial, however, the court permitted the prosecution to introduce the prior guilty plea as evidence against the defendant.\textsuperscript{140} As might be expected, Kercheval was convicted and once again sentenced to three years in prison.\textsuperscript{141}

On appeal to the Supreme Court, the Justices concluded that Kercheval’s withdrawn guilty plea was inadmissible.\textsuperscript{142} More importantly, however, the Court commented on the plea bargain between the government and the defendant.\textsuperscript{143}

A plea of guilty differs in purpose and effect from a mere admission or an extrajudicial confession; it is itself a conviction. Like a verdict of a jury it is conclusive. More is not required; the court has nothing to do but give judgment and sentence. Out of just consideration for persons accused of crime, courts are careful that a plea of guilty shall not be accepted unless made voluntarily after proper advice and with full understanding of the consequences. . . . But, on timely application, the court will vacate a plea of guilty shown to have been unfairly obtained or given through ignorance, fear or inadvertence.\textsuperscript{144}

Rather than positing that prosecutors are afforded more leeway to bargain for guilty pleas than might be permitted in the context of confessions to police officers, the Court indicated that even greater care should be taken to ensure guilty pleas are voluntary.\textsuperscript{145} Once again, the Court reiterated the long-standing principle that guilty pleas must be voluntary and may not be induced by “ignorance, fear or inadvertence.”\textsuperscript{146}

As the criminal dockets swelled during the early twentieth century and prosecutors began to offer plea bargains in greater numbers, the Supreme Court remained on the sidelines. It was not until 1941 that the Court again addressed the issue of bargained justice, this time in its most direct manner yet.\textsuperscript{147} In 1936, Jack Walker pleaded guilty to an indictment charging him with armed robbery of a

\textsuperscript{137} Id. During the negotiations, the prosecutor also asserted that the court would impose the recommended sentence. Id.
\textsuperscript{138} See id.
\textsuperscript{139} Id. (“After hearing evidence on the issue, the court declined so to change the sentence, but on petitioner’s motion, set aside the judgment and allowed him to withdraw his plea of guilty and to plead not guilty.”).
\textsuperscript{140} Id. at 221–22.
\textsuperscript{141} Id. at 222.
\textsuperscript{142} See id. at 223, 225.
\textsuperscript{143} See id. at 223–24.
\textsuperscript{144} Id.
\textsuperscript{145} See id.
\textsuperscript{146} Id.
\textsuperscript{147} See Walker v. Johnston, 312 U.S. 275 (1941).
national bank. After being sentenced to twelve years in prison, Walker filed a writ of habeas corpus. In his writ, he alleged that he had been induced to plead guilty by promises of leniency and threats of harsher sanctions for a failure to cooperate. According to the Court:

[The District Attorney] told him to plead guilty, warning him that he would be sentenced to twice as great a term if he did not so plead . . . . In view of the District Attorney’s warning, and in fear of a heavy prison term, he told the District Attorney he would plead guilty.

While these arrangements seem common by today’s standards, the Supreme Court found the prosecutor’s actions to be improper in 1941. The Court concluded that prosecutor’s threats had resulted in Walker involuntarily pleading guilty, a clear violation of the standard established in earlier precedent regarding voluntariness, confession, and pleas of guilt. The Court remarked:

[Walker] was deceived and coerced into pleading guilty when his real desire was to plead not guilty or at least to be advised by counsel as to his course. If he did not voluntarily waive his right to counsel, or if he was deceived or coerced by the prosecutor into entering a guilty plea, he was deprived of a constitutional right.

As a result, the Supreme Court remanded the case for a hearing regarding Walker’s allegations. While the language of Walker should have given prosecutors pause before continuing to offer defendants plea bargains, the rising tide of plea-bargaining and the courts’ growing reliance on such swift justice was likely already unstoppable.

A year later, the Supreme Court again addressed the standards applicable to guilty pleas. In Waley v. Johnston, the Supreme Court examined allegations that Federal Bureau of Investigation (FBI) agents had physically threatened a defendant and had threatened to “publish false statements and manufacture false evidence that the kidnapped person had been injured, and by such publications and false evidence to incite the public and to cause the State of Washington to hang the petitioner . . . .” According to Waley, the government’s threats that he would be

148 Id. at 279.
149 Id.
150 Id. at 280–81.
151 Id. at 280.
152 See id. at 286–87.
153 See id. at 286; see also Hallinger v. Davis, 146 U.S. 314, 324 (1892) (requiring that the defendant voluntarily avail himself of the option to plead guilty).
154 Walker, 312 U.S. at 286.
155 Id. at 287.
157 Id. at 102.
thrown out a window and beaten were unpersuasive. However, the threats that false evidence would be offered to increase his punishment to death were compelling and induced him to plead guilty as demanded.

In ordering the lower court to conduct a hearing regarding Waley’s writ of habeas corpus, the Supreme Court concluded that while the allegations may “tax credulity,” they certainly called for further inquiry. In describing the applicable considerations during the lower court hearing, the Supreme Court reiterated the requirement that guilty pleas be entered voluntarily. According to the Court, “[i]f the allegations are found to be true, petitioner’s constitutional rights were infringed. For a conviction on a plea of guilty coerced by a federal law enforcement officer is no more consistent with due process than a conviction supported by a coerced confession.” As late as 1940, therefore, the voluntariness standard as applied uniformly to guilty pleas and confessions remained the same and coercive tactics and incentives rendered any such admissions unconstitutional.

In 1958, the issue of plea bargaining again reached the Supreme Court, but, on this occasion, the case was never decided for fear the result would throw the entire American criminal justice system into chaos. The case for consideration was that of Paul Shelton, who, despite consistently maintaining his innocence, pleaded guilty to one count of interstate transportation of stolen vehicles after being promised leniency. The prosecution had induced Shelton to confess his guilt by promising a sentence of no more than one year and dismissal of all other pending charges. Though these promises sound similar to those made to Lea

---

158 See id. (“The petitioner stated generally that threats of . . . agents to throw petitioner out of a window and ‘beat [him] up’ ‘didn’t bother [him].’”).
159 Id. (“But [the petition] specifically alleged that petitioner’s plea of guilty had been induced by the threats of a named [FBI] agent to publish false statements and manufacture false evidence that the kidnapped person had been injured . . . .”).
160 See id. at 104.
161 See id.
162 Id.
163 See id.
165 Shelton, 242 F.2d at 102.
166 Id. at 102 & n.1. According to the opinion, Shelton was induced to plead guilty based on the following promises:

(1) government counsel’s promise to arrange for dismissal of all other federal charges, (2) government’s counsel’s ‘guarantee’ of a sentence of not more than one year in the instant case, (3) government’s counsel’s threat that it would take defendant longer to get tried on the other pending charges than it would take him to serve a one year sentence, and (4) defendant’s confused, anxious, desperate, and incompetent state of mind resulting from prolonged confinement in county jails while awaiting trial as aforesaid.
Fastow in 2004, the United States Court of Appeals for the Fifth Circuit concluded that Shelton’s plea must be vacated because the prosecutor’s promises of leniency likely meant the plea was not entered “voluntarily.”

There is no doubt, indeed it is practically conceded, that the appellant pleaded guilty in reliance on the promise of the Assistant United States Attorney that he would receive a sentence of only one year. The court, before accepting the plea, did not ascertain that it was in truth and in fact a voluntary plea not induced by such promise. It necessarily follows that the judgment of conviction must be set aside and the plea of guilty vacated.

In the opinion of the judges on the Fifth Circuit in 1957: “Justice and liberty are not the subjects of bargaining and barter.”

Interestingly, the panel decision from the Fifth Circuit was later overturned en banc, and the case proceeded to the Supreme Court. The Court never addressed the challenge to plea-bargaining, however, because the government filed an admission that the guilty plea may have been improperly obtained and the case was remanded to the District Court without further discussion. According to Professor Albert Alschuler, evidence indicates that the government likely confessed its error for fear that the Supreme Court would finally make a direct ruling that all manner of plea bargaining was wholly unconstitutional.

By 1963, when the Supreme Court ruled in the Lynumn case that her confession was involuntarily coerced due to threats regarding her children and her potential prison sentence, one other significant plea bargaining case had come before the Court. In 1962, the court considered whether it was appropriate for a

Id.

See id. at 113 (“If a plea of guilty is made upon any understanding or agreement as to the punishment to be recommended, it is essential, we think, that, before accepting such plea, the district court should make certain that the plea is in fact made voluntarily.”).

Id.

Id.

Shelton v. United States, 246 F.2d 571, 573 (5th Cir. 1957) (en banc).

See Shelton v. United States, 356 U.S. 26, 26 (1958) (“The motion for leave to proceed in forma pauperis and the petition for writ of certiorari are granted. Upon consideration of the entire record and confession of error by the Solicitor General that the plea of guilty may have been improperly obtained, the judgment of the United States Court of Appeals for the Fifth Circuit is reversed and the case is remanded to the District Court for further proceedings.”).

Alschuler, supra note 65, at 37 (“In 1958, the Solicitor General (or perhaps some other official in the Justice Department) may have assessed the probable votes of individual Supreme Court Justices, may have sensed a substantial likelihood that the Court would hold the practice of plea bargaining unlawful, and may have sought to foreclose this ruling through a confession of error on narrow and disingenuous grounds.”).

United States Attorney to promise a defendant that in return for pleading guilty his sentence would be no more than twenty years in prison.\textsuperscript{174} In the alternative, the prosecutor stated that if the defendant did not cooperate, “certain unsettled matters concerning two other robberies would be added to the petitioner’s difficulties.”\textsuperscript{175} The Supreme Court remanded the *Machibroda* case for further fact finding, but concluded that if the allegations were true, the petitioner was “clearly entitled to relief.”\textsuperscript{176}

There can be no doubt that, if the allegations contained in the petitioner’s motion and affidavit are true, he is entitled to have his sentence vacated. A guilty plea, if induced by promises or threats which deprive it of the character of a voluntary act, is void. . . . Like a verdict of a jury [a plea of guilty] is conclusive. . . . Out of just consideration for persons accused of crime, courts are careful that a plea of guilty shall not be accepted unless made voluntarily after proper advice and with full understanding of the consequences.\textsuperscript{177}

With such strong language regarding the requirement that plea bargains be voluntary and the historical context that guilty pleas and confessions were treated in the same manner, it is little surprise that a year later the Supreme Court struck down Beatrice Lynumm’s confession as a statement made after threats of punishment and promises of leniency.\textsuperscript{178}

In 1968, the Supreme Court once again struck a blow to plea-bargaining, though this did not stop the acceleration of its growth. In *United States v. Jackson*,\textsuperscript{179} the Court examined a federal statute that differentiated between the punishment available to those who pleaded guilty and those who put the government to its burden.\textsuperscript{180} The law, 18 U.S.C. section 1201(a), reads as follows:

> Whoever knowingly transports in interstate . . . commerce, any person who had been unlawfully . . . kidnaped and held for ransom . . . or otherwise . . . shall be punished (1) by death if the kidnapped person has not been liberated unharmed, and if the verdict of the jury shall so recommend, or (2) by imprisonment for any term of years or for life, if the death penalty is not imposed.\textsuperscript{181}

As illustrated in the statute’s text above, the death penalty was only available where a jury convicted the defendant and recommended capital punishment.

\textsuperscript{174} *Id.* at 489–90.
\textsuperscript{175} *Id.*
\textsuperscript{176} *Id.* at 493, 496.
\textsuperscript{177} *Id.* at 493 (internal quotation marks omitted).
\textsuperscript{180} *Id.* at 570–72.
\textsuperscript{181} *Id.* at 570–71.
Where the defendant pleaded guilty, however, the judge was limited to imposing a sentence of life in prison. In October 1966, a federal grand jury in Connecticut charged Charles Jackson and two others with a kidnapping that resulted in an injury to the victim. The District Court, however, struck the indictment because it found that the statute unconstitutionally made the “risk of death” the “price for asserting the right to jury trial.”

On appeal to the Supreme Court, a majority of the Justices agreed that the statute imposed an “impermissible burden upon the exercise of a constitutional right.”

It is no answer . . . that federal trial judges may be relied upon to reject coerced pleas of guilty and involuntary waivers of jury trials. For the evil in the federal statute is not that it necessarily coerces guilty pleas and jury waivers but simply that it needlessly encourages them. A procedure need not be inherently coercive in order that it be held to impose an impermissible burden upon the assertion of a constitutional right. Thus the fact that the [statute] tends to discourage defendants from insisting upon their innocence and demanding trial by jury hardly implies that every defendant who enters a guilty plea to a charge under the [statute] does so involuntarily. The power to reject coerced guilty pleas and involuntary jury waivers might alleviate, but it cannot totally eliminate, the constitutional infirmity in the capital punishment provision of the [statute].

This is perhaps one of the most significant statements ever made about plea-bargaining by the Supreme Court. According to the Court, and consistent with earlier precedent, voluntary guilty pleas are permissible. In fact, the Court states that the power of the lower court’s to accept a guilty plea is traditional and “necessary for the . . . practical . . . administration of the criminal law.” The Court offers as an example of a voluntary plea motivated by constitutional considerations the case where an individual pleads guilty to “spare themselves and their families the spectacle and expense of protracted courtroom proceedings.” The Court implies, however, that guilty pleas induced by coercive threats of punishment or offers of leniency are unconstitutional. With regard to the federal kidnapping statute, the threat of death only for those who refuse to confess their

---

182 See id.
183 Id. at 571 & n.1.
184 Id. at 571.
185 Id. at 572. The Court, however, disagreed that the statute as a whole was unconstitutional and, instead, severed the provision from the remainder of the statute. Id.
186 Id. at 583.
187 Id. at 584–85.
188 Id.
189 Id. at 584.
190 See id. at 583.
guilt is an example of a coercive incentive that makes any resulting guilty plea invalid.\textsuperscript{191}

In a dissenting opinion, Justice White argued that the kidnapping statute should not be held unconstitutional.\textsuperscript{192} Relying on the majority’s conclusion that “not every plea of guilty or waiver of the jury trial would be influenced by the power of the jury to impose the death penalty,” Justice White stated that trial courts need only carefully examine a defendant’s guilty plea.\textsuperscript{193} In so doing, the lower court can make an individual determination “to make sure that they have been neither coerced nor encouraged by the death penalty power in the jury.”\textsuperscript{194} For the majority, however, there was no legitimate justification for the imposition of death upon only those who seek a trial and, therefore, while Justice White’s approach may be appropriate in most cases, it was not sufficient to correct the unconstitutional construction of this particular statute.\textsuperscript{195}

By 1968, the Supreme Court had struck down every guilty plea induced by threats of punishment or promises of leniency that had arrived on its docket.\textsuperscript{196} The necessity of the plea bargaining machine, however, had never been greater. As the implementation of the procedural rights given to defendants during the Due Process Revolution slowed trial proceedings and criminal dockets continued to swell,\textsuperscript{197} prosecutors had no choice but to continue to utilize a constitutionally questionable system. Finally, in 1970, the Supreme Court directly addressed the issue of plea-bargaining’s constitutionality, but the result was stunning in that it was inconsistent with over a century of precedent and the Court’s prior animosity towards bargained justice.

\textsuperscript{191} Id.
\textsuperscript{192} Id. at 592 (White, J., dissenting).
\textsuperscript{193} Id.
\textsuperscript{194} Id.
\textsuperscript{195} Id. at 583. Following Jackson, the Supreme Court ruled in the 1969 case of McCarthy v. United States that federal courts must personally address the defendant to satisfy the requirements of Rule 11. McCarthy v. United States, 394 U.S. 459, 463–64 (1969). Interestingly, in McCarthy, one sees that even in the late 1960s, prosecutors were weary of creating plea agreements that contained direct promises of leniency. See id. at 461–62. In McCarthy, the defendant agreed to plead guilty to one charge of tax evasion. Id. at 461. After the guilty plea was accepted, the prosecutor moved to dismiss two other pending tax charges. Id. at 461. In so doing, the prosecutor made clear to the court that the dismissal was primarily motivated by the defendant’s agreement to pay all taxes, penalties, and interest with regard to any remaining tax burden. See id. at 461–62. In fact, the prosecutor went on to insist that the judge ask the defendant if he had been induced by any threats or promises. Id. at 461. The defendant stated that he had not and, therefore, the lower court concluded the plea was voluntary. See id.; see also Boykin v. Alabama, 395 U.S. 238, 242–43 (1969) (ruling that Rule 11 requirements also apply to state court proceedings and stating that “[i]gnorance, incomprehension, coercion, terror, inducement, subtle or blatant threats might be a perfect cover-up of unconstitutionality”).
\textsuperscript{196} See supra notes 115–199 and accompanying text.
\textsuperscript{197} See infra notes 230–233 and accompanying text.
B. The Supreme Court’s Great Compromise of 1970

In 1970, a case came before the Supreme Court that would help usher in the final stages of plea-bargaining’s triumph. In 1959, Robert Brady was charged with kidnapping in violation of 18 U.S.C. § 1201(a), the same criminal statute that had been at issue in the Jackson case in 1968.198 As previously discussed, the statute allowed for the death penalty only when recommended by a jury.199 As a result, a defendant who pleaded guilty and avoided a jury trial could successfully avoid the death penalty in favor of the alternative maximum sentence of life in prison.200 The victim in the Brady case had not been liberated unharmed and, therefore, Brady was subject to the death penalty provision if recommended by a jury.201 At first, Brady pleaded not guilty and intended to proceed to trial.202 Brady later learned that his co-defendant had agreed to plead guilty and intended to testify against him at trial.203 As such, Brady changed his plea to guilty and was sentenced to fifty years in prison.204 Brady later sought relief from the Supreme Court, claiming that his plea was involuntary because the statute coerced him into avoiding a jury trial to ensure that he would not be sentenced to death.205 In examining Brady’s claim, the lower court concluded, “no representations [were] made with respect to a reduced sentence or clemency.”206 Thus, the lower court held the guilty plea voluntary.207

The Supreme Court began its analysis by noting that guilty pleas must be voluntary and intelligent, a determination that must be conducted based on the facts of the case.208 The first question examined by the Court, therefore, was whether Brady’s plea was involuntary because he faced a more severe sentence if found guilty after trial.209 The Court concluded that just because the enhanced penalty led to Brady’s decision, it does not follow that the plea was necessarily involuntary.210 Rather, Brady’s decision was similar to one made by a defendant who pleads guilty in the face of overwhelming evidence rather than subjecting his

199 See id.
200 See id.
201 Id.
202 Id.
203 Id.
204 Id. at 743–44.
205 Id. at 744.
206 Id. at 745.
207 Id.
208 Id. at 748.
209 Id. at 749.
210 Id. at 749–50 (“But even if we assume that Brady would not have pleaded guilty except for the death penalty provision of § 1201(a), this assumption merely identifies the penalty provision as a "but for" cause of his plea. That the statute caused the plea in this sense does not necessarily prove that the plea was coerced and invalid as an involuntary act.”).
or her family to the expense of trial. The Court continued, however, with a caveat regarding coercive tactics by the prosecution.

Of course, the agents of the State may not produce a plea by actual or threatened physical harm or by mental coercion overbearing the will of the defendant. But nothing of the sort is claimed in this case; nor is there evidence that Brady was so gripped by fear of the death penalty or hope of leniency that he did not or could not, with the help of counsel, rationally weigh the advantages of going to trial against the advantages of pleading guilty.212

According to the Court, where the defendant pleads guilty in return for the “certainty or probability of a lesser penalty [as contained in the statute] rather than face the wider range of possibilities extending from acquittal to conviction and a higher penalty authorized by law for the crime charged,” the plea is voluntary and constitutionally sound.213 The Court concluded that a guilty plea to avoid a possible death penalty under the applicable statute is valid.214

Based on its statements that a defendant may plead guilty in return for the promise or possibility of a reduced sentence, the Supreme Court appears to have adopted language crafted by Judge Tuttle of the United States Court of Appeals for the Fifth Circuit to describe the applicable test for “voluntariness” moving forward.215

[A] plea of guilty entered by one fully aware of the direct consequences, including the actual value of any commitments made to him by the court, prosecutor, or his own counsel, must stand unless induced by threats (or promises to discontinue improper harassment), misrepresentation (including unfulfilled or unfulfillable promises), or perhaps by promises

211 Id. at 750 (“The State to some degree encourages pleas of guilty at every important step in the criminal process. For some people, their breach of a State’s law is alone sufficient reason for surrendering themselves and accepting punishment. For others, apprehension and charge, both threatening acts by the Government, jar them into admitting their guilt. In still other cases, the post-indictment accumulation of evidence may convince the defendant and his counsel that a trial is not worth the agony and expense to the defendant and his family. All these pleas of guilty are valid in spite of the State’s responsibility for some of the factors motivating the pleas . . . .”).

212 Id. at 750.

213 Id. at 751.

214 Id. at 755.

215 Id. at 755. The Court attempts to explain the apparent inconsistency between its decision in Brady and the prior precedent of Bram. According to the Court, the decision in Bram is distinguishable because there was no counsel present, the threats and promises were made face-to-face, and there was a hazard of an impulsive and improvident response to a seeming but unreal advantage. Id. at 754–55.
that are by their nature improper as having no proper relationship to the prosecutor’s business (e.g. bribes)."216

By the end of 1970, therefore, confession law and the law of guilty pleas had diverged significantly. No longer was any promise of leniency thought to corrupt a guilty plea as the Court had ruled throughout the nineteenth and early twentieth centuries. Though the plea must still be “voluntary,” the term had shifted to mean merely that the plea could not be induced “by actual or threatened physical harm or by mental coercion overbearing the will of the defendant.”217 A dramatic shift had occurred in the law, and, although there was still an upper limit beyond which inducements to plead guilty could not venture, that limit was much higher than it had been just a few short years before.218 Plea-bargaining, it seemed, had finally established itself as an unseverable part of the American criminal justice system.

It is clear that the Supreme Court radically departed from its previous decisions to craft its decision in \textit{Brady}, but it is unclear why the Court felt it necessary to make such a significant departure from past precedent. The answer, it appears, has much to do with plea-bargaining’s rise to dominance by 1970 and the realities of an overburdened criminal justice system.

Despite the hostility expressed towards bargained justice by the Supreme Court prior to the \textit{Brady} decision in 1970, the plea-bargaining machine continued to evolve and to increase in influence. Early records indicate that in the first half of the nineteenth century only 10 to 15 percent of cases resulted in a guilty plea.219 While plea-bargaining may have occurred in certain of these cases, it is likely many were merely cases in which the defendant did not wish to endure trial when the evidence appeared overwhelming, such as occurred in the \textit{Hallinger} case in 1892.220 Shortly thereafter, however, the rate of convictions by guilty plea began an upward climb.221 By 1908, 50 percent of federal criminal cases resulted in pleas

\footnotesize{216 Id. at 755. Interestingly, this test appears to be plucked from the \textit{en banc} decision in \textit{Shelton v. United States}, the case the solicitor general removed from the Supreme Court’s jurisprudence by offering a confession of error for fear the Supreme Court as composed in 1958 would strike plea bargaining as unconstitutional in its entirety. \textit{Shelton v. United States}, 246 F.2d 571, 572 n.2 (5th Cir. 1957) (en banc), rev’d per curiam on confession of error, 356 U.S. 26 (1958).

217 See \textit{Brady}, 397 U.S. at 750–51; see also Becker, supra note 100, at 759 (“Although the Court agreed that guilty pleas were invalid if not ‘voluntary,’ its treatment of voluntariness cut that concept loose from its moorings in the law.”).

218 See \textit{Brady}, 397 U.S. at 757–58; see also Becker, supra note 100, at 759 (“In \textit{Brady v. United States} and its companion cases, however, the Court denied the relevance not merely of some but all prior doctrine.”).

219 See Alschuler, supra note 65, at 10 (noting that Boston Police Court records from 1824 indicated that only 11 percent of defendants entered pleas of guilty); id. at 18 (noting that in 1839, 15% of all defendants in Manhattan and Brooklyn pleaded guilty).

220 \textit{Hallinger} v. Davis, 146 U.S. 314 (1892).

221 See Alschuler, supra note 65, at 27 (“In the federal courts, the statistics date from 1908, when only about 50% of all convictions were by plea of guilty. This percentage remained fairly constant until 1916, when it increased to 72%.”).}
of guilty, and by 1916 this number had increased to 72 percent.222 By 1960, the decade just prior to the Brady decision, approximately 90 percent of cases in the American criminal justice system were resolved through a guilty plea.223 Perhaps the increase in guilty pleas reflected a growing desire by defendants to confess their sins, but, more likely, it is a reflection of the growing dominance of plea-bargaining throughout this period.

Given that 90 percent of all convictions by 1970 were the result of a guilty plea, most acquired in return for some benefit from the prosecution, the Supreme Court was tasked with a very difficult decision in Brady. If the Court continued with the century-old line of cases regarding confession, plea-bargaining would certainly have been found unconstitutional. The results of such a ruling, however, would be to destroy the criminal justice system, a system that was already struggling to dispose of its ever-growing docket even with 90 percent of cases ending without a trial.224 The shadowy rise of plea-bargaining during the late nineteenth and early twentieth centuries helped solidify its role in the criminal justice system, because, by the time the Supreme Court came to address the issue specifically, all court systems in the United States relied on bargained justice to survive.225

---

222 Id.
223 Borteck, supra note 79, at 1439 n.43 (2004) (citing Corinna Barrett Lain, Accuracy Where It Matters: Brady v. Maryland in the Plea Bargaining Context, 80 WASH. U. L.Q. 1, 1 (2002) (noting that since the 1960s the plea bargaining rate has been around 90 percent)); see also ABA PROJECT ON STANDARDS FOR CRIMINAL JUSTICE, supra note 77, at 1–2 (“The plea of guilty is probably the most frequent method of conviction in all jurisdictions; in some localities as many as 95 per cent of the criminal cases are disposed of in this way.”). Today, pleas of guilty account for over 95 percent of all federal cases. See U.S. SENTENCING COMM’N, supra note 44.
224 See F. Andrew Hessick III & Reshma M. Saujani, Plea Bargaining and Convicting the Innocent: The Role of the Prosecutor, the Defense Counsel, and the Judge, 16 BYU J. PUB. L. 189, 192 (2002) (noting that if plea-bargaining were not available, “the legal system would crumble under the weight of the cases requiring juries and judges”); Thomas R. McCoy & Michael J. Mirra, Plea Bargaining as Due Process in Determining Guilt, 32 STAN. L. REV. 887, 895 n.40 (1980) (citing John Kaplan, American Merchandising and the Guilty Plea: Replacing the Bazaar with the Department Store, 5 AM. J. CRIM. L. 215, 220 (1977)) (stating that the costs of administering the criminal justice system in the 1970s would have doubled if plea-bargaining had been reduced even from 90 percent to 80 percent of all cases).
225 The Court also likely realized that even if it explicitly ruled that plea-bargaining was unconstitutional, the reluctance of prosecutors, defendants, and courts to observe the Court’s earlier rejection of bargained justice meant that plea-bargaining would simply slip back into the shadows as an unofficial means of dealing with the pressures of the system. See Bordenkircher v. Hayes, 434 U.S. 357, 365 (1978) (“[A] rigid constitutional rule that would prohibit a prosecutor from acting forthrightly in his dealings with the defense could only invite unhealthy subterfuge that would drive the practice of plea bargaining back into the shadows from which it has so recently emerged.”).
As if the criminal justice system were not already bogged down with growing caseloads, in part due to overcriminalization, the Supreme Court had just finished handing defendants a number of significant victories during the due process revolution of the 1960s. For instance, the Supreme Court imposed the “exclusionary rule” for violations of the Fourth Amendment, granted the right to counsel, and imposed the obligation that suspects be informed of their rights prior to being interrogated. The result was not only to increase the complexity of criminal prosecutions, but also to increase the length of trials. In many jurisdictions, the length of trial almost doubled from the early 1960s to the late 1960s. According to one commentator, “A major effect of the ‘due process revolution’ was to augment the pressures for plea negotiation. For one thing, the Supreme Court’s decisions contributed to the growing backlog of criminal cases... In addition, the Court’s decisions probably contributed to the increased length of the criminal trial.”

The Supreme Court considered the necessity of plea-bargaining as a tool to promote judicial economy in reaching its decision in *Brady*. The Court noted that 90 to 95 percent of all convictions are the result of a guilty plea and that “with the avoidance of trial, scarce judicial and prosecutorial resources are conserved for those cases in which there is a substantial issue of the defendant’s guilt.” The Court also went on to cite an American Bar Association report on plea-bargaining issued in 1968. The report stated, “It may be noted... that a high proportion of plea of guilty and *nolo contendere* does benefit the system. Such pleas tend to limit

---

226 See generally Dripps, supra note 76, at 1156–61 (discussing the relationship between broadening legal rules and plea-bargaining); Stuntz, supra note 76, at 519–20 (2001) (discussing the influence of broader laws on the rate of plea-bargaining).


228 See Alschuler, supra note 65, at 38 (“In the District of Columbia, the length of the average felony trial grew from 1.9 days in 1950 to 2.8 days in 1965, and in Los Angeles the length of the average felony jury trial increased from 3.5 days in 1964 to 7.2 days in 1968.”).

229 Id.


231 Brady, 397 U.S. at 752 & n.10; see also Santobello v. New York, 404 U.S. 257, 260 (1971) (“The disposition of criminal charges by agreement between the prosecutor and the accused, sometimes loosely called ‘plea bargaining,’ is an essential component of the administration of justice. Properly administered, it is to be encouraged.”).  

232 Brady, 397 U.S. at 752 n.9 (citing ABA PROJECT ON STANDARDS FOR CRIMINAL JUSTICE, supra note 77, at 37–52); see also Bordenkircher v. Hayes, 434 U.S. 357, 363 (1978) (citing ABA PROJECT ON STANDARDS FOR CRIMINAL JUSTICE, supra note 77, at 60–69 in support of its conclusion that a recommendation of leniency is permitted as part of a plea bargain).
the trial process to deciding real disputes and, consequently, to reduce the need for funds and personnel.”

Faced with a daunting decision, the Supreme Court chose to make a great compromise. Defendants had already been afforded significant additional rights during the 1960’s due process revolution. Now the Court would permit defendants, armed with their new rights, the ability to bargain for a more lenient sentence when the government had significant evidence against them. At the same time, plea-bargaining would provide courts with a mechanism for keeping up with their growing case loads and allow them to continue dispensing justice in a timely manner.

III. PLEA BARGAINING’S INNOCENCE PROBLEM

A. The Coercive Consequences of the Great Compromise

In the same year the Supreme Court decided Brady, it also handed down another plea-bargaining decision that helped to solidify bargained justice as a major facet of the American criminal justice system. In North Carolina v. Alford, the Court stated that a defendant could plead guilty in return for some benefit, such as a reduced sentence, while continuing to maintain his or her innocence. The Court inserted a caveat, however, requiring the “record before the judge contain[] strong evidence of actual guilt” to ensure the rights of the innocent are protected and guilty pleas are the result of “free and intelligent choice.”

In 1977, Kerry Max Cook was arrested for the rape and murder of a woman in Tyler, Texas. Though Cook was convicted by a jury and sentenced to death a year later, he continued to profess his innocence. During his time on death row, Cook was abused, raped, and attempted to commit suicide twice. After one suicide attempt, prison officials found a note that stated, “I really was an innocent man.” His initial conviction was eventually overturned and a second trial in

---

233 See ABA PROJECT ON STANDARDS FOR CRIMINAL JUSTICE, supra note 77, at 2.
235 Id. at 37; see also Andrew D. Leipond, How the Pretrial Process Contributes to Wrongful Convictions, 42 AM. CRIM. L. REV. 1123, 1156 (2005) (“An Alford plea, where the defendant pleads guilty but simultaneously denies having committed the crime, clearly puts the court on notice that this guilty plea is problematic . . . .”).
236 Alford, 400 U.S. at 37, 38 n.10; Leipond, supra note 235. Currently, the federal system, the District of Columbia, and forty-seven states permit Alford pleas. See Stephanos Bibas, Harmonizing Substantive-Criminal-Law Values and Criminal Procedure: The Case of Alford and Nolo Contendere Pleas, 88 CORNELL L. REV. 1361, 1372 n.52 (2003).
238 Id.
239 Id.
240 Id.
1992 resulted in a hung jury.\textsuperscript{241} The government retried Cook again, and the death penalty was imposed by a jury a second time.\textsuperscript{242} In 1996, Cook’s conviction was again overturned, this time due to police and prosecutorial misconduct dating back to the initial investigation and trial of Cook in 1978.\textsuperscript{243} Despite the numerous setbacks, the government moved forward to retry Cook a fourth time.\textsuperscript{244} Due to the prior misconduct, however, the prosecutor in the case would no longer be able to use the testimony of a central witness in the case.\textsuperscript{245} As the trial for Cook’s life approached, the prosecution conceded that the case was looking increasingly weak.\textsuperscript{246}

Having discussed the evolution of the plea-bargaining machine it will come as no surprise that the prosecutor responded to the significant likelihood of losing a trial by offering Cook a plea deal.\textsuperscript{247} In return for pleading guilty, Cook would receive a sentence of time served and walk out of prison.\textsuperscript{248} Cook refused, however, continuing to profess his innocence.\textsuperscript{249}

Kerry [Max Cook] looked [his attorney] in the eye and said, “I want to be free, I want this behind me, but I will go back to death row, I will let them strap me to the gurney and put the poison in my veins before I lie, before I plead guilty.”\textsuperscript{250}

In response, the prosecutor offered Cook the same deal in return for an \textit{Alford} plea.\textsuperscript{251} Cook could now plead guilty, while, at the same time, continuing to maintain his innocence.\textsuperscript{252} Cook took the plea agreement and, twenty-two years after being placed in prison, walked free.\textsuperscript{253} Two months later, a DNA test conclusively demonstrated that Cook was not a match to forensic evidence

\begin{flushright}
241 \textit{Id.} \\
242 \textit{Id.} \\
243 \textit{Id.} (The Texas Court of Criminal Appeals “wrote that the investigation was intentionally misleading, the testimony of the key witness, Robert Hoehn, was prejudicial, and the first conviction was obtained through fraud and in violation of the law.”). \\
244 \textit{Id.} \\
245 \textit{Id.} (“The irony was that the testimony of the witness, Robert Hoehn, who had died in the meantime, was the example cited by the Texas Court of Criminal Appeals as prejudicial and contradictory.”). \\
246 \textit{See id.} \\
247 \textit{Id.} \\
248 \textit{Id.} \\
249 \textit{Id.} \\
250 \textit{Id.} \\
251 \textit{Id.} \\
252 \textit{Id.} \\
253 \textit{Id.}
\end{flushright}
obtained at the scene of the crime in 1977.254 Though he had been induced to plead guilty, Cook was, in fact, innocent.255

Today, over 95 percent of defendants in the criminal justice system plead guilty and, in most cases, such confessions are prompted by offers of leniency or other benefits from the prosecution.256 It is unclear how many of these defendants are innocent, but it is clear that plea-bargaining has an innocence problem.257 At

254 Id.
255 Id. (“So, you know, was it worth it? Sometimes when I’m holding my son I can say yes. Sometimes, when I’m by myself, I say no. They won.”). Cook is certainly not the only innocent defendant coerced into pleading guilty by overpowering plea offers. A 1999 Los Angeles Times article describing newly discovered evidence in a police corruption case discussed the case of two defendants who may have been framed by police. See Samuel H. Pillsbury, Even the Innocent Can be Coerced Into Pleading Guilty, L.A. TIMES (Nov. 28, 1999), http://articles.latimes.com/1999/nov/28/opinion/op-38287. According to the article, though the new evidence indicated that both men were innocent, each had previously pleaded guilty to the fabricated charges to avoid the risk of serving lengthy prison sentences if convicted at trial. Id. (“Oscar Peralta (aka Jose Perez) pleaded guilty to assault on a police officer in a shooting incident, although he believed that he was framed in the case to cover up improper police conduct. Peralta admitted guilt to win a promised sentence of time served—10 months in county jail—and avoid the life sentence he faced had he been convicted at trial.”); see also Morris B. Hoffman, The Myth of Factual Innocence, 82 CHI.-KENT L. REV. 663, 672 (2007) (“Do innocent people plead guilty? Of course. Innocent people sometimes even preempt their false guilty pleas by falsely confessing, both with and without overbearing interrogation.”).

256 See U.S. SENTENCING COMM’N, supra note 44, at fig.C (stating that in 2009, 96.3 percent of federal defendants pleaded guilty).
257 See Baldwin & McConville, supra note 64, at 296 (“[T]he results of our research suggest that some defendants who perhaps were innocent, and a larger group who probably would have been acquitted had the case gone to trial, were nonetheless induced to plead guilty.”); Russell D. Covey, Signaling and Plea Bargaining’s Innocence Problem, 66 WASH. & LEE L. REV. 73, 74 (2009) (“Plea bargaining has an innocence problem.”); Michael O. Finkelstein, A Statistical Analysis of Guilty Plea Practices in the Federal Courts, 89 HARV. L. REV. 293, 295 (1975) (“On the basis of the analysis that follows, I conclude that the pressure on defendants to plead guilty in the federal courts has induced a high rate of conviction by ‘consent’ in cases in which no conviction would have been obtained if there had been a contest.”); Daniel Givelber, Meaningless Acquittals, Meaningful Convictions: Do We Reliably Acquit the Innocent?, 49 RUTGERS L. REV. 1317, 1343–44 (1997) (arguing “that America’s criminal justice system creates a significant risk that innocent people will be systematically convicted”); Leipond, supra note 235, at 1154 (“Yet we know that sometimes innocent people plead guilty . . . .”); Robert E. Scott & William J. Stuntz, Plea Bargaining as Contract, 101 YALE L.J. 1909, 1949–51 (1992) (discussing plea-bargaining’s innocence problem); David L Shapiro, Should a Guilty Plea Have Preclusive Effect?, 70 IOWA L. REV. 27, 27 (1984) (discussing the opposition to argument that plea bargain should “be preclusive against the person who entered the plea”); see also Oren Gazal-Ayal, Partial Ban on Plea Bargains, 27 CARDOZO L. REV. 2295, 2295–96 (2006) (arguing a partial ban on plea-bargaining would assist in preventing innocent defendants from being forced to plead guilty by forcing prosecutors to allocate their limited resources towards only strong cases).
least one study has concluded that as many as 27 percent of defendants who plead guilty would not have been convicted at trial, though this estimate seems exceptionally high. Another more empirically driven study examined DNA evidence in capital rape-murder cases and determined that between 3.3 and 5 percent of those convicted, either through trial or a guilty plea, were factually innocent. Other studies have placed the number of defendants who plead guilty as a result of inducements by the government but who are factually innocent between 1.6 percent and 8 percent. Taking even the lowest of these estimates, the reality is striking and means that in 2009 there were over 1,250 innocent defendants forced to falsely admit guilt in the federal system alone. Extrapolated out to the entire American criminal justice system since 1970, there are

---

258 See Shapiro, supra note 257, at 45; see also Finkelstein, supra note 257, at 309–10 (concluding that “at least one-third of all defendants pleading guilty in high-[plea-bargaining] rate districts would ultimately have escaped conviction if they had refused to consent”).

259 D. Michael Risinger, Innocents Convicted: An Empirically Justified Factual Wrongful Conviction Rate, 97 J. CRIM. L. & CRIMINOLOGY 761, 778 (2007) (demonstrating through analysis of DNA exonerations for capital rape-murders from 1982 to 1989 that the minimum factually wrongful conviction rate was 3.3 percent); see also Brandon L. Garrett, Judging Innocence, 108 COLUM. L. REV. 55, 74 (2008) (noting that nine of the first two hundred individuals exonerated by the Innocence Project had pleaded guilty); Scott & Stuntz, supra note 257, at 1948 n.134 (“[I]f acquittal is correlated (albeit not perfectly) with innocence, and there is an unusually high percentage of acquittals in murder cases, then there may be an unusually high number of innocent defendants on trial for murder. Since risk aversion should lead many (perhaps most) innocent defendants to plead guilty, even given a substantial acquittal rate, the pool of all murder defendants probably includes an unusually high number of innocents.”).

260 See Baldwin & McConville, supra note 64, at 296–98 (discussing plea-bargaining’s innocence problem in England); Givelber, supra note 257, at 1343–44 (estimating nearly 8 percent of all convictions would be false if “innocent people plead guilty at as high a rate as innocent people are convicted following trial”); id. (“These figures represent convictions of innocent people following trial. In this country at least, 90 percent of convictions are secured through guilty pleas. Here, the criminal justice system is truly operating without a compass. If it is modestly assumed that, at the least, there are as many total convictions of innocent defendants resulting from pleas as there are from trials, one arrives, using the English study percentages, at a figure of 1.6 percent. This figure is a slightly higher estimate than that of the criminal justice officials surveyed by Huff, Rattner and Sagarin. If the more aggressive assumption is made that innocent people plead guilty at as high a rate as innocent people are convicted following trial, then nearly 8% of all convictions would be false.”); George C. Thomas III, Two Windows Into Innocence, 7 OHIO ST. J. CRIM. L. 575, 577–78 (2010) (“McConville and Baldwin concluded that two percent of the guilty pleas were of doubtful validity. As there were roughly two million felony cases filed in 2006, if two percent result in conviction of an innocent defendant, 40,000 wrongful felony convictions occur per year.”).

261 U.S. SENTENCING COMM’N, supra note 44, at tbl.10 (stating that in 2009, there were 81,372 defendants convicted in the federal system, of which 96.3 percent pleaded guilty).
conservatively tens of thousands of innocent persons who have been induced to plead guilty by overpowering plea bargains. If this is the case, plea-bargaining certainly has a significant and unacceptable innocence problem.

In 1991, John Dixon pleaded guilty to first degree kidnapping, first degree robbery, two counts of first degree aggravated sexual assault, and unlawful possession of a weapon in the third degree, and was sentenced to forty-five years in prison. Dixon had not entered a guilty plea because he was remorseful for something he had done. Dixon pleaded guilty because he was threatened with the possibility of a higher sentence if he lost at trial and offered the promise of leniency if he relieved the government of its burden at trial. In 2001, Dixon’s conviction was vacated after DNA testing revealed that he could not have been the assailant in the crime. A week later, after ten years in prison, he was released. Dixon reminds us of the real people behind the statistics and demonstrates that with significant regularity defendants are being compelled to falsely plead guilty to offenses they did not commit. The question remains, however, is there a remedy to plea bargaining’s innocence problem or did, as some believe, the Supreme Court grant prosecutors unlimited powers to bargain in *Brady*?

---

262 *The Innocence Project - Know the Cases: Browse Profiles: John Dixon*, INNOCENCE PROJECT, http://www.innocenceproject.org/Content/86.php (last visited Feb. 7, 2012) (describing the story of John Dixon, who pleaded guilty to rape charges for fear he would receive a harsher sentence if he proceeded to trial, but was later exonerated by DNA evidence).

263 *Id.; see also* Richard Klein, *Due Process Denied: Judicial Coercion in the Plea Bargaining Process*, 32 Hofstra L. Rev. 1349, 1408 (2004) (“The defendant pleads guilty to reduce his punishment; it’s a matter of expediency. Remorse has very little to do with it, and judges, who, if anything, are more cynical then [sic] the rest of us, are fully aware of that.”).

264 *The Innocence Project - Know the Cases: Browse Profiles: John Dixon*, INNOCENCE PROJECT, http://www.innocenceproject.org/Content/86.php (last visited Feb. 7, 2012); *see also* Klein, *supra* note 263, at 1398 (“By the time of the plea allocution it is clear that the defendant has decided to take the plea bargain and knows or has been instructed by counsel to tell the court that he did indeed do the crime. Predictably, the National Institute of Justice survey found that judges rejected guilty pleas in only two percent of cases. Since efficiency and speed is the name of the game, it is not unexpected that meaningful questioning of the defendant does not occur and it is not surprising that the Institute concluded that the plea allocution procedure is ‘close to being a new kind of pious fraud.’”); Wright, *supra* note 76, at 93 (“But when it comes to the defendant's ‘voluntariness’— the second half of the formula—courts have walked away. The proper knowledge, together with a pro forma statement from the defendant that her guilty plea was not coerced, normally suffices.”).

265 *The Innocence Project, supra* note 264.

266 *Id.

267 *See id.; Leipold, supra* note 235, at 1154 (“[S]ometimes the prosecutor offers such a generous discount for admitting guilt that the defendant feels he simply can’t take the chance of going to trial.”).
B. The Constitutional Limits of the Great Compromise
and a Litmus Test for the Brady Safety-Valve

While the Brady decision signaled a shift away from wholesale rejection of bargained justice, it contained an important limitation regarding how far the Court would permit prosecutors to venture in attempting to induce defendants to plead guilty. In the concluding paragraphs of the decision, the Supreme Court discussed its vision for the utilization of plea-bargaining in the criminal system. Plea-bargaining was a tool for use only in cases where the evidence was overwhelming and where the defendant, unlikely to succeed at trial, might benefit from the opportunity to bargain for a reduced sentence. Plea-bargaining, however, was not to be used to overwhelm defendants into pleading guilty where their guilt was in question.

For a defendant who sees slight possibility of acquittal, the advantages of pleading guilty and limiting the probable penalty are obvious—his exposure is reduced, the correctional processes can begin immediately, and the practical burdens of a trial are eliminated. For the State there are also advantages—the more promptly imposed punishment after an admission of guilt may more effectively attain the objectives of punishment; and with the avoidance of trial, scarce judicial and prosecutorial resources are conserved for those cases in which there is a substantial issue of the defendant’s guilt or in which there is substantial doubt that the State can sustain its burden of proof.

According to the Court, if the government were to begin offering significant incentives to defendants whose guilt was uncertain in an effort to motivate them to plead guilty, defendants who might in fact be innocent, the Court would be forced to reconsider its approval of the plea-bargaining system.

This is not to say that guilty plea convictions hold no hazards for the innocent or that the methods of taking guilty pleas presently employed in this country are necessarily valid in all respects. This mode of conviction is no more foolproof than full trials to the court or to the jury. Accordingly, we take great precautions against unsound results, and we should continue to do so, whether conviction is by plea or by trial. We would have serious doubts about this case if the encouragement of guilty pleas by offers of leniency substantially increased the likelihood that

269 See id. at 752.
270 See id. at 750, 752.
271 Id. at 752 (emphasis added).
272 See id. at 757–58.
defendants, advised by competent counsel, would falsely condemn themselves.273

While the Supreme Court was willing to permit a compromise in the interests of justice and judicial economy in 1970, it also created a safety-valve.274 Safety-valves are intended to relieve pressure when forces within a machine become too great and, thereby, preserve the integrity of the machine.275 The Brady safety-valve serves just such a purpose by placing a limit on the amount of pressure that can constitutionally be placed on defendants to plead guilty. According to the Court, however, should plea-bargaining become so common that prosecutors offer deals to all defendants, including those whose guilt is in question, and the incentives to bargain become so overpowering that even innocent defendants acquiesce, then the Brady safety-valve will have failed and the plea-bargaining machine will have ventured into the realm of unconstitutionality.276

Interestingly, the Brady decision is not the only Supreme Court case from the 1970’s that discusses the Court’s concerns regarding innocent defendants and that makes mention of a safety-valve or threshold regarding acceptable incentives to bargain.277 In Alford, the Court stated that Alford plea agreements must only be

273 Id. The sentiment that innocent defendants should not be encouraged to plead guilty has been echoed by academics. See, e.g., Bibas, supra note 236, at 1382 (“Even if innocent defendants want to plead guilty, the law should not go out of its way to promote these unjust results.”); Leipold, supra note 235, at 1158 (supporting the statements of Stephanos Bibas regarding innocent defendants and plea bargaining).
274 See Brady, 397 U.S. at 757–58.
276 See Brady, 397 U.S. at 757–58.
277 See Bordenkircher v. Hayes, 434 U.S. 357, 363–65 (1978) (discussing plea bargains and noting there are “undoubtedly constitutional limits” to the exercise of plea bargaining); North Carolina v. Alford, 400 U.S. 25, 37–39 (1970) (holding as valid a guilty plea made to avoid the death penalty and discussing when a court should accept guilty pleas”). Even the 1968 ABA report regarding plea-bargaining accepted the premise that plea-bargaining should be for defendants whose guilt is demonstrated by overwhelming evidence, thus freeing resources for the trial of those defendants whose guilty is less certain.

[A] high proportion of pleas of guilty and nolo contendere does benefit the system. Such pleas tend to limit the trial process to deciding real disputes and, consequently, to reduce the need for funds and personnel. If the number of judges, courtrooms, court personnel and counsel for prosecution and defense were to be increased substantially [if necessitated by a decreased use of plea bargains], the funds necessary for such increases might be diverted from elsewhere in the criminal justice process. Moreover, the limited use of the trial process for those cases in which the defendant has grounds for contesting the
utilized in cases where the evidence regarding actual guilt is strong and sufficient to negate the defendant’s claim of innocence.\textsuperscript{278} Where there is doubt as to the defendant’s culpability, the case must be sent to trial and no plea bargain may be struck so as to protect the “innocent and... insur[e] that guilty pleas are a product of free and intelligent choice.”\textsuperscript{279} Unfortunately, despite this clear standard from the Supreme Court, Kerry Max Cook’s false confession of guilt was accepted even though the offer itself signaled the prosecution’s belief that it would lose at trial.\textsuperscript{280}

In 1978, the Supreme Court addressed plea-bargaining once more in \textit{Bordenkircher v. Hayes},\textsuperscript{281} and, once again, expressed concern regarding innocent defendants. Paul Hayes was indicted on a charge of uttering a forged instrument, an offense under Kentucky law punishable by two to ten years in prison.\textsuperscript{282} During plea negotiations, the prosecution stated that if Hayes pleaded guilty, he would recommend a sentence of five years.\textsuperscript{283} If he did not “save the court the inconvenience and necessity of a trial,” however, the prosecution threatened to re-indict Hayes under a habitual offender law that carried a life sentence.\textsuperscript{284} Hayes refused to plead guilty and, as a result, was re-indicted, found guilty at trial, and sentenced to life in prison.\textsuperscript{285}

In holding that the prosecutor in the \textit{Bordenkircher} case had not acted inappropriately, the Supreme Court began by clearly stating that “[t]o punish a person because he has done what the law plainly allows him to do is a due process violation of the most basic sort.”\textsuperscript{286} The Court continued, however, “in the ‘give-and-take’ of plea bargaining, there is no such element of punishment or retaliation so long as the accused is free to accept or reject the prosecution’s offer.”\textsuperscript{287} As long

\begin{flushright}
matter of guilt aids in preserving the meaningfulness of the presumption of innocence.
\end{flushright}

\textsuperscript{278} See \textit{Alford}, 400 U.S. at 37–38.
\textsuperscript{279} Id. at 38 n.10; see also Bibas, \textsuperscript{\textit{supra}} note 236, at 1382 (“It should go without saying that it is wrong to convict innocent defendants. Thus, the law should hinder these convictions instead of facilitating them through \textit{Alford} and \textit{nolo contendere} pleas.”).
\textsuperscript{280} See \textit{infra} notes 237–255 and accompanying text (discussing the Kerry Max Cook case).
\textsuperscript{281} \textit{Bordenkircher}, 434 U.S. 357.
\textsuperscript{282} Id. at 358.
\textsuperscript{283} Id.
\textsuperscript{284} Id. at 358–59.
\textsuperscript{285} Id. at 359.
\textsuperscript{286} Id. at 363.
\textsuperscript{287} Id.; see also Corbitt v. New Jersey, 439 U.S. 212, 225 (1978) (“Finally, we are unconvinced that the New Jersey statutory pattern exerts such a powerful influence to coerce inaccurate pleas non vult that it should be deemed constitutionally suspect. There is no suggestion here that Corbitt was not well counseled or that he misunderstood the choices that were placed before him. Here, as in \textit{Bordenkircher}, the State did not trespass on the defendant’s rights ‘so long as the accused [was] free to accept or reject’ the choice presented to him by the State that is, to go to trial and face the risk of life imprisonment or
as the incentives to bargain are sufficiently reasonable and the defendant retains this sense of freedom, the Court determined that it is unlikely an innocent defendant would be “driven to false self-condemnation.” Of course, the Supreme Court was wrong regarding its estimation of the impact of prosecutors’ offers on innocent defendants. As John Dixon demonstrates, it is very likely that innocent defendants will plead guilty in return for the incredible incentives currently being offered by the prosecution, incentives that strip away a defendant’s ability to freely decide whether to accept or reject the government’s advances.

While the Supreme Court in *Brady, Alford,* and *Bordenkircher* noted that plea bargains were only intended for cases with strong evidentiary support and not as a tool to induce all defendants, including the innocent, to plead guilty, modern day plea-bargaining is clearly at odds with this central edict of the Supreme Court’s great compromise of 1970. Today, defendants whose guilt is uncertain are offered plea bargains as often, if not more often, than those for whom the evidence is particularly strong. Not only are these defendants of questionable guilt offered to seek acceptance of a *non vult* plea and the imposition of the lesser penalty authorized by law.

---

288 *Bordenkircher,* 434 U.S. at 363 (“Indeed, acceptance of the basic legitimacy of plea bargaining necessarily implies rejection of any notion that a guilty plea is involuntary in a constitutional sense simply because it is the end result of the bargaining process.”).

289 See supra notes 262–267 and accompanying text (discussing the John Dixon case).


291 See John G. Douglass, *Fatal Attraction? The Uneasy Courtship of Brady and Plea Bargaining,* 50 EMORY L.J. 437, 449 (2001) (The prosecutor “will offer greater sentencing concessions in those cases where conviction is less likely, and fewer concessions where she is more confident of conviction.”); Gazal-Ayal, supra note 257, at 2298–99 (“When the case is weak, meaning when the probability that a trial would result in conviction is relatively small, [the prosecutor] can assure a conviction by offering the defendant a substantial discount—a discount big enough to compensate him for foregoing the possibility of being found not guilty.”); Hessick & Saujani, supra note 224, at 199 (“Assuming the innocent defendant is more likely to be acquitted at trial, the prosecutor has much higher incentives to enter a plea agreement with the innocent defendant than with a guilty defendant because the prosecutor perceives the innocence as a lack of evidence.”); Yin, supra note 90, at 443 (“Curiously, the arena of plea bargaining pits the concepts of duress and consideration against each other: a large sentencing differential makes it more likely that a defendant is coerced into pleading guilty, and yet it also increases the benefit
plea bargains, their bargains are often more generous and, therefore, more irresistible than those offered to defendants against whom the government believes it has a strong chance of success at trial.\footnote{292} As a result, while the Supreme Court in 1970 did not desire that plea-bargaining be utilized to induce the innocent to plead guilty, it is precisely these defendants who are placed under the most severe pressure to confess today.\footnote{293}

It is important to note that while plea bargaining’s innocence problem serves as an indication that we have gone beyond the types of incentives and persuasions to plead guilty permitted by the \textit{Brady} safety-valve, the Supreme Court did not intend that this threshold on pressure apply only to the innocent. Rather, the innocent serve as a litmus test for plea-bargaining more generally. If prosecutors are offering incentives that are sufficient to force innocent defendants to plead guilty, these same types of incentives are equally unconstitutional when offered to guilty defendants because both sets of defendants are involuntarily entering their pleas.

Recall that while \textit{Brady} drastically changed the landscape of American criminal justice by permitting plea-bargaining, it retained the requirement that all pleas be “voluntary.” Clearly, the \textit{Brady} court did not intend “voluntary” in the plea-bargaining context to retain the restrictive meaning it had held when pleas of guilt were considered legally identical to confessions, because such a definition would make almost every offer of leniency impermissible. Instead, “voluntary” took on a new, although woefully undefined, meaning.\footnote{294} To discern what “voluntary” meant after \textit{Brady} one must examine the use of this term throughout Supreme Court precedent.\footnote{295} In so doing, one learns that the term has less to do with the type of incentives offered and more to do with the effect of such offers on defendants. This is why plea-bargaining’s innocence problem serves as a litmus test for voluntariness generally.

\footnote{\textit{See infra} note 291.}

\footnote{\textit{See Covey, supra} note 257, at 74, 79–80 (“Plea bargaining has an innocence problem. . . . In short, as long as the prosecutor is willing and able to discount plea prices to reflect resource savings, regardless of guilt or innocence, pleading guilty is the defendant’s dominant strategy. As a result, non-frivolous accusations—not proof beyond a reasonable doubt—is all that is necessary to establish legal guilt. This latter point forms the root of plea-bargaining’s ‘innocence problem,’ which refers here not merely to the fact that innocent people plead guilty, but that the economics of plea bargaining drives them to do so.”).}

\footnote{\textit{See supra} notes 198–218 and accompanying text (discussing \textit{Brady v. United States}).}

\footnote{\textit{See supra} Part I.
Cases dating back to the eighteenth century, such as *Rex v. Warickshall* (1783), imparted a central tenet that has remained true throughout modern times—admissions of guilt forced from defendants by overwhelming incentives strip such defendants of free will and must be discredited. In the earliest examples of American case law regarding guilty pleas and confessions, the Court trumpets this same overarching theme of free will. In cases such as *Hallinger v. Davis* (1892) and *Bram v. United States* (1897), the Court specifically required confessions and pleas of guilty to be more than just voluntary—they had to be entered “freely.” By the mid-twentieth century, in cases such as *Walker v. Johnson* (1941), *Waley v. Johnson* (1942), and *United States v. Jackson* (1968), the Supreme Court began using the term “coercive” to describe guilty pleas and confessions that were not given freely, but the intent was the same. After all, since the fifteenth century “coerce” has meant to dominate by force or threat. By 1962, in *Machibroda v. United States*, the Supreme Court was once again drawing a clear line between permissible pleas and those “induced by promises or threats which deprive [the plea] of the character of a voluntary act . . . .” Perhaps no case better summarizes the concept that guilty pleas must be entered into freely than the first case discussed in this Article—the case of Beatrice Lynumn. The *Lynumn* Court

---

296 See Alschuler, supra note 65, at 12–13.
297 Hallinger v. Davis, 146 U.S. 314, 324 (1892); Bram v. United States, 168 U.S. 532, 542–43, 544–48 (1897) (“But a confession, in order to be admissible, must be free and voluntary: that is, must not be extracted by any sort of threats of violence, nor obtained by any direct or implied promises, however slight, nor by the exertion of any improper influence . . . for the law cannot measure the force of the influence used, or decide upon its effect upon the mind of the prisoner . . . .” (quoting 3 RUSSELL ON CRIMES 478 (6th ed.))).
298 See Waley v. Johnston, 316 U.S. 101, 104 (1942) (“For a conviction on a plea of guilty coerced by a federal law enforcement officer is no more consistent with due process than a conviction supported by a coerced confession.”); Walker v. Johnston, 312 U.S. 275, 286–87 (1941) (“If he did not voluntarily waive his right to counsel, or if he was deceived or coerced by the prosecutor into entering a guilty plea, he was deprived of a constitutional right.”); c.f. United States v. Jackson, 390 U.S. 570, 583 (1968) (“[T]he evil in the federal statute is not that it necessarily coerces guilty pleas and jury waivers but simply that it needlessly encourages them. A procedure need not be inherently coercive in order that it be held to impose an impermissible burden upon the assertion of a constitutional right.”).
301 Id. at 493 (“There can be no doubt that, if the allegations contained in the petitioner’s motion and affidavit are true, he is entitled to have his sentence vacated. A guilty plea, if induced by promises or threats which deprive it of the character of a voluntary act, is void. . . . Like a verdict of a jury[, a plea of guilty] is conclusive. . . . Out of just consideration for persons accused of crime, courts are careful that a plea of guilty shall not be accepted unless made voluntarily after proper advice and with full understanding of the consequences.”).
302 See Lynumn v. Illinois, 372 U.S. 528 (1963); see also supra notes 2–22 and accompanying text.
stated, where the defendant’s “will was overborne at the time [s]he confessed, . . . [the] confession cannot be deemed ‘the product of a rational intellect and a free will.’” 303 That Lynumn had a choice when confronted with the possibility of losing her children was clear, but, as the Supreme Court recognized, the choice was illusory. 304 The incentive to bargain, which she believed meant retaining custody of her children, was so powerful as to be overbearing and remove any elements of freedom in her decision. Her confession, though gladly offered, was nonetheless involuntary. 305

When the Supreme Court addressed plea-bargaining in Brady, almost two centuries of precedent established that the incentives offered in return for a plea must not be so powerful as to remove the defendant’s ability to consider the choice and act with “free will.” 306 Though the Court permitted a major shift in plea bargaining policy in Brady, it certainly did not remove this requirement. In fact, Brady, Alford, and Bordenkircher all specifically retained the threshold requirement of free will by continuing to require that all pleas be entered into voluntarily. 307 In Brady, the Court stated, there may be no “mental coercion overbearing the will of the defendant.” 308 The same year, in Alford, the Court reiterated that guilty pleas must be “free and intelligent,” not irrational acts forced

---

303 Lynum, 372 U.S. at 534 (“We think it clear that a confession made under such circumstances must be deemed not voluntary, but coerced.”).

304 In his article regarding judicial intervention in plea bargaining, Richard Klein briefly discusses the similarities between coerced confessions and the activities prohibited by Miranda v. Arizona. See Klein, supra note 263, at 1419 (“This is an inherently coercive atmosphere; one which could well cause even an innocent individual to realize that his most sensible option, indeed perhaps his only option, is to take the plea deal that will never be offered again. The setting and the pressures are strikingly similar to those the Supreme Court found to be intolerable in Miranda v. Arizona, where the entire thrust of the suspect’s interrogation ‘was to put the defendant in such an emotional state as to impair his capacity for rational judgment. . . . [T]he compelling atmosphere of the in-custody interrogation, and not an independent decision on his part, caused the defendant to speak.’”).

305 See supra notes 2–22 and accompanying text (discussing the Lynumn confession).

306 See supra notes 296–305 and accompanying text.

307 Brady v. United States, 397 U.S. 742, 758 (1970); North Carolina v. Alford, 400 U.S. 25, 37–39 (1970); Bordenkircher v. Hayes, 434 U.S. 357, 363 (1978). Even the American Bar Association report from 1968 seemed to adhere to this basic tenet of plea bargaining as it was then understood. See ABA PROJECT ON STANDARDS FOR CRIMINAL JUSTICE, supra note 77, at 29 (“The court should not accept a plea of guilty or nolo contendere without first determining that the plea is voluntary. . . . The court should [] address the defendant personally and determine whether any other promises or any force or threats were used to obtain the plea.”).

308 Brady, 397 U.S. at 750 (emphasis added) (“Of course, the agents of the State may not produce a plea by actual or threatened physical harm or by mental coercion overbearing the will of the defendant.”).
from the defendant. Finally, eight years later, in *Bordenkircher*, the Court stated, the key criteria for acceptance of guilty pleas is the determination that the defendant, in making his or her decision, was “free to accept or reject the prosecution’s offer.”

“Voluntary,” therefore, even after *Brady*, means that the incentives offered to defendants may not be so coercive as to overbear the individual’s ability to act freely and decide in a rational manner whether to accept or reject the government’s offer.

When one considers this definition of “voluntary,” the reason the Supreme Court created a litmus test regarding innocent defendants to monitor the operation of the *Brady* safety-valve becomes evident. Innocent defendants serve as the most efficient monitor of whether defendants are being afforded the opportunity to exercise their free will and decide in a rational manner whether to accept or reject a plea bargain. As the *Bordenkircher* Court stated, innocent defendants should not be “driven to false self-condemnation” by the plea-bargaining system.

Instead, innocent defendants should challenge the government in most cases and, though trials are not perfect, exercise their right to be proven guilty beyond a reasonable doubt before a jury. Where, however, innocent defendants in significant numbers are not exercising these rights, but are being coerced to plead guilty, such acts serve as a strong indication that the *Brady* safety-valve has failed and defendants, both innocent and guilty, are being overborne by unconstitutional incentives.

It is important to note that just because plea-bargaining has an innocence problem, it does not necessarily follow that all plea bargains are unconstitutional. Rather, the innocence litmus test serves only to alert the Court that plea-bargaining as an institution has begun to drift into impermissible territory. For instance, few would argue that a defendant offered six years in prison for pleading guilty or seven years in prison if he or she losses at trial would be stripped of the ability to make a rational and free decision because of the polarity of these options. But, perhaps, the offer in the Lea Fastow case does represent an impermissible inducement. Did Lea Fastow really have a choice after she was presented with her

---

309 *Alford*, 400 U.S. at 38 n.10 (cautioning against accepting “pleas coupled with claims of innocence” without a factual basis for the plea, so as to protect the “innocent and . . . insur[e] that guilty pleas are a product of free and intelligent choice . . . .”).

310 *Bordenkircher*, 434 U.S. at 363 (“[I]n the ‘give-and-take’ of plea bargaining, there is no such element of punishment or retaliation so long as the accused is free to accept or reject the prosecution’s offer.”).

311 *Id.* (“Indeed, acceptance of the basic legitimacy of plea bargaining necessarily implies rejection of any notion that a guilty plea is involuntary in a constitutional sense simply because it is the end result of the bargaining process.”).

312 It is important to note that this Article is not arguing that innocent defendants should never be permitted to plead guilty. In rare cases, a defendant who is innocent may nonetheless have an exceedingly small chance of success at trial because of the available evidence. An example would include a case involving a mistaken, yet convincing, eyewitness identification. These defendants should be given the opportunity to plead guilty in return for a benefit. Where significant numbers of defendants are pleading guilty, however, this serves as an indication of a larger and more systemic problem.
options? Did Beatrice Lynumn have such a choice? The distinction, therefore, between plea bargains that are permitted by *Brady* and those that violate the great compromise of 1970 is the size of the sentencing differential in each case. At some point, the sentencing differential becomes so large that it destroys the defendant’s ability to act freely and decide in a rational manner whether to accept or reject the government’s offer. The solution to the failure of the *Brady* safety-valve, therefore, is to limit the size of the sentencing differentials used to induce defendants to plead guilty.\(^{313}\)

It should come as no surprise that the *Brady* safety-valve is failing and pleas are being entered involuntarily by both innocent and guilty defendants, because prosecutors have been gaining increased powers to bargain for over half a century.\(^{314}\) As discussed in Part I.B, as prosecutors’ powers to bargain increase, so too does their ability to create startling sentencing differentials as inducements for defendants to plead guilty.\(^{315}\) Consider for a moment the sentencing differential for Lea Fastow: ten years in prison and parentless children versus five months in jail and a guarantee that one parent will always be at home.

While the Supreme Court did not discuss the relationship between voluntary plea-bargaining and sentencing differentials in its *Brady* opinion, the 1968 American Bar Association report on plea-bargaining, which the Court cited with approval, had some very noteworthy comments.

Assuming that two defendants have engaged in the same conduct under essentially the same circumstances and that the usual presentence information as to the two does not materially differ, is it proper to give a somewhat lower sentence to one defendant because he has consented to enter a plea of guilty? . . . [I]t appears that most judges consider such leniency proper if the sentence disparity is not unreasonable.\(^{316}\)

---

\(^{313}\) As various commentators have noted, abolishing plea-bargaining in its entirety would not only crush the existing criminal justice system, it would eliminate a system that is beneficial in many ways for defendants, prosecutors, and courts. *See*, e.g., Covey, *supra* note 257, at 83–86 (noting that abolishing plea-bargaining does not solve the innocence problem, and does not improve the quality or accuracy of trials); Gazal-Ayal, *supra* note 257, at 2299 (“As many scholars have shown, a total ban on plea bargaining is hardly feasible in the overloaded American criminal justice system.”).

\(^{314}\) *See supra* Part I.B (discussing the increases in prosecutorial power during the twentieth century).

\(^{315}\) *See supra* Part I.B; *see also* Russell D. Covey, *Fixed Justice: Reforming Plea Bargaining with Plea-Based Ceilings*, 82 Tul. L. Rev. 1237, 1238–41 (2008) (arguing that large sentencing differential forced some innocent defendants to accept plea bargains); Jenia Iontcheva Turner, *Judicial Participation in Plea Negotiations: A Comparative View*, 54 Am. J. Comp. L. 199, 204 (2006) (“Two features of our current federal system are especially concerning in this respect—harsh sentences and steep discounts for pleading guilty. Together, they may induce even defendants with good odds of prevailing at trial to accept a plea bargain.”).

\(^{316}\) *See ABA Project on Standards for Criminal Justice, supra* note 77, at 37–38 (emphasis added).
Thus, even the ABA’s report praising plea-bargaining in 1968 made clear that when sentencing disparities reach unreasonable levels they reflect an unacceptable type of bargained justice. The Supreme Court agreed and, therefore, while it embraced plea-bargaining in the Brady decision, it also created a safety-valve and a litmus test to monitor its success. As the Court understood in 1970, one can hardly be expected to make a rational and free decision regarding an offer to plead guilty when the differences between the offered leniency and the threat of punishment are staggering and life altering.

Various scholars have discussed the means by which the size of sentencing differentials may be limited. Some have proposed requiring judges to reject plea bargains that contain significant benefits, while others have argued for “fixed discounts” for those who admit their guilt in court. Regardless of how one achieves reduced sentencing differentials, however, such reforms are key to eliminating plea-bargaining’s innocence problem and, at the same time, ensuring that plea bargains are entered into “voluntarily” by defendants as required by the Supreme Court’s great compromise of 1970.

Exactly how small sentencing differentials must be to ensure they do not exert pressure in excess of that permitted by the Brady safety-valve is still unclear and outside the scope of this Article. What is clear, however, is that the sentencing differentials being offered in a significant number of cases today are too large. As explained under the express terms of Brady, defendants whose guilt at trial is almost assured should not require large sentencing differentials to plead guilty and defendants whose culpability is questionable should not be offered such benefits as an inducement to waive their right to walk free unless proven guilty by a jury. To offer staggering sentencing differentials to all defendants violates the spirit of the

317 See id.

318 See, e.g., Covey, supra note 315, at 1241–43 (discussing the potential of using a fixed-discount system such as “plea-based ceilings” to limit sentencing differentials); Gazal-Ayal, supra note 257, at 2299–2300 (discussing the possibility of a “partial ban” that “would allow prosecutors to extract guilty pleas when defendants are almost certainly guilty, while forcing them to conduct jury trials when they bring more questionable charges”).

319 See, e.g., Gazal-Ayal, supra note 257, at 2299–2300 (“The best way to cope with the innocence problem is to allow plea bargaining only in strong cases and to ban plea bargaining in weak cases. . . . How would this partial ban work? . . . Courts only have to reject plea bargains that result in substantial concessions.”).

320 See Covey, supra note 315, at 1241 (“Fixed discounts regularize the guilty-plea process by establishing a fixed and nonnegotiable discount for pleading guilty.”). Covey goes on to discuss the advantages to the plea bargaining system that might result from limiting the size of sentencing differentials. See id. at 1245–46.

321 The author is currently engaged in research designed to determine the impact of particular sentencing differentials on defendants and ascertain how the relative size of differentials influences the likelihood that innocent defendants will plead guilty. The results of this research will be published in a future article that will draw from the definition of “voluntariness” proposed herein.
Supreme Court’s great compromise and represents an abandonment of the fundamental tenets underlying the Constitutional rights afforded to all who are accused of committing a crime.

CONCLUSION

If any number of attorneys were asked in 2004 whether Lea Fastow’s plea bargain in the Enron case was constitutional, the majority would respond with a simple word—Brady. Yet while the 1970 Supreme Court decision Brady v. United States authorized plea-bargaining as a form of American justice, the case also contained a vital caveat that has been overlooked by scholars, practitioners, and courts for almost forty years. Brady contains a safety-valve that caps the amount of pressure that may be asserted against defendants by prohibiting prosecutors from offering incentives in return for guilty pleas that are so coercive as to overbear defendants’ abilities to act freely. Further, as a means to discern whether the safety-valve fails in the future and prosecutors are offering unconstitutional incentives, the Brady Court in 1970 created a litmus test regarding innocent defendants. The Court stated that should the plea-bargaining system begin to operate in a manner resulting in a significant number of innocent defendants pleading guilty the Court would be forced to reexamine the constitutionality of bargained justice. That plea-bargaining today has a significant innocence problem indicates that the Brady safety-valve has failed and, as a result, the constitutionality of modern day plea bargaining is in great doubt.322

322 While plea-bargaining began its rise in the mid-1800s, relatively few plea bargaining cases reached the Supreme Court until the mid-1900s. There is a procedural reason for this phenomenon. Those who are coerced into accepting plea bargains, the innocent and the guilty alike, are happy merely for the chance to receive a benefit rather than risk the astounding penalties they believe await them if they lose at trial.
RECANTATIONS RECONSIDERED:
A NEW FRAMEWORK FOR RIGHTING WRONGFUL CONVICTIONS

Adam Heder* & Michael Goldsmith**

Forward
Daniel S. Medwed

It is an honor for the University of Utah to publish one of Professor Michael Goldsmith’s last research projects before his untimely death. Michael was a devoted member of the state’s criminal justice community, a first-rate scholar, a terrific teacher, and an even better human being. Generations of law students had the good fortune to study with him at Brigham Young University, and countless scholars have benefited from the wisdom contained in his scholarship. We are grateful to Adam Heder for completing this paper and, most of all, to Michael for being a mensch in all aspects of his life.

INTRODUCTION

Even the most ardent supporters of our criminal justice system must acknowledge that wrongful convictions do occur and that innocent people suffer the consequences.1 During the past decade, DNA exonerations have exposed this unfortunate truth.2 But DNA-based exonerations are likely to decrease as the

* © 2012 Adam Heder.
** Prior to the completion of this Article, Professor Michael Goldsmith passed away on November 1, 2009 after a lengthy and courageous battle with Lou Gehrig’s disease (Amyotrophic lateral sclerosis). With the permission of his gracious widow, Carolyn, and pursuant to his wishes, I have completed the Article that he and I set out to write before his passing. Prof. Goldsmith was a tremendous personal example of integrity. He gave his fullest efforts to his scholarship, his students, his teaching, and his efforts to raise money and promote awareness of Lou Gehrig’s disease in the final years of his life. He blessed the lives of all of his students, and I am fortunate to count myself as one of them.

1 See, e.g., Press Release, George H. Ryan, Governor Ryan Declares Moratorium on Executions, Will Appoint Commission to Review Capital Punishment System (Jan. 31, 2000), available at http://www.illinois.gov/PressReleases/ShowPressRelease.cfm?SubjectID=3&RecNum=359 (“I now favor a moratorium, because I have grave concerns about our state’s shameful record of convicting innocent people and putting them on death row,” Governor Ryan said. ‘And, I believe, many Illinois residents now feel that same deep reservation. I cannot support a system, which, in its administration, has proven to be so fraught with error and has come so close to the ultimate nightmare, the state’s taking of innocent life. Thirteen people have been found to have been wrongfully convicted.”).

availability of DNA evidence at the pretrial stage increases. This decline, however, threatens to mask the realities that wrongful convictions also stem from nonbiological evidence. Indeed, according to at least one study, nearly eighty-four percent of all wrongful convictions result from mistaken witness identification, false statements, and witness confusion.

Thus, as DNA exonerations decline, other potentially reliable sources acquire heightened importance in reducing and overturning wrongful convictions. In many cases, for example, recanted testimony provides compelling evidence of innocence. However, a long tradition of judicial skepticism towards recanted testimonies

(“Over the past fifteen years, the use of deoxyribonucleic acid (DNA) testing in criminal cases has helped to expose the problem of wrongful convictions in the United States, resulting in the post-conviction exoneration of 162 innocent defendants. Indeed, dozens of state legislatures have recently enacted laws implementing procedures through which prisoners may request and secure access to post-conviction testing of biological evidence available in their cases. By creating mechanisms that allow for the possibility that factually innocent prisoners may have entrée to DNA testing and thereby gain their freedom, these statutes both further the cause of justice at the individual level and bolster the institutional credibility of the criminal justice system as a whole.” (footnotes omitted)).

3 Id. at 657 (“Post-conviction innocence claims based on DNA testing therefore represent a small proportion of innocence claims generally, and this percentage is bound to diminish in the future. That is, the growing availability of DNA technology at the pre-trial stage should filter out a greater number of innocent defendants up-front and, in so doing, decrease the volume of post-conviction exonerations via DNA.” (footnotes omitted)).

4 Id. at 657–58.

5 See BARRY SCHECK, PETER NEUFELD & JIM DWYER, ACTUAL INNOCENCE: FIVE DAYS TO EXECUTION AND OTHER DISPATCHES FROM THE WRONGLY CONVICTED 42–44, 73 (2000) (examining seventy-four wrongful conviction cases); see also id. at 42 (citing HUGO MUNSTERBERG, ON THE WITNESS STAND (1908)). Munsterberg argued in 1908 that scientific evidence showed eyewitnesses were just as likely to be wrong as right. Id. Other studies showed that even the best recollection of an event erred on about twenty-six percent of significant details; the worst recollection erred on nearly eighty percent of these same details. Id. at 42–43. In a separate experiment with 2,145 participants, only 14.1 percent correctly identified the “thief” in a lineup. Id. at 43–44; see also EDWIN M. BORCHARD, CONVICTING THE INNOCENT: ERRORS OF CRIMINAL JUSTICE, at xiii (1932). Edwin Borchard, in his groundbreaking work, found witness misidentification contributed to more than half of the sixty-five cases of “completely innocent” persons convicted. Id. Borchard’s research concluded that identification of the accused by the victim or other witness as the “major source” of the conviction of innocent persons. Id. In his research, juries readily credited the “veracity and reliability of the victims of an outrage more than any amount of contrary evidence by or on behalf of the accused.” Id. (emphasis added). Victims or eyewitnesses, Borchard claimed, were “so disturbed by [their] extraordinary experience that [their] powers of perception become distorted and [their] identification most untrustworthy.” Id. In eight of the cases he studied, the accused and the real criminal bore no semblance to each other; in twelve other cases, the resemblance was fair, “but not at all close.” Id.

6 Recantations, had they been addressed, could have exonerated at least some of the defendants later cleared through DNA evidence. See Shawn Armbrust, Rerevaluating
testimony has prompted courts to summarily reject recantations as grounds for winning exoneration. Modern courts have maintained this tradition, even in the face of corroborating evidence and other mitigating factors. As a result, recanted testimony has rarely served as the basis for judicial exonerations. This Article challenges this judicial trend and proposes remedies for reform.

Two recent cases highlight the difficult judicial atmosphere surrounding witness recantations and provide glimpses of potential avenues of reform. The first concerns the murder trial of Fernando Bermudez. In 1991, Bermudez was arrested, tried, and convicted for the murder of Raymond Blount. The prosecution’s entire case rested on the testimonies of five eyewitnesses, as there was no physical evidence linking Bermudez to the crime.

Two years after trial, these same five witnesses recanted their testimony and supported Bermudez’s claim of innocence. They provided affidavits stating that police and prosecutors had manipulated and pressured them into identifying Bermudez. The state court denied even a hearing on the matter, basing its decision on the notion that recantations typically lack credibility, are often coerced, and disturb the finality of judgments.

Recanting Witnesses: Why the Red-Headed Stepchild of New Evidence Deserves Another Look, 28 B.C. THIRD WORLD L.J. 75, 75–79 (2008) (noting the story of Gary Dotson, whose accuser in a rape case recanted, but was not exonerated until DNA evidence was presented); see also id. at 87–98 (discussing a number of cases where courts erroneously discredited recanted statements, while also noting that “200-plus DNA exonerations have provided sound bases to reevaluate the current treatment of recanting witnesses.”); Medwed, supra note 2, at 657–58 (“Although post-conviction innocence claims hinging on DNA testing have captured the attention of state legislators and the broader public, the far more pervasive issue of innocence claims in cases that lack biological evidence has largely escaped notice. The same problems that led to the wrongful convictions of those innocent prisoners later freed through DNA—erroneous eyewitness identifications, false confessions, witness perjury, ineffective assistance of counsel, and the like—presumably appear in the scores of convictions procured without biological evidence. In these non-DNA cases, prisoners must find alternative means to support their innocence claims, frequently, ‘newly discovered evidence’ not susceptible to a test tube, such as confessions by the actual perpetrator, statements by previously unknown witnesses, and/or recantations by trial participants.” (citations omitted)).
decision “upon the long standing reluctance of courts to overturn a jury conviction on the basis of recantation evidence.” 14 The trial court observed, “[i]t strains credulity to believe that five unshaken trial witnesses would suddenly claim that they had testified falsely under oath.” 15 Remarkably, the judge maintained that, if Bermudez had presented fewer than five recanting witnesses, his claim would have been more credible. 16 Without addressing this anomaly, an appellate court unanimously affirmed Bermudez’ conviction, again “relying on the inherent unreliability of recantations.” 17

In 2004, Bermudez again relied on these recantations, this time to support his petition for a writ of habeas corpus in federal court. 18 Though the court held an evidentiary hearing, it was bound both by the limited nature of federal habeas review and the state court’s factual findings. 19 Despite some rather compelling corroborating evidence, 20 the court rejected the recanted testimony as a basis for habeas relief, 21 again reciting the judicial mantra that recantations are “looked upon with utmost suspicion.” 22

This did not mark the end of Fernando Bermudez’s epic legal battle. In 2009, Bermudez sought additional post-conviction relief in New York state courts, arguing that the results of an investigatory report by the district attorney’s office (which ineffective assistance of counsel, perjury by prosecution witnesses, witness recantations by criminals, investigator misconduct, and prosecutorial misconduct in withholding evidence. Id. at *1–2.

16 Id. The trial court characterized Bermudez’s motion as a “scattershot” of arguments and innuendo that sought to “locate a valid attack on the guilty verdict.” Bermudez I, 2004 U.S. Dist. LEXIS 5427, at *33–34.
17 People v. Bermudez (Bermudez III), 667 N.Y.S.2d 901, 901 (1997) (“The motion was based entirely on the affidavits of recanting witnesses, which the court properly rejected, relying on the inherent unreliability of recantations as well as the highly suspicious circumstances, viewed in context of events at the trial, under which these recantations occurred. Since the affidavits are unworthy of belief, we reject each of defendant’s various claims that rely upon facts asserted therein.” (citations omitted)).
19 Id. at *138 (“[A]t the habeas corpus hearing held in connection with the instant application for a writ, the Court had the opportunity to observe the demeanor of each of the recanting witnesses, to assess each witness’ testimony and to reach a determination about each witness’ credibility.”).
20 For example, after the trial, Bermudez’s attorney discovered the existence of a man that more fully matched the name and description of the killer given by the prosecution’s main witness. Bermudez II, 2009 WL 3823270, at *6. Even though the prosecutor’s office eventually conceded in 2009 that Bermudez was not the man implicated in their main witness’s testimony, the existence of this man was brought to the attention of the state courts well before then. Id.
22 Id. at *138–39.
revealed the existence of a man who more fully matched the description of the murderer given at trial) constituted newly discovered evidence. In a remarkable turn of events, the New York state court found the new evidence material and corroborative of the earlier recantations. Based on that, the court found Bermudez had “demonstrated his actual innocence,” vacated his conviction, and dismissed the indictment with prejudice.

The second case involves the murder trial of Troy Davis. In 1991, Davis was convicted for the 1989 murder of a police officer in Georgia. A jury convicted him largely on the basis of eyewitness accounts and without any corroborating physical evidence. Five years later, and well after Davis lost his direct appeal, seven of the nine witnesses against him recanted their testimony; all claimed the police pressured them to testify falsely. Georgia courts, however, declined to hold an evidentiary hearing to consider the recantations. In affirming this outcome, the Georgia Supreme Court relied on case law holding recantations admissible for purposes of granting a new trial only if the previous testimony was “in every material part . . . purest fabrication.” Just as the courts had done in Bermudez, the Georgia Supreme Court emphasized the recantations’ “general lack of credibility” and concluded the original trial testimony was not “purest fabrication.” After Davis succeeded in postponing his execution due to a number of last-minute appeals, the United States Supreme Court, in a very rare move, ordered the case back to the federal district court in Georgia to hold an evidentiary hearing to examine the validity and credibility of the recanted statements. Nearly

24 Id. at *8.
25 Id. at *38.
26 Davis v. State (Davis I), 426 S.E.2d 844, 845 & n.1 (Ga. 1993).
28 Davis I, 426 S.E.2d at 845 n.1, 849.
29 Davis II, 660 S.E.2d at 359–60.
30 Id.
31 Id. at 357.
32 Id. at 358 (quoting Norwood v. State, 542 S.E.2d 373, 374 (Ga. 2001)).
33 Davis II, 660 S.E.2d 354, 358 (Ga. 2008).
34 Two hours before Davis’s scheduled execution, the United States Supreme Court stayed his execution pending review of Davis’s appeal. Davis v. Georgia (Davis III), 555 U.S. 967 (2008); Tristan Smith & Elliot C. McLaughlin, Execution Put on Hold for Man Convicted in Cop’s Murder, CNN (Oct. 24, 2008, 1:47 PM), http://www.cnn.com/2008/CRIME/10/24/troy.davis.stay.execution/index.html. The Court denied certiorari, but Davis immediately filed a new habeas claim in the Eleventh Circuit. Id.
35 In re Davis (Davis IV), 130 S. Ct. 1, 1 (2009). Ultimately, Davis lost that challenge and, after a last minute appeal to the United States Supreme Court was denied, Davis was executed in Georgia by lethal injection on September 21, 2011. Colleen Curry & Michael S. James, Troy Davis Executed After Stay Denied by Supreme Court, ABC NEWS (Dec. 21, 2011), http://abcnews.go.com/US/troy-davis-executed-stay-denied-supreme-court/story?id=14571862.
fifteen years after the recanting witnesses came forward, Troy Davis finally succeeded in obtaining a hearing on that evidence.

Recantations often warrant judicial scrutiny, but these two cases reflect the judiciary’s tendency to reject recantations based more on historic skepticism, or unduly burdensome procedural barriers, than on careful consideration of whether a particular recantation is reliable. This unfortunate trend stands inapposite to a fundamental premise of Anglo-American criminal justice: “It is better that ten guilty persons escape than one innocent suffer.” This Article demonstrates that modern courts have uncritically deferred to conventional wisdom debunking recanted testimony, and it proposes a new analytical framework. This new framework acknowledges grounds for judicial skepticism but requires courts to consider whether corroborating circumstances lend credence to any particular recantation.

Part I of this Article reviews the reasons the courts have historically viewed recantations with suspicion and why some of those reasons have become outdated. Part II considers the principal procedural contexts in which recantations arise, the differing legal standards that correspond to each procedural setting, and some problems inherent with those procedures. Part III examines the Bermudez and Davis cases as a vehicle for highlighting some of the problems with judicial practice in this area. Part IV discusses various scholars’ past efforts to address this problem. Part V offers this Article’s proposed framework for analyzing recantations, in which we suggest statutory reform that could help mitigate the effects of courts’ reluctance to consider recanted testimony. Part VI applies the proposed framework to the Davis and the Bermudez cases to show how those two cases could have been handled differently.

There are many competing interests in this area of law. It is our hope that under the statutory scheme proposed in this Article, we can create a more precise system of filtering the unreliable recantations from the reliable ones, and come one step closer to our society’s ultimate goal of convicting only the truly guilty.

I. BASIS FOR JUDICIAL SKEPTICISM

One commentator has observed that judicial skepticism of recantations “has become so universal that it appears to have given rise to an inference that recantation evidence is not trustworthy and should be treated as such absent the movant’s ability to persuade otherwise.” That courts mistrust recantations can hardly be doubted. As the New York Court of Appeals bluntly stated: “There is

36 See infra Part I.
37 CHRISTIANSON, supra note 9, at 17 (quoting 4 WILLIAM BLACKSTONE, COMMENTARIES 358 (1765–1769)).
39 See, e.g., Armbrust, supra note 6, at 82 (noting the extreme suspicion with which courts view recantations). Indeed it can hardly be doubted that this skepticism is virtually
no form of proof so unreliable as recanting testimony. In the popular mind it is often regarded as of great importance. Those experienced in the administration of the criminal law know well its untrustworthy character.”

Such reluctance to credit recantations stems from a variety of legal and policy considerations that can generally be grouped into two categories: (1) lack of trustworthiness and (2) the principle of finality.

universal amongst both federal and state courts. See, e.g., United States v. Provost, 969 F.2d 617, 619–20 (8th Cir. 1992) (holding that statements made under oath when pleading guilty are conclusive, and the judge may reject without a hearing a defendant’s later contention that his sworn statements were untrue); United States v. DiPaolo, 835 F.2d 46, 49 (2d Cir. 1987) (“Courts are particularly reluctant to grant such motions [for new trials based on newly discovered evidence] where the newly discovered evidence consists of a witness recantation as such recantations are ‘looked upon with the utmost suspicion.’” (citation omitted)); United States v. Ramsey, 726 F.2d 601, 605 (10th Cir. 1984) (“[R]ecanted testimony is properly viewed with suspicion.”); Lemire v. McCarthy, 570 F.2d 17, 21 (1st Cir. 1978) (“[R]ecantations are viewed with considerable skepticism.”); United States v. Mackin, 561 F.2d 958, 961 (D.C. Cir. 1977) (“[R]ecantations by witnesses for the prosecution are viewed with suspicion.”); United States v. Johnson, 487 F.2d 1278, 1279 (4th Cir. 1973) (citing United States v. Lewis, 338 F.2d 137, 139 (6th Cir. 1964)) (“Where a motion for a new trial is based upon recantation of testimony given at the trial, such recantation is ‘looked upon with the utmost suspicion.’”). See generally Tim A. Thomas, Annotation, Recantation of Testimony of Witnesses as Grounds for New Trial—Federal Criminal Cases, 94 A.L.R. Fed. 60, §1(a) (1989) (listing numerous examples of judicial reluctance to recanted testimony). This skepticism has led to a large burden for any defendant wishing to introduce recantations post-trial as a means of exoneration. Consequently, both federal and state courts have erected high evidentiary standards through which an appealing defendant must pass. See generally R.P.D., Annotation, Statements by a Witness after Criminal Trial Tending to Show That his Testimony was Perjured, as Grounds for New Trial, 158 A.L.R. 1062 (1945) (discussing standards and applicable cases). In short, courts disfavor recantations and afford them virtually no weight upon review.

41 See Sharon Cobb, Comment, Gary Dotson as Victim: The Legal Response to Recanting Testimony, 35 EMORY L.J. 969, 981–982 (1986) (discussing the roots of “[t]his distrustful attitude toward recantations”); see also Armbrust, supra note 6, at 82 (noting the extreme suspicion with which courts view recantations). Armbrust outlines the reasons for this suspicion as follows:

1. the perception that any witness who recants is untrustworthy,
2. determinations about the demeanor of the witness during an evidentiary hearing,
3. findings that the other evidence in the case supports the initial guilty verdict,
4. fears that the witness has recanted under duress or because of coercion,
5. close relationships between defendants and witnesses,
6. a desire for finality and concerns about judicial economy, and
7. the desire to prevent the manipulation of courts.

Id.
A. Lack of Trustworthiness

Judges have historically viewed recantations as inherently unreliable based on several principles that reflect common sense and prior experience. First is the intuitive notion that witnesses who have perjured themselves cannot be trusted.42

Second, recanting witnesses often have ample “incentives” to change their stories.43 Bribery, coercion, threats, and other impropriety are among the incentives most often cited.44

Finally, courts have found recantations especially untrustworthy when a close relationship exists between the defendant and the recanting witnesses; depending on the circumstances, such relationships often prompt an inference of intimidation or coercion45 or foster a protective instinct.

B. Preserving the Principal of Finality

The principle of finality is fundamental to every system of justice, and all courts have a strong interest in preserving the integrity of their verdicts and conserving scarce judicial resources.46 With the passage of time, witnesses become

---

42 As the Second Circuit observed in upholding a trial court’s denial of a new trial motion based on recanted statements, the judge “had no reason to believe that on one occasion more than another he was telling the truth.” United States v. Troche, 213 F.2d 401, 403 (2d Cir. 1954).

43 See, e.g., United States v. Johnson, 487 F.2d 1278, 1279 (4th Cir. 1973) (listing “coercion, bribery, or misdealing” as common means of obtaining recanted testimony).

44 See United States v. Goodwin, 496 F.3d 636, 641 (7th Cir. 2007) (noting that inmate composed the recanting affidavit under threat of coercion and later withdrew it); United States v. Leibowitz, 919 F.2d 482, 483 (7th Cir. 1990) (describing affidavit as “not worthy of belief” in light of affiant’s psychological frailty and the threats likely communicated by defendant); see also Armbrust, supra note 6, at 82; Cobb, supra note 41, at 983–87. This view is so widespread it led one commentator to remark that courts view recantations “as untrustworthy, acts not of conscience, but of sympathy or bribery or coercion.” Zielbauer, supra note 15, at A1. In a particularly illustrative case, a man convicted for rape appealed to the Minnesota Supreme Court the trial court’s denial of a new trial based on the recanted testimony of his accuser. After the trial, the prosecution’s witness received a number of threats, some over the phone and others through the intimidating behavior of some of the convicted’s acquaintances. Cobb, supra note 41, at 983–84 (citing State v. Hill, 253 N.W.2d 378, 383–84 (Minn. 1977)). The court remarked, “[c]ourts have traditionally looked with disfavor on motions for a new trial founded on alleged recantations unless there are extraordinary and unusual circumstances. This rule is particularly relevant where possible changes in testimony have been occasioned by threats, pressure, and intimidation.” State v. Hill, 253 N.W.2d 378, 384 (Minn. 1977) (citations omitted). In light of the evidence of coercion and intimidation, the court found the recanted statements untrustworthy and did not reverse the trial court’s denial of a new trial. Id.

45 Armbrust, supra note 6, at 82 (noting that courts often distrust recantations because of the close relationships between defendants and the recanting witnesses).

unavailable, memories fade, and evidence is no longer preserved. These factors often preclude retrials or render them highly problematic. For these reasons, the Supreme Court of the United States has stated that “[d]ue process does not require that every conceivable step be taken, at whatever cost, to eliminate the possibility of convicting an innocent person.” Otherwise, chaos and systemic paralysis would ensue. These concerns are especially salient in recantation cases, given the judiciary’s distrust of such testimony.

C. Problems with this Judicial Suspicion

Judicial reluctance to credit recantations reflects very real concerns. The difficulty, however, is that not all recantations are unreliable. Simply because many recantations are false does not preclude the possibility that some recantations are true. Yet the judiciary’s historic concerns lead to dismissals even when certain factors are present that lend credence to a particular recantation.

There are good reasons to reject broad application of such judicial skepticism of recantations. There is often no justifiable reason to presume a recantation unreliable. For instance, in one case, the recanting witness came forward years after the trial and, under threat of perjury, testified that he was “tired of lying and did not want another man’s blood on his hands.” In that case, the recanter was convicted on charges arising from the same conduct and was also on death row. The state court noted the recanter’s interest in the case, stating that “[i]f this was not a planned murder but simply a robbery gone bad, then [the recanter] is probably not eligible for the death penalty, so his recantation could potentially affect the imposition of the death penalty . . . .” On the other hand, the chances that the recantation would have actually led to a reduction in the recanter’s sentence were small and the recanter may, as he claimed, have been motivated simply by guilt. In sum, it would have been difficult, relying on nothing more than

---

47 See, e.g., Herrera v. Collins, 506 U.S. 390, 417 (1993) (noting that the need to protect finality of judgments is linked to the burdens of litigating with stale evidence).
48 Id. at 399 (alterations in original) (quoting Patterson v. New York, 432 U.S. 197, 208 (1977)).
49 Id.
50 See discussion supra note 7.
51 See infra Part III.
53 Id. at *29.
the inherent unreliability of the recantation, to conclude that the statements were
ingcredible on their own terms.54

In his article on this topic, Shawn Armbrust agrees with the premise that a
presumption against recantations can be problematic and lead to questionable
results.55 He gives four basic reasons. First, he discusses the various cognitive
biases that affect judges, noting past psychology studies that suggest judges are not
likely to find their own previous decisions false.56 Second, he notes that recanting
witnesses are often those that were incentivized to testify against the defendant in
the first place (such as a jailhouse informant testifying in exchange for a reduced
sentence), and because such testimony is unreliable in the first place, it makes little
sense to presume that the recantation is equally unreliable.57 Third, he notes that
police and others are sometimes guilty of coercing witness statements.58 Therefore,
it is illogical to presume that recantations are always the product of defendant
coercion and not simply the consequence of earlier police coercion.59 Finally,
Armbrust remarks that experience shows “trial judges are not always adept at
judging the demeanor of recanting witnesses.”60 He thus concludes, and we concur,
that there are often just as many reasons to trust a recantation as there are to
mistrust one.61

Moreover, reliance on judicial mistrust is not always an altogether logical
approach to recantations. For all the reasons just discussed, there are good reasons
to mistrust recantations. Broad invocation of those reasons to dismiss a recantation,
however, ignores logical flaws in the historical approach. Examination of Rule
804(b)(3) of the Federal Rules of Evidence provides a helpful illustration on this
point. Rule 804(b)(3) governs the admissibility of hearsay statements made against
the declarant’s penal interests (i.e., statements against interest). Historically,
hearsay statements, like recantations, are presumed invalid and untrustworthy.62
Statements made against the declarant’s penal interests were no exception. As the
advisory notes to the Rule 804(b)(3) state, the common law distrusted “evidence of
confessions by third persons offered to exculpate the accused.”63 Courts suspected

54 The district court did hold an evidentiary hearing in that case and based its decision
on evidence brought out at the adversarial proceeding. Id. at *30 (“Based on the Court’s
experience, common sense, and personal observations of Patrick Bonifay [the recanter], the
Court is satisfied that this new testimony is false. After listening to Mr. Bonifay, observing
his demeanor, and analyzing his testimony, the Court does not believe his recantation.”).
55 See Armbrust, supra note 6, at 89–98.
56 Id. at 89–91 (relying heavily on Medwed, supra note 2, at 700–08).
57 Id. at 91–94.
58 Id. at 94–97.
59 Id.
60 Id. at 97.
61 See id. at 98–102, 104.
62 See, e.g., Fed. R. Evid. 802 (“Hearsay is not admissible unless any of the following
provides otherwise: a federal statute; these rules; or other rules prescribed by the Supreme
Court.”).
63 Id. at 804 advisory committee’s note to subdivision (b).
defendants could fabricate both the confession and its contents and then rely on the unavailability of the witness to more easily admit the fabricated hearsay evidence.

Nevertheless, the Federal Rules of Evidence allow statements made against the declarant’s penal interest to be admitted (subject to several prerequisites, which are discussed infra) upon the “assumption that persons do not make statements which are damaging to themselves unless satisfied for good reason that they are true.” Thus, notwithstanding historical mistrust of such statements against interest, jurists now recognize that such statements can, and often do, bear indicia of reliability which may justify their occasional admission.

Recantations often bear these same indicia of reliability. Recanting witnesses expose themselves to perjury charges by admitting that they lied at trial. And just as courts presume that individuals will not make damaging statements about themselves “unless satisfied for good reason that they are true,” they could similarly conclude that witnesses will not recant and expose themselves to perjury charges unless the recantations are true.

The challenge, of course, is to develop a standard for assessing recantation reliability that will not unduly impinge upon the principle of finality. In short, courts need a more precise filtering system for distinguishing between recantations involving wrongful convictions and those suffering from classic defects of unreliability. Before delving into our proposal, however, it is necessary to detail the procedures surrounding recantations and how those processes further complicate defendants’ efforts in this regard.

II. PROCEDURAL CONTEXTS

Standards governing judicial review of recanted testimony depend upon when the defense first learns of the recantation. The timing of this event determines the corresponding procedure for raising this issue, which, in turn, dictates the applicable standard of review. This section reviews the role of timing in determining the procedures and standards that govern this issue.

In the typical case, a criminal judgment becomes final when the jury has rendered a verdict and the defendant has exhausted his direct appeals.
Recognizing, however, that defendants might discover exculpatory evidence after the fact, most jurisdictions (though not all) allow defendants a period of time to move for a new trial. \(^{70}\) Once the applicable timeframe has passed, the process of introducing recanted testimony becomes much more difficult. In those cases, defendants have only limited opportunities to introduce the newly discovered evidence through state and federal collateral remedies. \(^{71}\) Thus, the nature of the remedy available to criminal defendants depends on whether they have made their motion for a new trial within the applicable timeframe.

A. Motions for Newly Discovered Evidence

When a recantation comes to a defendant’s attention, he may file a motion for a new trial based on newly discovered evidence (unless the respective jurisdiction does not allow such motions). \(^{72}\) Federal and state criminal justice systems employ a variety of statutory and common-law standards for considering such motions. \(^{73}\) For example, federal law requires the motions be filed within three years of a verdict and provides that “[u]pon the defendant’s motion, the court may vacate any judgment and grant a new trial if the interest of justice so requires.” \(^{74}\) Both state and federal courts generally derive their standards from two leading cases, Berry v. State \(^{75}\) and Larrison v. United States. \(^{76}\) With slight variations, the Berry and

---

\(^{70}\) See 23A C.J.S. Criminal Law § 1931 (2006); see also discussion infra notes 92–95.

\(^{71}\) See 6 WAYNE R. LAFAVE, JEROLD H. ISRAEL & NANCY J. KING, CRIMINAL PROCEDURE § 28.1(a) (2d ed. 1999) (describing the nature of collateral remedies and the typical situations when they are sought).

\(^{72}\) See Janice J. Repka, Comment, Rethinking the Standard for New Trial Motions Based Upon Recantations as Newly Discovered Evidence, 134 U. PA. L. REV. 1433, 1438 (1986) (“[N]ew trials can now be sought in nearly all jurisdictions on the ground of newly discovered evidence.”); see also Armbrust, supra note 6, at 80 (“The primary vehicle for a defendant with newly discovered evidence of innocence is a motion for new trial. In some states, newly discovered evidence also can be brought to a court’s attention through collateral attack proceedings. This article focuses primarily on motions for a new trial because they are the primary vehicle for litigating newly discovered evidence claims and are conceptually similar to collateral attack proceedings and, therefore, can be treated similarly.”).

\(^{73}\) Thomas, supra note 39, § 2 (discussing various standards set forth by federal courts to determine whether a new trial is to be granted in a criminal case on the ground of a recantation by a witness of his or her trial testimony); see, e.g., Drake v. State, 287 S.E.2d 180, 182 (Ga. 1982) (employing the six part test first employed in Berry v. State); State v. Prudholm, 446 So.2d 729, 735 (La. 1984) (employing a slightly different four part test).

\(^{74}\) FED. R. CRIM. P. 33(a).

\(^{75}\) 10 Ga. 511, 512–13 (1851) (“It is pretty generally agreed, that it is incumbent on a party who asks for a new trial, on the ground of newly discovered evidence, to satisfy the Court, 1st. That the evidence has come to his knowledge since the trial. 2d. That it was not
Larrison tests focus principally on whether the newly discovered evidence would have produced a different result if it had been presented at trial.  

B. Collateral Review

If a defendant’s motion for a new trial on the basis of newly discovered evidence fails, or if a defendant has not come forward with the newly discovered evidence until after the applicable timeframe has passed, defendants are typically limited only to whatever collateral remedies exist in the respective jurisdiction. Every state offers some form of collateral relief to convicted criminals, though the

owing to the want of due diligence that he did not acquire it sooner. 3d. That it is so material, that it would probably produce a different verdict. 4th. That it is not cumulative only, viz.: speaking to facts in relation to which there was evidence on the trial. 5th. The affidavit of the witness himself should be procured or its absence accounted for; and 6th. The new trial will not be granted if the only effect of the evidence will be to impeach the credit of a witness.”).

76  24 F.2d 82, 87–88 (7th Cir. 1928) (“We shall approach this question on the assumption that a new trial should be granted when, (a) The court is reasonably well satisfied that the testimony given by a material witness is false. (b) That without it the jury might have reached a different conclusion. (c) That the party seeking the new trial was taken by surprise when the false testimony was given and was unable to meet it or did not know of its falsity until after the trial.”).

Virtually all state and federal courts adopt one of these two standards or some derivative thereof. Repka, supra note 72, at 1439–40 (noting that most states have adopted one of the Berry or Larrison tests or some slight variation of one of them).

77 Wolf, supra note 46, at 1926 (“[T]he vast majority of courts have steadfastly adhered to the long-established rule that newly discovered evidence will not entitle the accused to a new trial unless that evidence would probably produce a different verdict.”); see, e.g., United States v. Lee, 68 F.3d 1267, 1273 (11th Cir. 1995) (“A new trial is warranted based upon circumstances coming to light after trial only if the following five part test is satisfied: (1) the evidence was in fact discovered after trial; (2) the defendant exercised due care to discover the evidence; (3) the evidence is not merely cumulative or impeaching; (4) the evidence was material; and (5) the evidence was of such a nature that a new trial would probably produce a different result.”); see also Evenstad v. Carlson, 470 F.3d 777, 783 n.6 (8th Cir. 2006) (“Most circuits, including this one, absent a finding the government knowingly sponsored false testimony, require a petitioner seeking a new trial to show the jury would have ‘probably’ or ‘likely’ reached a different verdict had the perjury not occurred. Other circuits, like the Minnesota courts, apply a ‘possibility’ standard granting relief whenever the discovery ‘might’ have produced an acquittal.” (citations omitted)); United States v. Rodriguez-Marrero, 390 F.3d 1, 14 (1st Cir. 2004) (“A motion for new trial on the basis of newly discovered evidence will ordinarily not be granted unless the moving party can demonstrate that: (1) the evidence was unknown or unavailable to the defendant at the time of trial; (2) failure to learn of the evidence was not due to lack of diligence by the defendant; (3) the evidence is material, and not merely cumulative or impeaching; and (4) [the evidence] will probably result in an acquittal upon retrial of the defendant.” (quoting United States v. Wright, 625 F.2d 1017, 1019 (1st Cir. 1980))).
rules involved differ slightly in some states. In federal court, the writ of habeas corpus constitutes the foremost collateral remedy. An analysis of that writ, therefore, provides a helpful tool for illustrating the procedural challenges facing a defendant attempting to introduce newly discovered evidence via a collateral challenge.

As a general matter, habeas relief is available only for defendants who can establish a constitutional error with their convictions. In 1996, Congress enacted the Anti-Terrorism and Effective Death Penalty Act (“AEDPA”), which limited the scope of habeas corpus by narrowing the constitutional claims cognizable in a federal habeas proceeding. AEDPA precludes habeas relief to persons whose claims were decided on the merits in state court, unless the judgment

---

78 See LAFAVE, ISRAEL & KING, supra note 71, § 28.1(a).
79 There is the additional remedy of the writ of Coram Nobis. Initially developed as vehicle for correcting errors of fact, it has largely been “displaced by the motion for a new trial based on newly discovered evidence, and an expanded writ of habeas corpus.” Id. § 28.1(c). It has mostly “fallen into disuse.” Id. Nonetheless, it remains a possibility for defendants. This Article does not delve into the writ of Coram Nobis because of its present limited utility and also because the other procedures outlined in this Article effectively cover any instance in which the writ could be raised. Additionally, this Article’s proposals do not hinge upon the availability of the writ of Coram Nobis. See id. §28.9(a) (“The Court’s opinion in Morgan suggested three prerequisites to coram nobis relief. First, the writ is only an option if § 2255 or other relief is not available or adequate, ruling out most claims by prisoners ‘in custody’ who are able to seek relief under § 2255. The writ of coram nobis remains useful, however, when the defendant is not ‘in custody.’ Second, the Court noted that coram nobis is available for ‘errors of the most fundamental character,’ and held that petitioner’s sixth amendment claim was such a claim. The Court later explained in Carlisle v. United States that the writ was ‘traditionally available only to bring before the court factual errors ‘material to the validity and regularity of the legal proceeding itself,’ such as the defendants being under age or having died before the verdict.’ Third, drawing upon the Court’s statement in Morgan that ‘sound reasons exist[ed] for [the petitioner’s] failure to seek appropriate earlier relief,’ lower courts have insisted that petitioners seeking coram nobis must exercise ‘reasonable diligence’ in seeking prompt relief. An additional requirement that has been adopted by most circuits is that an applicant must demonstrate ‘lingering civil disabilities’ from a conviction. Given this daunting array of prerequisites, the Court understandably concluded in Carlisle that ‘it is difficult to conceive of a situation in a federal criminal case today where [a writ of coram nobis] would be necessary or appropriate.’” (emphasis added) (citations omitted)); see also Repka, supra note 72, at 1437 (“The highly technical writ of coram nobis, which lies to correct errors of fact not apparent in the record that would have prevented judgment had they been known, proved to be ineffective because most courts rejected its application in the recantation context. The writ of habeas corpus also proved to be an impotent remedy in this area, as the courts held that it would not be issued unless state officials ‘knowingly’ used the perjured testimony.”).

80 See 28 U.S.C. § 2241(c)(3) (2006); see also Stone v. Powell, 428 U.S. 465, 475–77 (1976) (noting the general history of habeas corpus and recognizing that its overarching purpose has been to correct constitutional errors in criminal convictions).
(1) resulted in a decision that was contrary to, or involved an unreasonable application of, clearly established Federal law, as determined by the Supreme Court of the United States; or (2) resulted in a decision that was based on an unreasonable determination of the facts in light of the evidence presented in the State court proceeding.  

The Supreme Court has interpreted this statute as allowing federal habeas relief if one of two circumstances has been met. First, federal courts may grant the writ if the “state court arrives at a conclusion opposite to that reached by [the Supreme Court] on a question of law or if the state court decides a case differently than [the Supreme Court] has on a set of materially indistinguishable facts.”  

Second, federal courts may grant the writ if the “state court identifies the correct governing legal principles from [the Supreme Court’s] decision but unreasonably applies that principle to the facts of the prisoner’s case.”

Therefore, if courts have denied a defendant’s motion for a new trial, the defendant may seek habeas corpus only if the state court’s denial of the motion either (1) produced a result inapposite to one reached by the United States Supreme Court or (2) the state court’s application of the law was unreasonable. This is significant, because a freestanding claim of innocence based on newly discovered evidence may or may not involve a constitutional challenge. As the Supreme Court explained:

Claims of actual innocence based on newly discovered evidence have never been held to state a ground for federal habeas relief absent an independent constitutional violation occurring in the underlying state criminal proceeding. . . . This rule is grounded in the principle that federal habeas courts sit to ensure that individuals are not imprisoned in violation of the Constitution—not to correct errors of fact.

---

81 28 U.S.C. § 2254(d)(1)–(2). Thus, a much broader range of constitutional claims were cognizable in a federal habeas court before passage of the AEDPA. The statute has now limited habeas claims to only those decisions that were “contrary to, or involved an unreasonable application of, clearly established Federal Law, as determined by the Supreme Court of the United States.” Id. The Supreme Court has offered narrowing interpretations of the statute in subsequent decisions. See Williams v. Taylor, 529 U.S. 362, 412–13 (2000).

82 Williams, 529 U.S. at 413.

83 Id. at 413.

84 Herrera v. Collins, 506 U.S. 390, 400 (1993); see, e.g., Johnson v. Bett, 349 F.3d 1030, 1038 (7th Cir. 2003) (“We have said that the ‘refusal to grant a new trial on the basis of newly discovered evidence is not actionable in habeas corpus.’”’ (quoting Guinan v. United States, 6 F.3d 468, 470 (7th Cir. 1993))); Coogan v. McCaughtry, 958 F.2d 793, 801 (7th Cir. 1992) (“Where the ‘newly discovered evidence’ consists of witness recantations of trial testimony or confessions by others of the crime, most courts decline to consider it in the absence of any showing that the prosecution knowingly proffered false testimony or failed to disclose exculpatory evidence, or that petitioner’s counsel was ineffective.”). Of course there might also be the writ of coram nobis, though this remedy is
Thus, absent constitutional grounds, habeas relief is not ordinarily available to challenge a state court’s denial of a new trial based on recanted testimony.\(^{85}\)

This is not to say criminal defendants may never be able to tether their recantations to a constitutional violation. For example, witness recantations could accompany allegations of police coercion, as in the *Bermudez* and *Davis* cases.\(^{86}\) Such allegations lend themselves to claims that the prosecution either knowingly proffered false testimony or failed to disclose exculpatory evidence—both of which qualify as constitutional claims.\(^{87}\) These are well-established constitutional principles, but require proof beyond the recantation itself. Other less well-established grounds based on due process and ineffective assistance of counsel claims also exist.\(^{88}\) In any event, because a freestanding claim of innocence based

typically covered by habeas proceedings and has fallen into disuse. See discussion supra note 79.

\(^{85}\) The Supreme Court has suggested only one possible exception to this rule:

We may assume, for the sake of argument in deciding this case, that in a capital case a truly persuasive demonstration of ‘actual innocence’ made after trial would render the execution of a defendant unconstitutional, and warrant federal habeas relief if there were no state avenue open to process such a claim. . . . [T]he threshold showing for such an assumed right would necessarily be extraordinarily high. *Herrera*, 506 U.S. at 417. The Court, therefore, has suggested that if a defendant can meet this “extraordinarily high” standard, habeas relief would be warranted. Despite this language in *Herrera v. Collins*, the Court did not decide whether a freestanding claim of innocence constitutes grounds for federal habeas relief. *Id.* Even if it did, however, *Herrera* claims would rarely succeed given the “extraordinarily high” showing required. See *LAFAVE, ISRAEL & KING*, supra note 71, § 28.3(f) (“The chances that such a [*Herrera*-type] claim will succeed, however, are extremely small, and the questions such a claim raises are numerous and daunting.”).


\(^{87}\) United States v. Agurs, 427 U.S. 83, 103 (1976) (relying on the rule articulated first in *Brady v. Maryland*, 373 U.S. 83 (1963), the Court stated, “In a series of subsequent cases, the Court has consistently held that a conviction obtained by the knowing use of perjured testimony is fundamentally unfair, and must be set aside if there is any reasonable likelihood that the false testimony could have affected the judgment of the jury” (footnotes omitted)); see also *Coogan*, 958 F.2d at 801.

\(^{88}\) Other defendants have argued that the courts’ refusal to grant new trials may “be a violation of the fundamental fairness embodied in the Due Process Clause” when the new “evidence is [sufficiently] compelling.” *Moore v. Casperson*, 345 F.3d 474, 491 (7th Cir. 2003) (“Mr. Moore [defendant] claims that the failure to grant him a new trial denied him due process of law in light of the new evidence of Kellner’s recantation.”). Other defendants have linked the delay in discovery of the recantations to the ineffective assistance of their counsel. See, e.g., *Coogan*, 958 F.2d at 801 (noting that an ineffective assistance of counsel claim should accompany a witness recantation). Though this list is
on a witness recantation rarely warrants habeas relief, the “Great Writ” does not stand out as an altogether effective means of obtaining postconviction relief on the basis of witness recantations.89

Where a jurisdiction does not allow motions for a new trial on the basis of newly discovered evidence or where the defendant fails to move for a new trial within the relevant timeframe, these procedural difficulties become even more acute. The general rule holds that a defendant may seek federal habeas relief only if he is not procedurally barred from doing so in state or federal court.90 Of course, a defendant who fails to introduce a witness recantation within the applicable timeframe or is simply disallowed by the respective jurisdiction from making such a motion altogether would be procedurally barred from pursuing habeas relief in a federal court. Though this rule is subject to several exceptions,91 the fact remains:

not exhaustive, it represents those constitutional claims most typically associated with witness recantations. See also Brief for Innocence Project et al. as Amici Curiae Supporting Petitioner, In re Davis, 565 F.3d 810 (11th Cir. 2008) (No. 08-16009) (arguing that in the face of compelling evidence, execution of an “innocent” man would violate the Eighth Amendment).

89 See supra note 85 and accompanying text.

90 See, e.g., Cone v. Bell, 129 S. Ct. 1769, 1780 (2009) (“Therefore, consistent with the longstanding requirement that habeas petitioners must exhaust available state remedies before seeking relief in federal court, we have held that when a petitioner fails to raise his federal claims in compliance with relevant state procedural rules, the state court’s refusal to adjudicate the claim ordinarily qualifies as an independent and adequate state ground for denying federal review.”).

91 The first exception is the “cause and prejudice” exception, and the second is the “fundamental miscarriage of justice” exception. The United States Supreme Court summarized the “cause and prejudice” standard:

In all cases in which a state prisoner has defaulted his federal claims in state court pursuant to an independent and adequate state procedural rule, federal habeas review of the claim is barred unless the prisoner can demonstrate cause for the default and actual prejudice as a result of the alleged violation of federal law.

Coleman v. Thompson, 501 U.S. 722, 724 (1991) (emphasis added). Subsequent case law has further refined and formulated the meanings of both “cause” and “prejudice.” See United States v. Frady, 456 U.S. 152, 167–70 (1982). This exception allows a defendant with a procedurally defaulted claim to nonetheless seek habeas relief if the default was caused by state interference. Murray v. Carrier, 477 U.S. 478, 488 (1986). The Court held that there must be a showing of interference by state officials that made compliance impracticable. Id. Such a showing would constitute “cause” for purposes of this exception. Id. After proving such interference, the defendant proceeds to show a “reasonable probability” that the verdict would have been different absent the constitutional violation. LAFAVE, ISRAEL & KING, supra note 71, § 28.4(f) (“In recent years, the standard most commonly applying in these cases requires a defense showing of a ‘reasonable probability’ that the outcome would have been different if not for the alleged constitutional violation. A ‘reasonable probability’ is described as ‘a probability sufficient to undermine confidence in
Thus, a defendant could conceivably obtain habeas review on the basis of newly discovered recantations only if he could show that the police or other state officers withheld exculpatory information or prevented witnesses from testifying sooner.

Recantations more commonly arise in the context of the second exception—the fundamental miscarriage of justice. This doctrine applies when the evidence suggests that an innocent defendant remains incarcerated or faces capital punishment. Carrier, 477 U.S. at 496 (“[I]n an extraordinary case, where a constitutional violation has probably resulted in the conviction of one who is actually innocent, a federal habeas court may grant the writ even in the absence of a showing of cause for the procedural default.”). The United States Supreme Court has clarified this fundamental miscarriage of justice/actual innocence doctrine through two important cases, Herrera v. Collins, 506 U.S. 390 (1993), and Schlup v. Delo, 513 U.S. 298 (1994). In Herrera, the court stated that an actual innocence claim was “a gateway through which a habeas petioner must pass to have his otherwise barred constitutional claim considered on the merits.” 506 U.S. at 404. If a defendant, therefore, has compelling evidence that he is actually innocent but is procedurally barred from pursuing habeas relief, that defendant may invoke the “actual innocence” exception and obtain a federal habeas hearing. Importantly, the defendant must, even if he has shown “actual innocence,” still allege a constitutional violation in order to obtain habeas relief. In Herrera, the defendant asserted only a claim of factual innocence without tying it to any alleged constitutional violation. Id. at 404–05. As a result, he failed to properly invoke habeas jurisdiction. Id.

In Schlup, the Court held that where a defendant’s claim of actual innocence was based on an ineffective assistance of counsel claim rather than on mere innocence, habeas relief was available. 513 U.S. at 314. The defendant, like in Herrera, was procedurally barred from seeking habeas relief, but argued that in light of compelling evidence of “actual innocence” he should be able to argue his otherwise procedurally barred constitutional claims. Id. Whereas Herrera lacked a constitutional element to his claim, the defendant in Schlup applied the “fundamental miscarriage of justice” doctrine precisely by using his claim of actual innocence merely as a gateway to litigate otherwise barred constitutional claims (i.e., a claim of ineffective assistance of counsel). Id. at 314–15. The Court formulated the standard for considering new evidence: the defendant must demonstrate that “it is more likely than not that no reasonable juror would have convicted him in the light of the new evidence.” Id. at 327. Assuming the defendant can do this, habeas relief becomes available and the court will consider the constitutional claims.

These doctrines have changed to some degree as a result of the 1996 AEDPA. Section 2264 of the AEDPA lays forth the standard to be applied by habeas courts in capital cases. 28 U.S.C. § 2264 (2006). In cases where a claim is procedurally defaulted, habeas courts are barred from considering the habeas petition unless

the failure to raise the claim properly is—(1) the result of State action in violation of the Constitution or laws of the United States; (2) the result of the Supreme Court’s recognition of a new Federal right that is made retroactively applicable; or (3) based on a factual predicate that could not have been discovered through the exercise of due diligence in time to present the claim for State or Federal post-conviction review.
Obtaining collateral review of recanted statements is difficult, and often impossible.

C. Practical Problems with the Procedural Schemes

Our examination of the procedures in this realm demonstrates that a substantial portion of defendants asserting recantations will never, indeed can never, have a day in court. First, not all jurisdictions within the United States allow for postconviction relief on the basis of newly discovered evidence, while others impose strict time limits. Though many of these jurisdictions may have some other mechanism available, criminal defendants in those jurisdictions nevertheless have either no or limited initial procedural mechanisms for introducing newly discovered evidence like witness recantations. Aside from the harshness of

Id. The first exception is essentially a rearticulation of the “cause and prejudice” standard. The third exception bears particular importance to petitioners presenting witness recantations. Many witnesses do not decide to recant until well after the statutory time period has expired. Criminal defendants facing this situation are not barred from presenting that evidence to the habeas court. Though again, a defendant may present evidence only to show a constitutional violation, not to show factual innocence.

Significantly, this specific section of the AEDPA, which applies only to convicts facing the death sentence, does not provide a statutory equivalent of the “fundamental miscarriage of justice” exception to the procedural default rule. Therefore, a convict facing the death penalty whose claims are procedurally defaulted has available to him only the three exceptions outlined above, and not the wider “fundamental miscarriage of justice” exceptions embodied in the Schlup rule. Very likely Congress intended this exact result, since Congress referred to the rule in other provisions of the statute. See 28 U.S.C. § 2244(b)(2)(B); see also LAFAYE, ISRAEL & KING, supra note 71, § 28.4(i) (“Perhaps the most significant feature of § 2264(a) is its failure to specifically take into account the impact of the alleged constitutional violation that was procedurally defaulted claim include an exception for a ‘miscarriage of justice.’ The absence of any ‘actual innocence’ component is especially striking since Congress included an actual innocence exception to two other provisions elsewhere in the Act, both of which would otherwise bar relief in cases in which only a petitioner’s liberty, not his life may be at stake.” (citation omitted)). As a result, in other noncapital criminal cases, the AEDPA does not change either the “cause and prejudice” or the “fundamental miscarriage of justice” exceptions to the rule prohibiting procedurally defaulted claims. See Id. (“Since the 1996 Act did not add a provision similar to §2264 to the statutory sections governing other types of cases, federal courts presumably will continue to apply the . . . cause and prejudice standard and its ‘fundamental miscarriage of justice’ exception to procedurally defaulted claims in all cases other than those regulated by §2264.”).


94 See Medwed, supra note 2, at 676–78 (“[M]any time limits governing motions for a new trial on the grounds of newly discovered evidence are remarkably brief. As a result,
entirely arbitrary deadlines, there is simply no reason to presume a recanter would come forward shortly after the trial instead of some significant amount of time later.\textsuperscript{95}

Finally, habeas relief is not generally available to those asserting freestanding claims of innocence based on recantations.\textsuperscript{96} Only where the defendant has tethered the recantation to a constitutional claim is he entitled to federal habeas review. Thus, where a defendant may only seek collateral review of a recantation, in the absence of a constitutional error, his efforts will generally be unavailing. But there is no reason to presume that a defendant who can tether his recantation to a constitutional argument has a more compelling case of innocence than the defendant who cannot.

these remedies are of limited utility to the bulk of criminal defendants who, in the immediate aftermath of their convictions, might not have the resources or the good fortune to find new evidence. According to Wilkes, litigants in seventeen states must file their motions within sixty days of judgment, and one state still follows the common law rule that motions for a new trial can only be filed during the court term in which the original judgment was entered. Seventeen other states and the District of Columbia have time restrictions on new trial motions spanning from one to three years. Of the remaining states, six allow new trial motions to be filed beyond three years after conviction, with four of those jurisdictions boasting waivable time limits of less than 120 days. A spare number of jurisdictions—nine—have no limitations period whatsoever. Also, many of the states that impose time limits on new trial motions view those conditions as jurisdictional in nature, barring trial courts from granting tardy motions even when such motions are based on newly discovered evidence. Some states have carved out exceptions to general statutory time limits when new trial motions involve newly discovered evidence. Similarly, if the alleged new evidence could not have been discovered prior to the end of the limitations period, several jurisdictions have held that the trial court has the latitude to consider the motion. Where the purported new evidence is coupled with a strong claim of innocence, some state courts gravitate toward entertaining the motion even if untimely filed pursuant to procedural norms. A number of states, however, have neglected to create exceptions to rigid time limits for new evidence claims, and a few states openly distinguish between scientific and nonscientific evidence in implementing statutes of limitations, treating the former with greater leniency. Only at a great cost, therefore, may state prisoners delay in pursuing potential non-DNA evidentiary leads and submitting their innocence claims to the trial court after conviction.” (footnotes omitted)).

\textsuperscript{95} In one case, for example (which is discussed more fully \textit{infra}), the recanter (the accused’s son) testified as a young child that his father had killed his sister. Eight years later, he recanted his testimony. Dobbert v. Wainwright, 468 U.S. 1231 (1984) (denying \textit{certiorari}). The recanter gave three reasons for not recanting earlier: (1) he was afraid of his father; (2) following the trial, he was living at a children’s home where he was heavily medicated; and (3) the staff of the children’s home pressured him to not recant his testimony against his abusive father. \textit{Id.} at 1232 (Brennan, J., dissenting). Even if none of these reasons had ultimately been truthful, they are all certainly reasonable and debunk any sort of concept that only recantations arising within some arbitrary deadline may be presumed worthy.

\textsuperscript{96} \textit{Herrera}, 506 U.S. at 400.
Without receiving any review of the recantations, therefore, defendants falling into any one of these categories are essentially precluded from relief on the basis of such recantations. Because we accept that not all recantations are false, this preclusion is unacceptable.

III. JUDICIAL PRACTICE

This Article proceeds on the following premise: even if most recantations are false, not all are. 97 Allowing summary rejection of recantations, therefore, is problematic. Unfortunately, the combination of intensely complex procedural requirements and judicial skepticism has at times effectively produced this result. Though there are many examples, 98 the Bermudez and Davis cases illustrate just how far the effects of the judicial suspicion of recantations reach. In neither case was there any physical evidence linking the defendants to the crime. 99 In Bermudez, all of the eyewitnesses recanted, 100 and in Davis, seven of the nine recanted. 101 Undoubtedly, the recantations, if true, would have materially affected the jury’s verdicts, as the recantations entirely undermined the only factual bases for the verdicts. Yet neither defendant received a hearing in state court. 102 Instead, the judges in the respective cases judged the credibility of the statements on the basis of their suspicion of recantations in general. 103 This was particularly true in Bermudez, where the trial court stated that “it strains credulity to believe that five unshaken trial witnesses would suddenly claim that they had testified falsely under oath,” 104 and denied even a hearing “based upon the long standing reluctance of courts to overturn a jury conviction on the basis of recantation evidence.” 105 In the absence of a factual finding regarding the credibility of the recantations, it hardly

---

97 See supra note 7 and accompanying text. Neither courts nor policymakers can ignore this fact when DNA testing has and will likely continue to exonerate defendants that had previously offered recantations.

98 See Armbrust, supra note 6, at 89–98 (discussing a number of cases where courts erroneously discredited recanted statements, while also noting that “200-plus DNA exonerations have provided sound bases to reevaluate the current treatment of recanting witnesses”).


100 Bermudez II, 2009 WL 3823270, at *22–32.

101 Davis II, 660 S.E.2d at 359–60.

102 Bermudez II, 2009 WL 3823270, at *8; Davis II, 660 S.E.2d at 357.

103 Davis II, 660 S.E.2d at 358; see Bermudez II, 2009 WL 3823270, at *8.


105 Bermudez II, 2009 WL 3823270, at *8. The Bermudez court is not alone. See also Bracy v. Gramley, 81 F.3d 684, 693–94 (7th Cir. 1996) (“[T]he more sweeping the recantation, the less credible it is likely to be.”); Nicks v. Cain, No. Civ.A. 04–0519, 2005 WL 1578024, at *7 (E.D. La. June 30, 2005) (upholding the trial court’s finding that the recantation lacks credibility); Hamilton v. Herbert, No. 01 CV 1703, 2004 WL 86413, at *18–19 (E.D.N.Y. Jan. 16, 2004) (rejecting the defendant’s claims summarily while stating that the “state court’s findings are presumed to be correct”).
makes sense to dismiss recantations that, if true, would undermine the entire basis of the jury’s finding. Thus, it is the case that judicial suspicion of recantations can lead to the absurd result where a defendant who was convicted in the absence of any physical evidence and whose sole accuser recants stands no greater chance of exoneration than the defendant convicted largely on the basis of physical evidence who had only one of ten accusers recant.

Of course, such a scheme also runs afoul of the Berry and the Larrison tests (assuming such is the rule in the relevant jurisdiction), both of which focus principally on whether the newly discovered evidence would have produced a different result if it had been presented at trial. See cases cited supra note 77. At least one other court has recognized that credibility can hardly be judged by the text of the words themselves. United States v. Earles, 983 F. Supp. 1236, 1254–58 (N.D. Iowa 1997). Instead, the court outlined various factors for a court to consider in determining a recanter’s credibility:

1. the interest of the witness in the result of the hearing;
2. the witness’s relation to any party in interest;
3. the witness’s demeanor upon the witness stand;
4. the witness’s manner of testifying;
5. the witness’s tendency to speak truthfully or falsely, including the probability or improbability of the testimony given his or her situation to see and observe;
6. the witness’s apparent capacity and willingness to tell truthfully and accurately what he or she saw and observed;
7. whether the witness’s testimony is supported or contradicted by other evidence in the case;
8. whether the recanted testimony implicated a family member;
9. whether the recanted testimony was consistent with other statements the witness had given;
10. whether the recanted testimony was substantially corroborated by other evidence; and
11. whether the witness was willing to assert the recantation in open court.

Fortunately, however, not all courts have engaged in this sort of haphazard reasoning. This second group of courts either assumes the recantation is true and analyzes it accordingly, or at least marshals the facts in support of their holding. The Eighth Circuit, in a recantation case, analyzed the recanting witness’s statement and held that even if it were true “would only serve to potentially impeach the credibility of these eye witnesses, but it could not prove [defendant] did not shoot [the victim].” Cox v. Burger, 398 F.3d 1025, 1031 (8th Cir. 2005); see also Ex parte Franklin, 72 S.W.3d 671, 678 (Tex. Crim. App. 2002) (noting that even if the recantation were true, it would serve only to impeach the witness; it provided no affirmative evidence of the witness’s innocence). Thus the court was justified in dismissing the recantation without further analysis. The Ninth Circuit engaged in a similar analysis in Smith v. Baldwin, 510 F.3d 1127, 1143 (9th Cir. 2007). Though the court noted the general unreliability of recantations, id. at 1141–42, the court took the appropriate measure of nevertheless assuming the recantation was true for purposes of explaining why it was nonetheless immaterial. Id. at 1142 (“[The] recantation changes little. A reasonable juror would still have sufficient evidence to believe that [defendant] was more likely than not the real killer, and [defendant] certainly cannot meet his burden by a preponderance of the evidence.”). The court proceeds to explain exactly why the recantation was neither material nor credible. Id. at 1141 (noting, among other things, that the recantations, even if true, failed to explain various inconsistencies in some of the trial testimonies, or why the recantation lacked corroboration when the original did
In *Davis*, Georgia state courts relied on little more than the inherent unreliability of recantations in affirming a death sentence.\(^{108}\) Relying entirely on “the general lack of credibility that should be assigned to recantation testimony,”\(^{109}\) the Georgia Supreme Court stated that it would credit the recanting testimony only if it could show “that the State’s witness’ [trial] testimony in every material part [was] purest fabrication.”\(^{110}\) Because recantation testimony merely “impeaches the witness’ prior testimony,” the court found it was “not the kind of evidence that proves the witness’ previous testimony was the purest fabrication.”\(^{111}\) Moreover, the court concluded, without weighing the cumulative effect of the recantations, that each individual recantation would only undermine the relevant testimony, not the overall case.\(^{112}\) This conclusion was dubious, as the entire case was built on the testimonies of those same eyewitnesses—indeed, there was no other evidence. But perhaps the conclusion was justified legally under the “purest fabrication” standard erected by Georgia state courts.

The approach pursued by the Georgia state courts in *Davis* is worrisome, because it effectively allows summary dismissal of recantations. Every recantation, by its very definition, is impeachable by prior trial testimony.\(^{113}\) Thus, it is difficult (if not impossible) to imagine an instance where a recantation could ever warrant relief under the “purest fabrication” standard.\(^{114}\) Accepting our premise that not all

---

\(^{108}\) *Davis II*, 660 S.E.2d at 358 (“The trial court relied on this principle [i.e., the mistrust of recantations] in disregarding the alleged recantations in Davis’s case.”).

\(^{109}\) *Id.*

\(^{110}\) *Id.* (citation omitted) (internal quotation marks omitted).

\(^{111}\) See *id.* (citation omitted) (internal quotation marks omitted).

\(^{112}\) *Id.* at 358–60.

\(^{113}\) Indeed, by its very definition, a recantation is impeachable. Recantation means “to make an open confession of error.” *Merriam Webster’s Collegiate Dictionary* 1038 (11th ed. 2003).

\(^{114}\) This consideration led Chief Justice Sears to state in his dissenting opinion:

We have noted that recantations by trial witnesses are inherently suspect, because there is almost always more reason to credit trial testimony over later recantations. However, it is unwise and unnecessary to make a categorical rule that recantations may *never* be considered in support of an extraordinary motion for new trial. The majority cites case law stating that recantations may be considered only if the recanting witness’s trial testimony is shown to be the ‘purest fabrication.’ To the extent that this phrase cautions that trial testimony should not be lightly disregarded, it has obvious merit. However, it should not be corrupted into a categorical rule that new evidence in the form of recanted testimony can never be considered, no matter how trustworthy it might appear. If recantation testimony, either alone or supported by other evidence, shows convincingly that prior trial testimony was false, it simply defies all logic and morality to hold that it must be disregarded categorically.
recantations are false, Georgia’s “purest fabrication” test, and others like it, lead to unacceptable results.

Another example is a 1984 death penalty case, for which the U.S. Supreme Court denied certiorari, *Dobbert v. Wainwright.* That case involved a man convicted of murdering his daughter based solely on the eyewitness testimony of his son. Eight years after trial, the son recanted his testimony, citing his fear of his father, pressure from the staff of a children’s home where he was staying, and an intense regimen of medication as his reasons for lying at trial. Echoing the approach from the *Davis* case, the state trial court there dismissed the recantation based entirely on its discussion of the prerecantation statements. The state court reached its conclusion without any discussion of the son’s recantation—this despite the fact that the recantation corroborated statements the son had made shortly after the death of the victim. Providing further illustration of the difficulties facing similarly situated defendants, the federal courts in Dobbert’s case were more or less bound by the factual determinations of the state courts. Thus, the Supreme Court denied certiorari in that case and Dobbert’s motion to stay the execution was not granted.

In a heated dissent, Justice Brennan argued that the execution should have been stayed in order to allow analysis of “the appropriate ground rules for considering the materiality and credibility of [the son’s] recantation.” Justice Brennan focused largely on the state court’s “complete failure to analyze the recanting evidence itself.” He also noted reasons one could, at least arguably, credit the recantations over the original trial testimony. For example, the son was terrified of his abusive father as a young boy at the time of the trial, at the time of the recantation, however, the son was “a grown man . . . [with] no reason to fear his father and has . . . recovered from the trauma to which he was subjected.” As a legal matter, Justice Brennan’s conclusion was questionable, as one can hardly argue that federal appellate courts are the best forum for challenging state courts’

*Dobbert v. Wainwright,* 660 S.E.2d at 363-64 (Sears, C.J., dissenting).


116 *Id.* at 1231–32 (Brennan, J., dissenting).

117 *Id.* at 1232.

118 *Id.* at 1233.

119 *Id.* at 1232–33 (“Although [the son’s] initial statements to the police contradicted his subsequent trial testimony, [the trial court] found that the inconsistencies were easily explained away in light of the boy’s terror of the father.”).

120 *Id.* at 1233.

121 *Id.* at 1231.

122 *Id.* at 1238.

123 *Id.* at 1236.

124 *Id.* at 1236–37 n.*.

125 *Id.*
factual conclusions. Nevertheless, his dissent highlights some important consequences of judicial practice in this area.

That a man could be executed in the face of a recantation without ever receiving any sort of meaningful review of the exact contents of that recantation seems “absurd on its face.” Yet judicial suspicion coupled with precedent that requires courts to focus exclusively on the prerecantation material can drive this result. Just as in Davis, Dobbert was denied any meaningful review of the recantation, because in both cases the state courts focused on the prerecantation testimony and largely ignored the contents of the recantation itself. This was particularly problematic in Dobbert, where the recanter’s testimony corroborated earlier statements and there were reasons to believe the recantation over the trial testimony. Again, accepting our premise that not all recantations are false, this is an unacceptable reality in our justice system. As Justice Brennan noted in his Dobbert dissent:

There may well be numerous reasons why [the son’s] recantation should not be entitled to controlling weight, but there is nothing here to suggest that [the trial court] considered those possible reasons, let alone relied upon them in reaching his conclusion. In the face of a specific recantation of critical testimony, a court must evaluate the recantation itself and explain what it is about that recantation that warrants a conclusion that it is not credible evidence.

The subtext underlying these three cases is simple: courts mistrust recantations. In Bermudez, the trial court refused to hold an evidentiary hearing, relying entirely on the inherent unreliability of recantations. In Davis, the Georgia courts, while noting the same principle, relied on precedent that makes it all but impossible to obtain meaningful review of a recantation. Such precedent alone stands out as evidence of the long-standing judicial skepticism of recantations. More importantly, it highlights the ease with which most courts can simply dismiss recantations. A simple principle can be derived from these cases: momentum drives courts to dismiss recantations based on only the historic reasons discussed in this Article. Thus, absent extraordinary circumstances, we can expect courts to presume recantations unreliable and dismiss them accordingly.

The combination of intensely complex procedural requirements and judicial skepticism has effectively allowed courts to categorically deny recantations. The solution, therefore, lies first in an honest recognition that “the recantation on occasion represents the truth,” and second, and more importantly, in ensuring

127 See Dobbert, 468 U.S. at 1235 (Brennan, J., dissenting).
128 Id. at 1236.
129 Id. at 1235–36.
that courts reject recantations only after properly granting “the process that attends a valid credibility assessment.”131 For courts to summarily dismiss recantations as unbelievable would be, in no trite sense, “bothersome.”132 The proper solution lies in crafting a system that, while not imposing an undue burden on courts that would both disturb judicial economy and the finality of judgments, makes it more difficult for judges to summarily dismiss recantations based only the historic mistrust of them.

IV. EXISTING SCHOLARSHIP

Before delving into the specifics of our recommended framework, we first examine various proposals that other authors have made. Generally, we are concerned the existing proposals do not go far enough in addressing the underlying sources of the problem: judicial suspicion and procedural complexities.

Many scholars have written at length about the problems inherent with the preexisting judicial standards, namely the Berry and Larrison standards. Daniel Wolf, for example, has criticized the existing frameworks as too burdensome and unjust.133 He puts forward an alternative standard, which he entitles the “significant chance” rule.134 He proposes that

[i]n order to grant a rule 33 motion for a new trial based upon newly discovered evidence of false testimony at trial, the court must be reasonably well satisfied that (a) testimony given by a material witness at trial was false or that the witness has become so thoroughly discredited that the court is unable to determine whether the testimony was true or false; (b) there is a significant chance that a jury with knowledge that the witness testified falsely would return a different verdict; and (c) the defense was taken by surprise by the false testimony and was unable to meet it or know of its falsity at the time of the trial.135

In a similar vein, Sharon Cobb argues that the present standards inappropriately shift the underlying factual question—i.e., whether the recanting witness is credible—to the judge.136 Thus, she argues for a new standard, wherein “new trial motions should be granted whenever there is substantial doubt as to

---

131 Id. at 601 (“But when a court rejects a plausible recantation out of hand, without any of the process that attends a valid credibility assessment, we owe the court’s finding no respect. To sustain such a summary determination, as we do today, is to unnecessarily exalt the sovereignty of state courts over due process and the pursuit of truth.”).
132 Id. at 600.
133 Wolf, supra note 46, at 1947.
134 Id. at 1947–48.
135 Id.
136 See Cobb, supra note 41, at 993–94.
which of the witness’s statements [i.e., the original trial statement or the recanting statement] is true . . . .

Similarly, Janice Repka proposes the “reasonable possibility approach.”

Like most other commentators on the subject, she agrees that the judicial presumption that recantations are untrustworthy has become too broad. To combat judicial standards that she views as too loose in this area, she proposes that “[w]here the recanting occurred under specific circumstances reasonably free from suspicion of duress or improper motive, and, upon a new trial the original trial testimony reasonably could be found false, a new trial will be granted . . . .”

We agree with several of the premises underlying these authors’ scholarship, namely that the judicial suspicion of recantations has become too broad and that the existing standards do not adequately equip courts to weigh recantations. Our concern, however, is that the approaches advocated by these scholars all suffer the same defect. As evidenced by the cases highlighted in this Article’s introduction, courts have an inherent suspicion and mistrust of recantations. Thus, any solution that seeks to compensate for this suspicion merely by articulating for judges another imprecise standard seems to miss the point. The focus, rather, should be on crafting a scheme with a built-in mechanism to compensate for the judicial tendencies that drive denials of recantations.

This is not to say that new judicial standards would have no effect; they likely would. For example, the Georgia courts in the Davis case might no longer be bound by the “purest fabrication” test, which, for the reasons discussed in Part III, could have made a significant impact in that case. Nevertheless, we merely emphasize that, regardless of the standard to be employed, courts do not trust recantations. Therefore, in the absence of exceptional circumstances, the presumption against recantations will prevail. For example, under Repka’s proposed “reasonable possibility” standard, a judge must determine whether “upon a new trial the original trial testimony reasonably could be found false.” In making this determination, however, we can expect judges to resolve evidentiary doubts against the recantations—both because of the judicial suspicion of recantations and to protect the finality of judgments. Likewise, Wolf’s proposed

---

137 Id. at 1001–02.
138 Repka, supra note 72, at 1454.
139 See id. at 1445–46.
140 Id. at 1454.
141 See cases cited supra note 17.
142 See supra notes 110–114 and accompanying text.
143 Repka, supra note 72, at 1454.
144 Medwed has written at length about the cognitive biases of a judge reviewing a recantation years after the trial ended. Medwed, supra note 2, at 699–709. He identified two contributing factors to this bias: (1) most judges reject the prospect of a wrongful conviction, because they see themselves as fair and competent individuals; and (2) judges often have “difficulty deviating from a prior decision because that decision has become the reference point to which they compare and contrast newfound information.” Id. at 701–02; see also infra note 175 (discussing how judges should not supplant the role of the jury).
“significant chance” rule focuses on whether “there is a significant chance that a jury with knowledge that the witness testified falsely would return a different verdict.” But of course, given the mistrust of recantations (which would remain unaffected by the new standard), it will rarely be the case that a judge would conclude, in the absence of extraordinary circumstances, that there is a significant chance that a jury would come to a different result than it already did in the first trial. Finally, Cobb’s proposed standard, which would grant a new trial anytime “there is substantial doubt as to which of the witness’s statements [i.e., the original trial statement or the recanting statement] is true . . . ,” suffers the same tragic flaw. Because long-established precedent forces courts to view recantations with great suspicion, we can expect courts to err on the side of finding that there is no substantial doubt about the witness’s statements.

We submit that, while these authors have crafted creative and helpful ways for courts to examine recantations, they fail to properly measure the effect of generations’ worth of judicial tradition as well as complex procedures. We thus view the proper solution as more mechanical than merely proposing a new standard—something courts could not overlook through simple invocation of the historical suspicion of recantations.

In a 2008 article, Shawn Armbrust advances a proposal, which, in our view, comes closer to achieving this goal than the other aforementioned proposals. Armbrust criticizes those standards that force a defendant to either prove the truth of the recantation or the falsity of the trial testimony. He notes several cases where the defendants were eventually exonerated through the use of DNA evidence, even though their accusers had recanted years earlier. He explains that, “[i]n those scenarios, proving the absolute truth of the recantation or absolute falsity of the trial testimony was virtually impossible until each defendant could use DNA evidence.” He continues, “[t]he simple explanation for this is that when only the witness knows the real truth, the courts essentially are forced to guess based on credibility judgments and other subjective factors.” Thus, he concludes that a new standard is necessary “to balance the interests in judicial economy with the reality that discerning truth based on witness demeanor is incredibly difficult.” He therefore proposes a corroboration requirement, in which a defendant’s motion for a new trial based on recanted statements will fail, unless that defendant corroborates the recantation with other independent evidence.

145 Wolf, supra note 46, at 1947.
146 Cobb, supra note 41, at 1001–02.
147 See Armbrust, supra note 6, at 98–103.
148 Id. at 99–100.
149 Id. at 99–101.
150 Id. at 99.
151 Id. at 99–100.
152 Id. at 100.
Armbrust relies on Wisconsin state court precedent to articulate a two-pronged corroboration requirement. 153 "The first element of Wisconsin’s . . . corroboration requirement is evidence of a feasible motive for providing false testimony at trial." 154 "The second element of Wisconsin’s corroboration requirement is circumstantial guarantees of trustworthiness." 155 Armbrust proceeds to outline a nonexhaustive list of factors to consider in determining whether there are sufficient circumstantial guarantees of trustworthiness, such as "the spontaneity of the statement, . . . the extent to which the statement is self-incriminatory and against the penal interest of the declarant, and the declarant’s availability to testify under oath and subject to cross-examination." 156 Armbrust concludes that under this approach of "objectively consider[ing] reliability both in the traditional context articulated by the Wisconsin courts and in the context of what we now know about the fallibility of the justice system, [judicial decisions] are likely to be fairer and to balance more appropriately the interests of defendants with interests of finality." 157

Armbrust recognizes that, in order for there to be meaningful change in this area of law, a more mechanical mechanism must be put in place to counterbalance judges’ natural tendencies in this area. 158 We agree. Though the judicial standards articulated by the aforementioned authors are novel and would certainly effect change, in the absence of some structural counterweight to judicial disapproval of recantations, meaningful review in this area will remain inconsistent.

We worry, however, that Armbrust’s proposed method for determining corroboration ultimately suffers from the same defect as the other proposals. He urges courts to focus on “evidence of a feasible motive for providing false testimony,” 159 and on “circumstantial guarantees of trustworthiness.” 160 As we have discussed throughout, there are almost always reasons to infer a recantor is lying. Thus, even if there is evidence of a feasible motive for providing false testimony, there will surely almost always be evidence of a feasible motive for lying in the recantation as well. Consequently, we fail to see how the first prong of the corroboration requirement compensates for the historical judicial trends.

---

153 Id. at 100–01.
154 Id. at 100.
155 Id. at 101.
156 Id. (citation omitted).
157 Id. at 101–02.
158 As a side note, we find irrelevant Armbrust’s second proposal, which is that “[appellate] courts should adopt a less deferential standard of review for summary denials of new trial motions based on recantation testimony.” Id. at 102. Without engaging this point in great detail, appellate courts are not in the business of engaging in intense factual finding, which any review of a recantation must necessarily include. Thus, in our view, the correct approach focuses more on the process that the trial courts undertake and less on the process adopted by the appellate courts. We have focused our efforts in this paper accordingly.
159 Id. at 100.
160 Id. at 101.
Likewise, for all the reasons discussed in Parts I and III, there are almost always reasons to infer that a recantation is untrustworthy. It would be a rare case indeed that a recantation suffered none of the traditional defects. Even if Armbrust’s proposed corroboration requirement forces courts to consider the potentially truthful implications of a recantation, it certainly does not require those courts to ignore all the traditional indicia of reliability associated with recantations. Thus, adopting Armbrust’s proposed corroboration requirement would simply place courts in a position where they are forced to recognize that there is competing evidence on both sides. Because a court’s inertia will push that court to dismiss the recantation and defer to the jury’s earlier factual findings, there is little reason to presume the courts, under the standard proposed by Armbrust, will change their behavior in this regard in any significant way. Thus, while we agree with Armbrust that a corroboration requirement is necessary, we would push harder and recommend the corroboration requirement be a more easily verifiable bright-line test. Though bright-line tests are not always preferred, it would provide the most guaranteed way of balancing judicial practice in this area.

These concerns lead us to a few simple conclusions. First, a court’s path of least resistance, when presented with a recantation, is to disregard it. Second, even where there are multiple explanations for why a witness would have recounted, we expect courts will, for the historical and policy reasons noted throughout this Article, defer to the explanation that discredits the recantation. Third, unless there is an automatic mechanism in place to compensate for this judicial gravity, we cannot expect meaningful change in this area.

Working from the scholarship and taking a practical view of judicial practice in this area, we outline a framework we believe would best accomplish this goal.

V. RECOMMENDED FRAMEWORK

Being sensitive to the need to preserve final judgments and protect scarce judicial resources, courts and legislatures can nonetheless craft mechanisms that will force judges to more precisely and validly distinguish between false recantations and truthful ones.

As an initial measure, courts can weed out a substantial amount of recantation cases through a simple “materiality” analysis. In other words, where a recantation, even if assumed true, could not materially affect a jury’s verdict, courts could easily dispose of such cases. Courts frequently do this already. Moreover, the Berry and Larrison standards, of which most jurisdictions have already adopted some variation, include such a requirement. Thus, imposition of

---

161 See, e.g., Smith v. Baldwin, 510 F.3d 1127, 1148 (9th Cir. 2007).
162 The Eighth Circuit, in a recantation case, analyzed the recanting witness’s statement and held that even if it were true it “would only serve to potentially impeach the credibility of these eye witnesses, but it could not prove [defendant] did not shoot [the victim].” Cox v. Burger, 398 F.3d 1025, 1031 (8th Cir. 2005).
163 See supra note 107 and accompanying text.
164 See supra note 77.
a strict requirement to consider materiality would not pose an administrative obstacle for most courts.

Furthermore, the U.S. Supreme Court has offered guidance on the implementation of this rule in an analogous setting. In *Franks v. Delaware*, the Court specifically addressed challenges to affidavits supporting a search warrant—an approach easily applicable in the recantation scenario. In concluding that a defendant may, in the rare circumstance, challenge the validity of affidavits, the Court held:

There is, of course, a presumption of validity with respect to the affidavit supporting the search warrant. To mandate an evidentiary hearing, the challenger’s attack must be more than conclusory and must be supported by more than a mere desire to cross-examine. There must be allegations of deliberate falsehood or of reckless disregard for the truth, and those allegations must be accompanied by an offer of proof. They should point out specifically the portion of the warrant affidavit that is claimed to be false; and they should be accompanied by a statement of supporting reasons. . . . [I]f these requirements are met, and if, when material that is the subject of the alleged falsity or reckless disregard is set to one side, there remains sufficient content in the warrant affidavit to support a finding of probable cause, no hearing is required. On the other hand, if the remaining content is insufficient, the defendant is entitled, under the Fourth and Fourteenth Amendments, to his hearing.

The challengers’ attacks, like recantations, are presumed to be untrustworthy. Nevertheless, the Court has articulated a specific standard for overcoming this presumption. The Court requires two things: (1) claimants must first point to the “specific” portions of the affidavit claimed to be false; and (2) claimants should accompany their allegations with a statement of supporting reasons. Then the court proceeds by assuming the allegations are true. If even in the absence of the false portions, probable cause would remain, the challenge fails. If, however, probable cause would no longer exist, then a hearing is warranted.

Courts could work from this precedent. As in *Franks*, courts would first require that the moving defendant point out how the recantation would specifically affect both the witness’s trial testimony and the overall evidence presented against him. Then the court would determine whether the recantation is material by asking whether, assuming the recantation is true, the recantation would alter the guilty verdict. If the answer is no, the motion for a new trial is denied. If the answer is

---

166 *Id.* at 171–72.
167 *Id.* (emphasis added).
168 See *supra* Part I.A.
169 *Franks*, 438 U.S. at 171.
170 Of course nearly all jurisdictions have already adopted some variation of this “materiality” requirement in their motions for a new trial on the basis of newly discovered
yes, however, the court proceeds to the more important step of determining the credibility of the recantation.

Pursuant to our earlier discussion, we believe that the most appropriate method for determining the credibility of a recantation must seek to counterbalance courts’ impulses to summarily deny recantations. Thus, instead of merely offering a new formulation of a court’s discretion in this area, the more precise response must effect a fundamental structural change—one that will force courts to consider the veracity of a recantation independent of its historic untrustworthiness. Because of judicial cognitive biases, we do not expect that the same judges who presided over the original trials will be ideally suited to be the final arbiters of a recantation’s trustworthiness. Rather, there needs to be a system in place, which automatically weeds out clearly unreliable recantations. Where there remains plausible doubt about the credibility of a recantation, the ultimate resolution of the question is better left in the hands of a different fact finder. We therefore second Armbrust’s proposal to add a corroboration requirement to the review process here. In other words, where a recantation has been sufficiently corroborated, determinations of that recantation’s credibility will go to a fact finder. Where a recantation is not sufficiently corroborated, the judge may dismiss the recantation as incredible.

The corroboration requirement should not be left to the discretion of the trial judge. Instead, we encourage courts, together with the relevant legislative and judicial committees in charge of drafting and approving the respective jurisdictions’ rules governing the use of evidence in postconviction hearings, to consider grouping recantations with other forms of unreliable evidence. Grouping recantations with other forms of evidence that are subject to their own set of
evidence. See supra note 77 and accompanying text. This aspect of the Article’s suggested approach would not diverge with the judicial and statutory approaches of most jurisdictions and thus would presumably not be difficult to execute.

171 See supra notes 103–131 and accompanying text.
172 See supra notes 55–61 and accompanying text.
173 Cobb, supra note 41, at 980, 994 (“One study indicates that the most reliable guide to truth-telling is the verbal content of testimony and not facial or non-verbal cues. Findings such as these suggest that the deference accorded to the trial judge, if not ill-founded, must be justified by policy grounds rather than those grounds generally articulated. . . . In determining which of the witness’s versions of events is true, the trial judge is preempting the function of the jury.” (footnote omitted)); Repka, supra note 72, at 1452–54 (detailing numerous cases and commentators discussing the desirability of juries over judges in making these sorts of credibility determinations); Wolf, supra note 46, at 1936–37 (“The right to trial by jury means that a jury, not a judge, must make the final determination of the innocence or guilt of the accused. The function of the trial judge is to decide only whether the newly discovered evidence should be submitted to the jury, not what its ultimate weight should be. Because there is no way for a court to determine that the perjured testimony did not have controlling weight with the jury, when the judge attempts to second-guess the jury be determining whether it probably would have acquitted, he usurps the jury’s function.” (footnote omitted) (internal quotation marks omitted)).
legislative guidelines and case law would remove them from their current limbo, and instead place them with other forms of traditionally unreliable evidence that have, for one reason or another, received more attention over the years. This would certainly go a long way in promoting uniformity in courts’ treatment of certain historically unreliable forms of evidence. It would also provide the requisite automatic mechanism to offset judicial momentum in this area.

We refer here to both the “prior inconsistent statement” and “statements against interest” rules of the Federal Rules of Evidence. Federal Rule 801(d)(1) exempts from the “hearsay rule” certain prior inconsistent statements. It makes admissible statements where, “[t]he declarant testifies at the trial or hearing and is subject to cross-examination concerning the statement, and the statement is . . . inconsistent with the declarant’s testimony, and was given under oath subject to the penalty of perjury at a trial, hearing, or other proceeding, or in a deposition . . .”. Logically, there is very little reason to distinguish prior inconsistent statements from subsequent inconsistent statements. As one commentator noted, “[t]he only difference is in the timing of the alteration of the witness’s testimony; in one case it occurs before trial, and in the other, afterwards.” Therefore, “[i]f our legal system deems it important that juries have knowledge of prior inconsistent statements by witnesses, it is difficult to understand why so many obstacles are placed in the path of a defendant who wishes to acquaint a jury with subsequent inconsistent statements.”

Federal Rule 804(b)(3), which governs the admissibility of hearsay statements made against the declarant’s penal interests (i.e., statements against interest), also provides useful comparisons. Historically, statements against interest constituted one subset of hearsay statements, which, like recantations, are presumed invalid and untrustworthy. Both types of statements, however, can bear indicia of reliability. Regarding statements against interest, the advisory notes to the Federal Rules of Evidence allow these statements against interest to occasionally be admitted upon the “assumption that persons do not make statements which are damaging to themselves unless satisfied for good reason that they are true.” Under this same logic, recantations bear similar indicia of reliability. Recanting

---

175 Fed. R. Evid. 801(d)(1).
176 Cobb, supra note 41, at 994.
177 Id. at 994–95.
178 See, e.g., Fed. R. Evid. 802 (“Hearsay Rule”) (“Hearsay is not admissible except as provided by these rules or by other rules prescribed by the Supreme Court pursuant to statutory authority or by Act of Congress.”). The common law distrusted “evidence of confessions by third persons offered to exculpate the accused.” Id. at 804 advisory committee’s note to subdivision (b). Courts suspected defendants could fabricate both the confession and its contents and then rely on the unavailability of the witness to more easily admit the fabricated evidence. Id.
179 Id. at 804 advisory committee’s note to subdivision (b).
witnesses often expose themselves to perjury charges by admitting that they lied at trial, or at least leave themselves vulnerable to individuals sympathetic to the accused. And just like courts presume that individuals will not make damaging statements about themselves "unless satisfied for good reason that they are true," they could similarly conclude that witnesses will not recant and expose themselves to perjury charges, or any other threats, unless the statements are true.

Admittedly, courts have traditionally viewed these realities of recantations as reasons to disbelieve them. As noted earlier, courts worry about the effect that the subsequent coercive threats could have on a recantation. That consideration merely underscores the conclusion that there are frequently competing and compelling reasons to both allow and disallow recantations, just as there are with statements against interest. That such considerations exist is not, of course, a reason to disallow recantations altogether.

We submit that recantations are logically very similar to both "prior inconsistent statements" and "statements against interest." Recantations often arise from suspicious circumstances, but nevertheless as a whole, can bear some indicia of reliability. We recognize, however, that simply allowing all recantations in the same way that many prior inconsistent statements are admitted at trial would disrupt the finality of judgments and vindicate many of the concerns discussed earlier. Fortunately, the Federal Rules of Evidence provide further insight into how to more precisely filter the reliable recantations from the unreliable.

For statements against interest, the Federal Rules of Evidence struck this balance by adding, among other things, a "requirement of corroboration." The corroboration requirement is an attempt to filter the reliable statements from the unreliable. The rule was drafted to read that an out-of-court statement against the declarant’s penal interest is admissible if (1) the declarant is unavailable as a witness, (2) the statement so far tends to subject the declarant to criminal liability that a reasonable person in the declarant’s position would not have made the.

---

180 Id.
181 See supra Part I.A.
182 Cobb, supra note 41, at 994–95 (“[T]he courts’ response to recanting testimony is not in accord with the practice of admitting prior inconsistent statements of witnesses at trial, although the situations are remarkably analogous. The only difference is in the timing of the alteration of the witness’s testimony; in one case it occurs before trial, and in the other, afterwards. But while jurors rarely get to hear recanting testimony, under the Federal Rules of Evidence prior inconsistent statements are routinely admissible for impeachment purposes, and sometimes for substantive purposes as well. One authority notes that the use of prior inconsistent statements or self-contradiction is the most effective and most frequently employed means of impeaching witnesses, on the theory that ‘talking one way on the stand and another way previously is blowing hot and cold, and raises a doubt as to the truthfulness of both statements.’ If our legal system deems it important that juries have knowledge of prior inconsistent statements by witnesses, it is difficult to understand why so many obstacles are placed in the path of a defendant who wishes to acquaint a jury with subsequent inconsistent statements.” (footnotes omitted)).
183 FED. R. EVID. 804 advisory committee’s note to subdivision (b).
statement unless he or she believed it to be true, and (3) corroborating circumstances clearly indicate the trustworthiness of the statement. Consequently, courts may not summarily dismiss statements against interest as unreliable. Rather, they must base their decisions to accept or dismiss statements against interest on whether there has been corroboration.

A similar calculus may be applied to recantations. Instead of merely allowing all recantations that materially contradict a trial statement or that could conceivably expose the declarant to liability, we recommend the adoption of a “corroboration” requirement in the recantation context. Such a requirement would provide the appropriate automatic and mechanical counterbalance to historical suspicion of recantations and would not be a far break from precedent (inasmuch as courts already apply such a requirement to statements against interest, which are analytically similar to recantations). Courts could, therefore, justify imposing a corroboration requirement for recantations under the same logic that they have done so for statements against interest.

To be sure, we are not recommending the wholesale adoption of the “statement against interest” test for evaluating recantations. They are, in the end, different forms of evidence requiring different modes of evaluation. The first requirement—that the declarant be unavailable—is likely irrelevant when it comes to recantations. A recanting witness will be required to offer the recanting testimony either at a hearing or as an affidavit—under oath in either event. And like with prior inconsistent statements, the previous statement—i.e., the trial testimony—must also have been given under oath and subject to cross-examination. Likewise, the second step can also be relaxed in the recantation context, because, notwithstanding their similarities, there are important contextual differences between a recantation and a statement against interest. First, recantations are not actually hearsay, and thus the normal rules regarding when hearsay does and does not bear indicia of reliability are not entirely relevant. Of course, inasmuch as the recantation involves the restatement of hearsay, the typical hearsay rules would apply. Second, and perhaps more importantly, determining motive would again put this question back within the broad discretion of a judge, who, for all the reasons noted throughout this Article, has every reason to dismiss the recantation. As we have noted, we seek to counterbalance that inertia with a more mechanical test, because there are often reasons to both allow and disallow any given recantation based on the declarant’s perceived motive. Therefore, leaving this question in the hands of the same trial judge to preside over the trial would suffer the same tragic flaw that the other approaches did. Finally, there is little logical reason to allow only those recantations where the recanters expose themselves to criminal liability. Such an approach would fail to account for situations where witnesses change testimonies as a result of seeing some new extrinsic evidence, which causes them to reevaluate what they claim to have seen.

184 Id. at 804(b)(3).
185 See id. 801(d).
The more difficult question examines the specifics of the proposed corroboration requirement. In order to provide more objective, easily administrable, and easily verifiable results, the required corroboration ought to come in the form of an independently verifiable piece of evidence, such as another testimony, a piece of physical evidence, or compelling circumstantial evidence. Hinging the admissibility of recantations on the existence of independent evidence will resolve several aspects of the debate. First, courts can easily identify whether such evidence exists. This would create a very easily administrable bright-line test—one that would weed out those recantations most likely unreliable while not requiring courts to expend too many resources. It would also allow those criminal defendants with compelling evidence of innocence a more reasonable chance to have their evidence reviewed. It would mitigate the generally unreliable nature of recantations by requiring independently verifiable evidence. Finally, it would provide an automatic mechanism for shifting the factual determination to a different judge (from the original trial judge) or a jury.

In sum, our proposal seeks to add elements of objectivity and precision to the review process of recantations. Because judges may currently dismiss recantations without any sort of meaningful review solely by relying on recantations’ general suspicious nature, our proposal seeks to counterbalance judges’ natural incentives to dismiss recantations. It accomplishes this by defining a concrete test for determining both the materiality and credibility of a recantation. The review process will become more transparent, as it will hinge upon the existence of corroborating evidence. Such a system, we hope, will sufficiently resolve the

186 Judge Flaum, speaking for the Seventh Circuit, emphasized this fact: “[T]he precise meaning of the corroboration requirement in Rule 804(b)(3) is uncertain, and is not much clarified by either legislative history or the cases.” United States v. Garcia, 897 F.2d 1413, 1420 (7th Cir. 1990) (quoting United States v. Silverstein, 732 F.2d 1338, 1346–47 (7th Cir. 1984)). Thus the question remains on how courts ought to judge the reliability and credibility of recantations. Nevertheless, case law has identified at least some factors relevant to the determination of whether certain evidence in fact corroborates the statement. 4 STEPHEN A. SALTBURG ET AL., FEDERAL RULES OF EVIDENCE MANUAL § 804.02[9] (9th ed. 2006).

Those factors include: (1) the timing and circumstances under which the statement was made; (2) the declarant’s motive in making the statement and whether there was a reason for the declarant to lie; (3) whether the declarant repeated the statement and did so consistently, even under different circumstances; (4) the party or parties to whom the statement was made; (5) the relationship between the declarant and the opponent of the evidence; and (6) the nature and strength of independent evidence relevant to the conduct in question.

187 In the end, it is not necessary for us to outline the details of the corroboration requirement because courts can simply tap into their respective body of case law surrounding the specifics of their respective corroboration requirement.
demands of judicial economy while accommodating the need to locate and expose the occasional truthful recantation.

In substance, our proposal comes quite close to that proposed by Armbrust. We both agree that a corroboration requirement is “necessary to balance the interests in judicial economy with the reality that discerning truth based on witness demeanor is incredibly difficult.”\(^{188}\) However, whereas Armbrust’s corroboration requirement examines “evidence of a feasible motive for providing false testimony at trial” and “circumstantial guarantees of trustworthiness,” our corroboration requirement attaches the body of law surrounding Rule 804(b)’s corroboration requirement to witness recantations. By doing so, we hope our approach does not suffer the same flaw past approaches have; instead of merely articulating another vague standard subject to the wide discretion of judges already inclined to reject recantations, we seek to have recantations treated the same as another similar form of historically unreliable evidence. Thus, in contrast to Armbrust’s approach, ours adds an element of uniformity and predictability that has heretofore been lacking in this area of jurisprudence.

In addition to the corroboration requirement, we recommend several procedural reforms. We have noted several procedural complexities, which preclude a significant number of defendants from ever presenting their recantations. As noted, defendants have a limited time frame, in which to “discover” recantations.\(^{189}\) Such statutory schemes risk ignoring the psychological complexities of a recanting witness.\(^{190}\) If a witness has perjured himself at trial, there is generally no compelling reason he would be more willing to come forward with his recantation sixty days, instead of say ten years, after the end of the trial. A defendant whose accusers recant after the applicable time limit has no remedies but to seek habeas relief, which is, as discussed, limited to constitutional claims.\(^{191}\)

Because a typical recantation scenario would not necessarily involve a constitutional violation, defendants in this situation remain incarcerated with no recourse. Imposing no limit is, of course, one possible solution, though this option may be less appealing to some concerned with efficiently obtaining a final judgment.\(^{192}\) States could consequently take the less dramatic step of eliminating a

\(^{188}\) Armbrust, supra note 6, at 100.

\(^{189}\) See supra notes 68–77 and accompanying text.

\(^{190}\) See supra Part II.C.

\(^{191}\) See supra Part II.C.

\(^{192}\) Additionally, states could create something akin to an “Extraordinary Motion for a New Trial on the Basis of Newly Discovered Evidence.” This would be similar to the motion available to Troy Davis in the Georgia state courts. See Davis II, 660 S.E.2d 354, 358 (Ga. 2008). When the motion for a new trial was no longer available, Mr. Davis filed an extraordinary motion for a new trial. Id. (“Extraordinary motions for new trial are ‘not favored,’ and ‘a stricter rule is applied to an extraordinary motion for a new trial based on the ground of newly available evidence than to an ordinary motion on that ground.’ A trial court’s ruling on such a motion ‘will not be reversed unless it affirmatively appears that the court abused its discretion.’ [Cit.]’ For the reasons set forth below, we conclude that the trial court did not abuse its discretion in denying Davis’s extraordinary motion for new trial
time limit for discovering nonphysical evidence. Either option would significantly reduce the possibility that the “occasional truthful” recantation will go unspotted.

VI. APPLICATION OF RECOMMENDED FRAMEWORK TO THE BERMUDEZ AND DAVIS CASES

We next examine how our proposed standard would have affected the Bermudez and Davis cases.

In Bermudez, there is no question that the recantations, if believed, would have had a material effect on the verdict. There was no physical evidence in that case, and every eyewitness recanted. Moreover, the prosecution’s key witness, Efraim Lopez, recanted his testimony, in which he identified Bermudez as “Wool Lou”—it being accepted that the individual known by the nickname Wool Lou was the killer. Two years after trial, Lopez recanted that testimony and identified a man named Luis Munoz as “Wool Lou.” None of the other four recanting witnesses testified about the identity of Wool Lou, but all claimed that the police had used suggestive identification procedures and that they had only ever had

without first conducting a hearing, particularly in light of the requirement under Timberlake [v. Georgia, 271 S.E.2d 792 (Ga. 1980)], that newly-discovered evidence be so material that it probably would result in a different verdict, and in light of the duty of a defendant to present in the affidavits supporting his or her extraordinary motion for new trial “facts sufficient to authorize that the motion be granted.” (citations omitted)).

193 Such a result makes sense because the existence of nonphysical evidence is in no way contingent upon the defendant’s efforts or lack thereof.

194 At least one other scholar has also recommended this approach. Daniel Medwed writes at length on this point. Medwed, surpa note 2, at 694–95 (“Ridding state post-conviction regimes of unforgiving time limits on newly discovered evidence claims not only accrues to the benefit of potentially innocent defendants, but also reflects well upon the criminal justice system itself, sending a message that courts are willing and able to look anew at legitimate claims of innocence and not hide behind a veil of procedural default. Furthermore, to offset the uncertainty created by the departure of statutory time limits, state courts would likely establish approximate, de facto guidelines as to how long litigants have after discovery of new evidence to file their claims. The implementation of somewhat flexible timing standards of this nature, which would start to run at the date of discovery and be determined judicially on a case-by-case basis, would advance the airing of bona fide innocence claims yet still largely sustain the systemic goal of finality. Short of this reform, simply pegging the statute of limitations period for new evidence claims to the time of discovery or the time when the evidence should have been discovered as opposed to the date of conviction—essentially tolling the time period during the search for new evidence—would be a step in the right direction.” (footnotes omitted)).


196 Id. at *6–10.

fleeting observations of the killer at night. Consequently, they all recanted. It is thus quite clear that, if believed, these recantations would have had a material effect on the jury’s verdict. Indeed, without those trial testimonies, there would have been no evidence with which to convict Bermudez.

Moving to the second part of the framework, the judge would have next considered whether there was corroborating evidence for the recantations. This determination will largely depend on the specifics of the “corroboration requirement” in the relevant jurisdiction. Nevertheless, however one looks at it, there was corroborating evidence in the Bermudez case. Aside from the fact that every single witness recanted independent of each other, the state judge ultimately found unduly suggestive procedures were used to obtain the testimonies, based on evidence available when the witnesses initially recanted. More importantly, there was significant evidence that Luis Munoz was Wool Lou, and not Bermudez. Lopez had identified Munoz as Wool Lou before the trial, and he had given what was, at least in part, an inconsistent testimony. Moreover, Bermudez’s attorney discovered the existence of a man shortly after trial that more fully matched the name and description of the killer given by the prosecution’s main witness. Additionally, three female acquaintances of Lopez testified after trial, independent of Lopez, that Bermudez was not Wool Lou. More importantly, Luis Munoz’s sister identified her brother as Wool Lou. Upon identifying the corroborating evidence, the trial judge would have allowed the ultimate question of the recantations’ credibility to go to a jury or a different judge from the original trial judge, which in turn would have determined the fate of Fernando Bermudez. Instead, the trial court dismissed the recantations “based upon the long standing reluctance of courts to overturn a jury conviction on the basis of recantation evidence.” In the end, Bermudez received a day in court and was exonerated. Unfortunately, it took nearly twenty years to do so. Our hope is that, under the framework we propose here, that timeframe could be reduced from twenty years to potentially months after the initial verdict is rendered.

We are loath to second-guess trial counsel’s strategy or to presume we better understand the record before the district court in each instance. We merely wish to highlight that, given the significant amount of corroborating evidence, Bermudez could have, and should have, secured an earlier chance at exoneration. At the very least, we wish to highlight the absurdity that can often result from simple dismissals of recantations on the basis of long standing reluctance to credit recantations.

199 Id.
200 Id.
201 Id. at *9.
202 Id. at *6.
204 Id. at *31–32.
We next look at the case of Troy Davis. As with Bermudez, there was no physical evidence tying Davis to the crime—only the testimonies of nine witnesses.\(^\text{206}\) Five years later, seven of those nine witnesses recanted their trial testimonies.\(^\text{207}\) Of the two who did not recant, one was the individual the defense had targeted as the actual killer.\(^\text{208}\) Though not as material as the recantations in the Bermudez case, these recantations, if believed, would have certainly tilted the verdict in Davis’s favor. The trial court, however, simply dismissed the recantations due to “the general lack of credibility that should be assigned to recantation testimony . . . .”\(^\text{209}\) In fact, Davis was denied a hearing altogether.\(^\text{210}\)

Under the second step of the analysis, the court would have considered the presence of any corroborating circumstances. The lack of physical evidence and the sheer number of recanters—all of whom testified independent of each other—could have provided this. This would have sent the question back to a jury or a different judge for ultimate resolution. Instead, the case languished in the appeals courts for fifteen years, finally resulting in the Supreme Court taking the unusual move of ordering an evidentiary hearing.\(^\text{211}\)

Again, we do not purport to understand all the complexities of the factual record in this case better than the respective state courts. We merely suggest that there is a more efficient, and more effective, process for weeding these recantation cases out. While it appears that in the end there was not enough evidence to prove Davis’s innocence,\(^\text{212}\) that does not change the reality that it took fifteen years for a man to obtain his day in court for recantations that, at least if believed, would have had a material effect on the original verdict.

**CONCLUSION**

Recanting witnesses face both judicial skepticism and complex procedural hurdles. This Article recognizes the validity of this skepticism. It also recognizes the valid judicial interest in preserving the finality of judgments. Nevertheless, just because most recantations lack credibility does not mean they all lack credibility. Thus courts, legislatures, and jurists alike have a common interest in ensuring that standards are in place that will precisely filter the reliable from the unreliable. This Article has identified a number of structural disadvantages facing recantations,

\(^{206}\) See *Davis II*, 660 S.E.2d 354, 357–60 (Ga. 2008).

\(^{207}\) See *Davis I*, 426 S.E.2d 844, 845 (Ga. 1993).

\(^{208}\) See *Davis II*, 660 S.E.2d at 357.

\(^{209}\) Id. at 358.

\(^{210}\) Id.

\(^{211}\) *Davis IV*, 130 S. Ct. 1, 1 (2009).

though the larger problem remains in the dogmatic skepticism perpetuated by so many courts. Instead of relying on the judicial standards in place for properly analyzing recantations, courts have too often summarily dismissed recantations as “inherently unreliable.” Thus, in order to compensate for such skepticism, we propose a corroboration requirement modeled after the one contained within Rule 803(b) of the Federal Rules of Evidence.

We see no reason that in the future our criminal justice system cannot either prevent injustices similar to that suffered by Fernando Bermudez or expedite processes similar to that endured by Troy Davis. Through the collaborative efforts of state judiciaries and legislatives, we believe our criminal justice system can.
REMEDYING THE MISUSE OF NATURE

Sanne H. Knudsen*

INTRODUCTION

Like other life forms, humans depend upon nature for their survival—for food, water, clothing, shelter, and the like. Necessarily, we alter nature considerably as we go about meeting these needs. Some of these alterations are unobjectionable, or even good. Other alterations are not so good; they entail misuses of nature. When nature is misused, environmental problems are created. Air and water get polluted; soil erodes; dead zones spread; species disappear; exotics become pests. Though it is no easy task to distinguish between uses and misuses of nature, the wisdom of curtailing such misuses is clear.

A century ago, we mostly viewed nature as a source of discrete natural resources.1 We then misused nature when we dirtied our homes and diminished the flows of minerals, plants, and animals through wasteful consumption.2 Today we understand nature much differently—we realize that landscapes are complex, functioning wholes.3 We know, or ought to know, that our long-term flourishing depends, not just on specific, valuable parts of nature, but on the ability of these

---

* © 2012 Sanne H. Knudsen. Assistant Professor of Law, University of Washington School of Law. I am grateful to Professor Eric Freyfogle for his invaluable comments and discussions on earlier drafts of this work.

1 CURT MEINE, CORRECTION LINES: ESSAYS ON LAND, LEOPOLD, AND CONSERVATION 48 (2004) (“Conservation in the Progressive Era rested, first and foremost, on utilitarian and anthropocentric premises. ‘The first principle of conservation is development, the use of the natural resources now existing on this continent for the benefit of the people who live here now.’”) (quoting GIFFORD PINCHOT, THE FIGHT FOR CONSERVATION 43 (1910)); see also DONALD WORSTER, NATURE’S ECONOMY: THE ROOTS OF ECOLOGY 268 (1977) (“In the history of progressive agriculture, wild creatures had never counted for much. They failed to conform to the farmer’s productive purposes and so were seen as useless when not seen as a threat.”).


3 See Robert B. Keiter, Beyond the Boundary Line: Constructing a Law of Ecosystem Management, 65 U. COLO. L. REV. 293, 294–95 (1994) (noting that land managers “[r]ecogniz[e] that natural systems often cross jurisdictional boundaries” and that lawmakers are “beginning to speak in ecosystem terms”). As an example of the interconnectedness of ecosystems, consider that the Inuits inhabiting remote arctic areas have “among the highest PCB levels in their blood of any community on earth.” JAMES RASBAND, JAMES SALZMAN & MARK SQUILLACE, NATURAL RESOURCES LAW AND POLICY 3 (2d ed. 2009).
landscapes to maintain their ecological functioning. We have also adopted widely held beliefs that certain parts of nature—rare species, for instance—can have value and deserve protection without regard for known human benefits.

While our appreciation for what constitutes a misuse of nature has evolved with the study of ecology, our environmental laws have not. The multitude of environmental laws, although important in their own right, do not protect nature as an interconnected whole. Rather, environmental laws dissect ecosystems into discrete elements—air, water, timber, and wildlife, for example. This dissection ignores the invaluable services that ecosystems provide when allowed to function without undue disruption.

Shortcomings in current laws are particularly acute in the case of lands and resources that are privately owned. There is a widespread belief that private owners have discretion to act as they see fit so long as they avoid overtly harming neighbors. Laws tell landowners to restrain their alterations of nature in only a few settings. As a result, nature on private lands is shaped by decisions and values held by individual landowners, despite the public interests at stake.

Often the problem of protecting ecosystems is made even more complicated by the belief that we should be able to devise a solution by tweaking existing tools and operating within an inherited worldview. But what if the solution lies in another frame entirely? Imagine an unencumbered opportunity to align ecological understanding, private interests, public needs, and the law. How might we approach the problem then? In particular, if we were to devise a reformed law for curtailing and remedying misuses of nature on private and public lands, what might that law look like? How might it protect ecosystem services like water

---

4 See, e.g., James Lovelock, Gaia: A New Look at Life on Earth 9 (1979) (“[T]he entire range of living matter on Earth, from whales to viruses, and from oaks to algae, could be regarded as constituting a single living entity . . . .”); see also infra Part I.B (discussing ecosystem services and the public’s interest in private land).

5 See, e.g., Arne Naess, The Deep Ecological Movement: Some Philosophical Aspects, in Deep Ecology for the Twenty-First Century: Readings on the Philosophy and Practice of the New Environmentalism, supra note 2, at 64, 64–84 (explaining that the Deep Ecology Movement is based on the view that nature is valuable for its own sake, and not merely for human interests).


7 Joseph L. Sax, Ownership, Property, and Sustainability, 31 Utah Envtl. L. Rev. 1, 15 (2011); see also Eric T. Freyfogle, On Private Property: Finding Common Ground on the Ownership of Land 6–9, 20–24 (2007) (discussing the “partial truths” that pervade American thought and property law, such as the right of landowners to exclude all others, the protection of liberty through the protection of property rights, and the absolute nature of private property rights).

8 For example, the Clean Water Act prohibits the alteration of wetlands without a permit, even when the wetlands happen to fall on private lands. 33 U.S.C. § 1344.

9 See discussion infra Part II.C.
filtration, carbon sequestration, and soil nitrification? How would the law function given that nature is an interconnected whole that knows no boundary between public and private lands? Would it freely acknowledge the public’s interest in preserving valuable ecosystem services, even those that happen to fall on private lands? How might a reformed law work with existing federal and state natural resource laws?

A property law that restrained land misuse would go beyond protecting the interests of individual owners. It would respect broader public interests in how nature is altered. It would recognize how one piece of nature is ecologically intertwined with other pieces of nature, near and far, and how uses of one may affect others. It would consider how people living in a landscape depend upon the healthy functioning of the entire landscape. And it would take seriously, and somehow protect, the interests of other life forms and future generations.

A legal system that embraced the goal of using nature responsibly would require more than just new substantive rules on how people could use nature. It would also require a widely applicable legal remedy—a statutory cause of action—that citizens could use to stop misuses of nature and otherwise gain relief. Given the limitations of the common law, a more explicit and broad-based approach is needed in order to remedy harms to ecosystems. Some scholars have recognized this need, and many have lamented the difficulties inherent in attempting to mend a broken view of private property. None, however, have suggested the approach proposed in this Article—a rethinking and expansion of natural resource damages law.

---

10 See infra Part II.C.


12 See, e.g., Eric T. Freyfogle, The Land We Share: Private Property and the Common Good 7 (2003) (“Much of today’s conflict about property rights has arisen precisely because land is so different in law and nature.”); see also Robert J. Goldstein, Green Wood in the Bundle of Sticks: Fitting Environmental Ethics and Ecology into Real Property Law, 25 B.C. ENVTL. AFF. L. REV. 347, 352–53 (1998) (arguing that courts should recognize “green wood” in the bundle of sticks that make up property rights as a means of balancing private property rights and environmental ethics); Wilgus, supra note 11, at 102 (“Outmoded notions of property law prevent a harmonization of ecology and law.”).

13 Although very few scholars have discussed the potential for natural resource damages to serve a broader role in remedying harms to nature, Professor Peter M. Manus remarked over a decade ago that natural resource damages law “has direct and real
As currently conceived, natural resource damages are limited in scope; even in combination they cannot adequately remedy misuses of nature. Even so, these damages provide a good starting point for assessing the promise and flaws embodied in existing laws. By identifying the limits of current resource-related remedies, the changes required to better protect ecosystem health become clearer.

In search of a reformed natural resource damages law, Part I of this Article begins by exploring the idea that we should not misuse nature. It surveys current literature and explains how the idea would—if taken seriously—recast the ways we think about private property. Part II sets the stage for a reformulated law of natural resource damages by noting the gaps and limits of key environmental statutes and of the common law of property. It focuses particularly on the law’s failure to respect ecological functioning and the public’s interest in private land. Part III surveys the law of natural resource damages, which provides useful elements for constructing a broader, more ecologically grounded remedy. Part IV draws together the Article’s various parts to outline an expanded remedy for misuses of nature. The assessment is necessarily broad brushed. Its contribution is not in proposing detailed answers but in getting the challenges on the table. Part IV considers eight such challenges, each of which is foundational to the problem of protecting the healthy functioning of ecosystems when those systems know no boundary between federal, state, tribal, and private lands.

I. THE MISUSE OF NATURE—PUBLIC AND PRIVATE

The very suggestion that we provide remedies for the misuse of nature assumes that we can decide which human alterations of nature qualify as misuse. That task, however, is not so easy.

How do extractive and destructive land uses—for example, mining—fair when we delineate between use and misuse? Is a piece of land misused simply because it is destroyed? Or, is land only misused when the benefits derived from its degradation do not outweigh the benefit derived from leaving the ecosystem more intact? On what basis do we decide whether one set of benefits outweighs another? Who receives the benefits and for what reason? To what will we turn for guidance in determining where to draw the line between use and misuse—science, economics, or ethics? Will we rely on some combination of idealism and pragmatism?

Even though science can provide objective information about the character, extent, and consequences of land use, some ethical framework is needed to decide potential to translate into financial liabilities the long-term, broad ranging, and aesthetic effects of human activities on nature.” Peter M. Manus, Natural Resource Damages from Rachel Carson’s Perspective: A Rite of Spring in American Environmentalism, 37 WM. & MARY L. REV. 381, 388 (1996); see also id. at 421 (“Perhaps more than any other environmental law concept, [natural resource damages] has the potential to represent progress toward a legal structure that incorporates [Rachel] Carson’s philosophy.”).
what changes are wise or foolish, moral or immoral. In search of such a
framework, the widespread tendency of some observers is to evaluate a human-
altered landscape by comparing it with what the landscape would have looked like
without any human-caused change.\textsuperscript{14} This is the wilderness ideal, and it implies
that all human change is bad and that less change is better. This normative standard
may have value in some settings—in landscapes that we want to preserve or
restore to wilderness-like conditions—but it makes no sense when applied to
landscapes that humans inhabit to meet their needs. All species change their
environments as they go about living.\textsuperscript{15} Humans are not, and need not be, different.

Once we discard this wilderness ideal as the baseline for good land use, we
are cast adrift—we must come up with some other standard for distinguishing
between legitimate uses of nature and abuses of nature. The subject has drawn the
attention of various commentators, and their work provides a starting point. For
now, let us set aside the question of how to distinguish between use and misuse of
nature. Let us work first to achieve consensus on the point that it is necessary and
proper to do so.

\textit{A. Avoiding the Misuse of Nature}

As people living in a community, we have a collective responsibility to avoid
the misuse of nature. This obligation must be accepted before the problem of
ecosystem conservation can even begin to be solved; without a shared ethical basis
for accepting limits on the use of nature, we will never be comfortable with the
sacrifices of economics or convenience that such limits will inevitably require.

Support for our shared obligation to avoid the misuse of nature is found in
literature, religion, culture, science, and law. In particular, there is a rich literature
of scholars contemplating the roots and extent of our ethical obligations to serve as
stewards of nature, rather than merely consumers of it. The best-known and
perhaps most influential call for an environmental ethic (a land ethic, as he termed
it) has come from Aldo Leopold (1887–1948). In much-quoted language, Leopold

\textsuperscript{14} For a general introduction to the large body of literature on the wilderness ideal, see
THE GREAT NEW WILDERNESS DEBATE (J. Baird Callicott & Michael P. Nelson eds., 1998)
(a collection of previously published works that approach the concept of wilderness in a
variety of ways); THE WILDERNESS DEBATE RAGES ON: CONTINUING THE GREAT NEW
WILDERNESS DEBATE (Michael P. Nelson & J. Baird Callicott eds., 2008) (a collection of
works contributing to the wilderness conversation). One much-cited essay in the debate is
William Cronon, \textit{The Trouble with Wilderness; or, Getting Back to the Wrong Nature}, in
UNCOMMON GROUND: RETHINKING THE HUMAN PLACE IN NATURE 69, 69–90 (William
Cronon ed., 1995) (discussing the wilderness ideal and arguing that the time is ripe to look
beyond it).

\textsuperscript{15} See Carolyn M. Malmstrom, Ecologists Study the Interactions of Organisms and
Their Environment, 1 NATURE EDUC. KNOWLEDGE, no. 8, 2010 at 9, available at
http://www.nature.com/scitable/knowledge/library/ecologists-study-the-interactions-of-
organisms-and-13235586.
summed up the substance of his proposed land ethic with this deceivingly poetic measure: “A thing is right when it tends to preserve the integrity, stability, and beauty of the biotic community. It is wrong when it tends otherwise.”16 In other writings, Leopold emphasized the place of humans as citizens within a larger land community; he urged us to recognize the importance of land health. In doing so, Leopold argued for treating nature as an interconnected whole so that it retains its capacity for self-renewal:

[T]he health of the land as a whole, rather than the supply of its constituent ‘resources,’ is what needs conserving. Land, like other things, has the capacity for self-renewal (i.e. for permanent productivity) only when its natural parts are present, and functional. It is a dangerous fallacy to assume that we are free to discard or change any part of the land we do not find ‘useful’ (such as flood plains, marshes, and wild floras and faunas). Too violent modification of the natural order has repeatedly disorganized the land’s capacity for self-renewal.17

The responsibility for maintaining land health, Leopold argued, is a shared responsibility, a collective duty of stewardship: “A land ethic, then, reflects the existence of an ecological conscience, and this in turn reflects a conviction of individual responsibility for the health of the land.”18 Leopold’s words resonate in the quest to protect ecosystem services and use nature in a way that preserves nature’s functions.

Modern day scholars, most notably Professor Eric Freyfogle, have echoed Leopold’s call to root conservation in the stewardship of land and a concern for community interests.19 Freyfogle artfully sums up Leopold’s work by explaining


17 FREYFOGLE, supra note 16, at 22 (citing Aldo Leopold, Conservation (Aug. 8, 1946) (unpublished manuscript) (on file with the University of Wisconsin Digital Collections) (attached to Letter from Horace S. Fries to Aldo Leopold (Aug. 8, 1946) (on file with the University of Wisconsin Digital Collections), available at http://digicoll.library.wisc.edu/cgi-bin/AldoLeopold/AldoLeopold-idx?type=goto&id=AldoLeopold.ALCorresAK&isize=M&submit=Go+to+page&page=510)).

18 LEOPOLD, supra note 16, at 221.

19 Leopold is the touchstone for legal scholars who argue in support of an environmental ethic focused on land stewardship. See, e.g., Craig Anthony (Tony) Arnold, The Reconstitution of Property: Property as a Web of Interests, 26 HARV. ENVTL. L. REV. 281, 281 (2002) (using Leopold to begin the article’s discussion questioning the modern metaphor that describes property as a bundle of sticks); FREYFOGLE, supra note 16, at 18 (“Leopold has hardly been conservation’s only major intellect, but he remains the dominant one, long after his death in 1948.”); Alyson C. Flournoy, In Search of an Environmental Ethic, 28 COLUM. J. ENVTL. L. 63, 77 (2003) (beginning the discussion of environmental
that “conservation for Leopold focused on the totality of nature as an interconnected whole and on the need to counteract the chief forces—market economics and private property above all—that fueled harmful land-use choices.”

Picking up Leopold’s call to critically analyze private property, Freyfogle has written extensively on property law and our mistaken perception that private land use is without limits. By placing modern views on private property in a historical context, Freyfogle argues in favor of a conservation ethic that is rooted in community and supported by a public rights perspective of property law.

Property law, Freyfogle explains, is full of legal arrangements that recognize ownership in one party and use rights in another. It would be a significant step, but hardly unprecedented, for lawmakers to declare that the public owns all of nature, with private owners holding something akin to use rights, tailored to respect the common good.

Professor Joseph Sax has similarly emphasized the necessity of respecting and incorporating ethical obligations into the evolution of environmental laws. In his well-known 1970 article, Sax recognized moral and ethical limits on the ownership of nature when he argued that the public trust doctrine is a viable tool for ethical frameworks with the “formative writings” of Leopold); Goldstein, supra note 12, at 391–92 (explaining that the concept of stewardship has been developed from Leopold’s writings and land ethic).

20 FREYFOGLE, supra note 16, at 19.
21 See, e.g., FREYFOGLE, supra note 12, at 58–63 (discussing liberty and government regulation); FREYFOGLE, supra note 7 (discussing the rights of landowners, how these rights change overtime, and how these rights intertwine with other rights); Eric T. Freyfogle, Property and Liberty, 34 HARV. ENVTL. L. REV. 75 (2010) (discussing how private property and liberty are intertwined, and what liberties should be secured in the realm of private property); Eric T. Freyfogle, The Particulars of Owning, 25 ECOLOGY L.Q. 574 (1999) [hereinafter Freyfogle, Particulars of Owning] (discussing private property, popular will, and public policy).
22 FREYFOGLE, supra note 16, at 17.
23 FREYFOGLE, supra note 12, at 239.
24 In its traditional formulation, the public trust doctrine provides that states hold title to certain lands (submerged lands) in trust for the benefit of citizens of the state. See Ill. Cent. R.R. Co. v. Illinois, 146 U.S. 387, 458–60 (1892); Charles F. Wilkinson, The Headwaters of the Public Trust: Some Thoughts on the Source and Scope of the Traditional Doctrine, 19 ENVTL. L. 425 (1989). Because certain resources are held in trust, their disposition and use must be consistent with public trust responsibilities of the state. See, e.g., Phillips Petrol. Co. v. Mississippi, 484 U.S. 469, 475 (1988) (“[I]t has been long established that the individual States have the authority to define the limits of the lands held in public trust and to recognize private rights in such lands as they see fit.”).
protecting certain “gifts of nature” from privatization or exploitation. Sax has also argued for a property system based on a usufructuary model, wherein the owner of the usufruct only has a “right to uses compatible with the community’s dependence on the property as a resource.” Because public resources fall on private lands, Sax more recently urged that we need to do some “fundamentally fresh thinking” about what it means to be a landowner and that “the public has a legitimate stake in the way in which owners use land.”

In their writings, Sax and Freyfogle both echo Leopold’s theme of humans being part of a land community. They urge the subordination of private land use to community interests where appropriate, as one might expect from a stewardship-based ethic. Inherent in stewardship, of course, is the assumption that the stewards will not misuse the nature entrusted to their care. In this sense, a stewardship-based ethic has strong roots in cultural and religious doctrines.

Christianity is a good example of a dominant western religion with roots in values of stewardship. In the Christian tradition, God is the creator of Earth and has entrusted humans with the responsibility of caring for His creation. Many scholars have argued that this relationship between God, humans, and creation gives rise to a stewardship model in which humans must actively work to manage God’s creation. Professor John Copeland Nagle has meaningfully contributed to

---


29 For a detailed discussion on how attitudes towards private property changed during the Industrial Revolution and how individual rather than community interests began to dominate, see Freyfogle, supra note 12, at 65 (“Industrialism was coming, and by the time it had fully made its mark late in the century, landowners would hold a bundle of rights far different from the one they held when the century began.”); also see Joseph L. Sax Lecture, supra note 27.

30 *Genesis* 1:26–28. While Genesis provides that God entrusted man with “dominion” over the Earth, and while some have grabbed this passage as giving man license to dominate nature, religious scholars have explained that “dominion” is used elsewhere in the scriptures to “refer to a peaceful rule designed to serve those living subject to it.” John Copeland Nagle & J.B. Ruhl, *The Law of Biodiversity and Ecosystem Management* 44 (2002).

31 See John Copeland Nagle, *From Swamp Drainage to Wetlands Regulation to Ecological Nuisances to Environmental Ethics*, 58 Case W. Res. L. Rev. 787, 804–08 (2008) (providing an excellent discussion of literature focused on Christian ideas of
this area and noted that “Christian ideas of stewardship build upon the twenty-six references to ‘steward’ or ‘stewardship’ that are contained in the Bible.”32

Nonchristian cultures have similarly deep traditions that teach respect for nature. For example, in the Native American culture, man is considered “one with nature,” giving rise to a relationship in which man is a steward rather than conqueror or owner:

The Indians saw themselves as one with nature. All of their traditions agree on this. Nature is the larger whole of which mankind is only a part. People stand within the natural world, not separate from it; and are dependent on it, not dominant over it. All living things are one, and the people are joined with trees, predators and prey, rocks and rain in a vast, powerful, interrelationship... Because of this deep kinship, Indians accorded to every form of life the right to live, perpetuate its species, and follow the way of its own being as a conscious fellow creature. Animals were treated with the same consideration and respect as human beings.33

The respect that Native American and other subsistence cultures have for natural systems has been deemed an important “precursor to a land ethic.”34

Buddhism is yet another example of religion echoing a common theme of stewardship. The most important principle of “Buddhist karma-based ethics is ahinsa, the principles of non-harming and of respect for life.”35 To that end, Buddhist monks take vows to follow moral perceptions that prohibit harming the environment; “[t]here are vows for protecting the purity of the water; for not killing sentient beings who live in the earth; for not killing insects, birds, and animals; for not starting forest fires; and for respecting the life of trees, particularly ancient ones.”36

stewardship). Notably, Bruce Babbitt spoke of the stewardship responsibility that is rooted in Christian tradition when he argued that religious values as well as our responsibility as stewards remain at the heart of the Endangered Species Act. Bruce Babbitt, Between the Flood and the Rainbow: Our Covenant to Protect the Whole of Creation, 2 ANIMAL L. 1 (1996), as reprinted in THE LAW OF BIODIVERSITY AND ECOSYSTEM MANAGEMENT, supra note 30, at 39, 39–43.


33 NAGLE & RUHL, supra note 30, at 46 (quoting J. DONALD HUGHES, AMERICAN INDIAN ECOLOGY 14–17 (1983)).

34 Goldstein, supra note 12, at 388.


36 Id.
In addition to religious traditions, the notion that we ought not misuse nature finds support in our willingness to protect the environment through statutes and the common law. In other words, we would not have bothered ourselves with such a complex and comprehensive web of environmental statutes unless we recognized that unfettered pollution of rivers or destruction of habitat is unsustainable and unacceptable. This is not to say that the values motivating our various environmental laws form a cohesive or articulated ethic. But despite common quibbles over what level of pollution control or land use regulation is acceptable, our willingness to accept an impressive amount of environmental regulation reflects a common understanding that some limits on industrial society and land use are necessary to preserve a healthy civilization.

Various common law doctrines also reflect a desire to limit the use of nature when that use crosses certain thresholds. For example, the natural use doctrine provides that “[a]n owner of land has no absolute and unlimited right to change the essential natural character of his land so as to use it for a purpose for which it was unsuited in its natural state and which injures the rights of others.” Similarly, by protecting private property owners from unreasonable interference by other owners, the nuisance doctrine recognizes that private lands do not exist in isolation and that certain limits are appropriate when land use adversely impacts others. For resources protected under the public trust doctrine, limits on private property use are also appropriate to promote broader public interests. The inherent limits on land use recognized by these common law doctrines support the normative claim that we ought not misuse nature, whether to protect state’s rights, the rights of other landowners, or the public trust.

Beyond law and religion, the claim that we ought not misuse nature finds support in science itself. The science of ecology educates us on the indispensable

---

38 See Goldstein, supra note 12, at 395–96.
39 Flournoy, supra note 19, at 66 (arguing “it is not clear that environmental laws do reflect any clearly articulated ethic that should be called environmental” and suggesting “[a]s a nation, we lack an adequate understanding of the values that undergird these laws”).
40 Goldstein, supra note 12, at 395 (“The whole body of environmental laws that have been enacted over the course of the past forty years are a testament to that policy and to the willingness of society to exact economic costs for the sake of that policy. . . . [These laws] demonstrate the movement of society toward the understanding that preservation and protection of our natural environment is a positive value, and a widespread one.”).
41 Just v. Marinette Cnty., 201 N.W.2d 761, 768 (Wis. 1972); see Freyfogle, supra note 12, at 94–97 (explaining and quoting Just, 201 N.W.2d at 767–68); see also Arnold, supra note 19, at 350–51 (discussing natural use doctrine). Several states, including Wisconsin, Florida, New Hampshire, New Jersey, and South Carolina, have adopted the natural use doctrine. Id. at 350 n.344.
value of ecosystems to human survival.42 “[I]t is no exaggeration to state that the suite of ‘ecosystem services’—purifying air and water, detoxifying and decomposing waste, renewing soil fertility, regulating climate, mitigating droughts and floods, controlling pests, and pollinating vegetation—quite literally underpins human society.”43 The importance of these services lends urgency to the call for a land ethic grounded in stewardship rather than unbridled consumption. Science has been a strong voice in this regard; Professor Robert J. Goldstein posits that “[t]he science of ecology has been the most significant factor in the development of environmental ethics over the course of the last century.”44

Science, religion, culture, and law all offer arguments against misuse of nature. The more difficult question, and the one that will undoubtedly evoke more debate, is where to draw the line between use and misuse. Tackling this uncomfortable and unclear question is inevitable given that people and nature are intertwined. We cannot set aside natural resources in their entirety. We depend on their use for our survival—yet human actions have consequences. Whether we want to be or not, we are stewards of the natural world. The question is whether we are good stewards or poor stewards. And this is how we come to the ultimate question of which consequences are acceptable and which are not.45

Consider the role of science. In setting limits for environmental protection and public health, we often turn to science for answers. For example, we turn to science on issues like whether to ban leaded gasoline because of public health impacts of lead exposure,46 whether the habitat of a given species is so depleted or fragmented so as to make that species in danger of extinction,47 or whether to implement aggressive legislation to reduce greenhouse gas emissions in an effort to curb climate change.48 Likewise, science arises in debates over what protections

42 See infra Part I.B.
44 Goldstein, supra note 12, at 387.
45 Freyfogle explains that the conservation movement needs to adopt a vision of good land use “in such a way that human needs are taken seriously and satisfied insofar as possible, not shunted to the side for others to worry about.” See FREYFOGLE, supra note 16, at 176.
48 Through the Department of Commerce Appropriations Act of 2008, Pub. L. No. 110-161, 121 Stat. 1844, Congress mandated that the National Oceanic and Atmospheric Administration (NOAA), together with the National Academy of Sciences, to establish a committee and produce a series of studies to ensure that climate decisions are informed by
ought to be afforded to ecosystem functions.\textsuperscript{49} Indeed, the science of ecology has been a catalyst for contemplating how environmental regulation, common law doctrines, and property law theories need to adapt to new knowledge about the interconnectedness and indispensable value of natural systems.\textsuperscript{50}

Science can predict the likely consequences of our actions, telling us whether they are reversible or irreversible, localized or widespread.\textsuperscript{51} But science does not provide complete answers. As Professor Freyfogle has explained, part of the problem is that science does not, and cannot, provide all the answers because science is \textit{descriptive} in nature—it tells us what nature does and what consequences of our actions might be, but does not tell us what is right or wrong in terms of degree of acceptable impacts to land.\textsuperscript{52} Nor does it tell us what risk levels we should be willing to accept to promote public health. For example, on the issue of endangered species protection, science could tell us that the proposed site of the Tellico dam in the 1970s would have destroyed the only remaining snail darter habitat.\textsuperscript{53} Science could not tell us whether this little fish and the associated ecological benefits were worth the economic loss of destroying a nearly completed dam that would produce 200 million kilowatt hours of hydroelectric power and save an estimated 15 million gallons of oil in the midst of a national energy crisis.\textsuperscript{54}

Similarly, science plays a central role in the discussion of climate change.\textsuperscript{55} But global climate systems are complex, and there are many questions still unanswered on this issue. For instance, what will be the consequences of climate change in Utah, Florida, or India? Science cannot necessarily answer these questions with certainty at this time, and there are other questions that science

\begin{flushright}
the best possible scientific knowledge. The final report, \textit{Comm. on America’s Climate Choices, National Research Council, America’s Climate Choices (2011), available at} http://dels.nas.edu/Report/Americas-Climate-Choices/12781, was released in May 2011.\textsuperscript{49} For example, the state of Washington has passed a Growth Management Act that requires local governments to use the best available science when reviewing and revising their policies and regulations on wetlands. \textit{Wash. Rev. Code §§ 36.70A.010–70A.904} (2011). The state’s Department of Ecology received funding from the U.S. Environmental Protection Agency to synthesize the science on wetlands in support of its mandate to local governments. \textit{See Wetlands: Nature’s Sponges, Nurseries, and Water Filters, Wash. St. Dep’t Ecol.}, http://www.ecy.wa.gov/programs/sea/wetlands/index.html (last visited Feb. 25, 2012).\textsuperscript{50}

\textsuperscript{51} See infra notes 197 and accompanying text for a brief survey of the scholarship aimed at aligning environment laws with ecological understanding.\textsuperscript{152}

\textsuperscript{52} \textit{Freyfogle, supra} note 16, at 159.


\textsuperscript{55} \textit{See, e.g., Ackerman & Heizerling, supra} note 46.
cannot answer at all, like how many species are we willing to put at risk through habitat loss? How much land are we prepared to lose along our coasts as a result of the rising sea levels predicted to accompany climate change? Similarly, how many communities in developing island nations are we willing to expose to flooding? These last questions turn on our willingness to accept certain risk levels, which are fundamental issues of policy or values, not science. Ultimately, our line drawing will be a blend of ethics and science.

Like the role that it has played in many issues before, science will undoubtedly play a central role in distinguishing between use and misuse of nature. Science will help decide when ordinary, expected, and necessary uses of nature cross the line into misuses that ought to be limited for the greater benefit of society. Science cannot, however, provide the wholesale answer. Some measure of ethics and values will have to be considered.

**B. The Public’s Interest in Private Land**

If we are to remedy the misuse of nature, we need more than an ethical framework guiding our actions; we need to be specific about the scope of lands for which we are prepared to accept responsibility. In particular, this means that any enduring remedial framework must address lands in private ownership. Over sixty percent of lands in the United States are privately owned. These lands are not neatly separated from public lands, but instead form a pattern of land ownership that resembles something of a patchwork quilt. Constructing a framework to remedy the misuse of nature that ignores limits on private land use would fail the most basic challenge of respecting land as an interconnected whole.

---

56 Professor Freyfogle has grappled with many of these issues. See Freyfogle, supra note 16, at 152–59 (suggesting some normative benchmarks for determining what is good land use. Prominent among those benchmarks are human utility, broadly defined; and ethical obligations for future generations).

57 See Flournoy, supra note 19, at 80–88 (describing the various ethical frameworks that drive the environmental philosophical debate, ranging from anthropocentric utilitarian to biocentric communitarian).

58 Freyfogle, supra note 12, at 16 (“Leading American conservationists have repeatedly offered a stern warning: because private land is so extensive, we are unlikely to achieve conservation goals or halt the processes of degradation unless the country revises (yet again, as we shall see) what it means to own nature.”).


60 For a case study example of how the comingling of public and private lands challenges land conservation goals, see John D. Erickson & Sabine U O’Hara, From Top-Down to Participatory Planning: Conservation Lessons from the Adirondack Park, United States, in BIODIVERSITY AND ECOLOGICAL ECONOMICS: PARTICIPATION, VALUES, AND RESOURCE MANAGEMENT 146–61 (Luca Tacconi ed., 2000).
It would amount to nothing more than an incomplete and fragmented vision, doing little to synthesize ecological understanding with legal recourse.61

Because no discourse on proper land use is complete without considering private lands, the normative claim introduced above should be more finely stated: we ought not misuse nature, whether on public or private land. Accepting this proposition requires us to engage in the more difficult task of healing schisms between private property and environmental stewardship. We must answer thorny questions like what limits are proper to the use of nature on private lands? Or, put differently, what public ownership rights are inherent in private land? Answers to these questions must be consistent with the common law of property and must respect private property as an institution.62 Otherwise, legislative efforts will undoubtedly meet staunch resistance from landowners and potentially run afoul of regulatory takings law.

The following subsections examine public ownership of nature on private lands. The first subsection considers from an ecological perspective why the public should have an interest in nature on private land. The second subsection goes on to examine the false dichotomy that clouds perceptions of private property, exploring property law’s capacity to protect public welfare on private lands.

1. The Ecological Case For Public Interest In Private Lands

There are certain elements of the natural world that are vital to a functioning society and therefore should not be entirely under private control. The law has recognized this truth in other settings for quite some time, as illustrated by the public trust doctrine’s protection of public rights in navigable waterways.63 As the study of ecology has evolved, we have begun to understand how other elements of nature also provide invaluable benefits to society at large and are therefore worthy of protection against unfettered private use.64 In particular, our understanding of ecology has matured to the point where we now appreciate that nature is an

61 Bradley C. Karkkainen, Collaborative Ecosystem Governance: Scale, Complexity, and Dynamism, 21 VA. ENVTL. L.J. 189, 212–13 (2001) (explaining the need to involve multiple landowners (public and private) because single parcels are too small to address ecosystem issues); see also Adler, supra note 59, at 302 (“Without conservation on private lands, meaningful ecological conservation cannot be achieved.”).

62 FREYFOGLE, supra note 12, at 122 (“The protection of lands and communities is a goal that inevitably needs balancing against the benefits that the community gets when landowners are reasonably secure in their entitlements.”).

63 See supra note 24 for a brief discussion of the public trust doctrine.

64 See, e.g., Nagle, supra note 31, at 789–97 (describing how an evolved ecological understanding of wetlands has changed its treatment in the law from one of disgust to one of reverence and protection); see also Gretchen C. Daily, Introduction: What Are Ecosystem Services?, in NATURE’S SERVICES: SOCIETAL DEPENDENCE ON NATURAL ECOSYSTEMS 1, 3–4 (Gretchen C. Daily ed., 1997).
interconnected whole and that vital societal functions are derived from protecting certain ecosystem services.65

“Ecosystem services” are generally defined as “the conditions and processes through which natural ecosystems, and the species that make them up, sustain and fulfill human life.”66 For example, freshwater systems supply drinking water, habitat, and instream benefits (like recreation, transportation, and flood control).67 Forests prevent erosion, reduce sedimentation, and sequester carbon.68 Grasslands conserve soils, maintain the genetic library, and stabilize the composition of the atmosphere.69 Soil moderates the climate, supports plants, and disposes of dead organic matter.70 Natural predators, parasites, and pathogens control pests that threaten agricultural crops.71

Because nature is a network of systems and services, all alterations of the land have consequences. One cannot destroy a wetland, for instance, without impacting the ecosystem service of water retention.72 Similarly, one cannot develop forested watersheds without degrading the service of water purification.73

Not only is nature interconnected, but nature and its services also cannot readily be substituted by technological enterprise and human engineering. In an article discussing the obstacles of regulating ecosystem services, Professor James Salzman recounts the failed Biosphere II experiment in the early 1990s. It is an example of the complex underpinnings of nature’s systems:

---

65 Daily, supra note 64, at 3.
66 Id.
67 Sandra Postel & Stephen Carpenter, Freshwater Ecosystem Services, in Nature’s Services: Societal Dependence on Natural Ecosystems, supra note 64, at 195, 195–207.
69 Id. Grasslands contribute to the genetic library because of the abundant biodiversity located in that ecosystem. Also, a large fraction of domesticated species originated from grasslands, and the wild populations related to those domesticated species (along with their associated pests and pathogens) continue to thrive in grasslands. These areas are, therefore, most likely to provide information and strains of species that are resistant to disease. Osvaldo E. Sala & José M. Paruelo, Ecosystem Services in Grasslands, in Nature’s Services: Societal Dependence on Natural Ecosystems, supra note 64, at 237, 242–43.
70 Gretchen C. Daily et al., Ecosystem Services Supplied by Soil, in Nature’s Services: Societal Dependence on Natural Ecosystems, supra note 64, at 113, 113–32.
71 Rosamond L. Naylor & Paul R. Ehrlich, Natural Pest Control Services and Agriculture, in Nature’s Services: Societal Dependence on Natural Ecosystems, supra note 64, at 151, 151–74.
73 Id.
In understanding the power and challenge of ecosystem services, it is best to start our story fifteen years ago, beneath the blazing Arizona desert sun. There, on September 26, 1991, walking through a crowd of reporters and flashing cameras, eight men and women entered a huge, glass-enclosed structure and sealed shut the outer door. Their 3.15 acre miniature world, called Biosphere II, had been designed with no expense spared to re-create the conditions of the earth. . . . Biosphere II sought to re-create a truly self-sustaining environment, complete with designer rainforest, ocean, marsh, savanna, and desert habitats. The eight plucky adventurers, so-called “Bionauts,” intended to remain inside this micro-world for two years. By sixteen months into their adventure, however, oxygen levels had plummeted 33%, nitrous oxide levels had increased 160-fold to levels causing brain damage, ants and vines had overrun the vegetation, and nineteen of the twenty-five vertebrate species had gone extinct, as well as all of the pollinators. The experiment was abandoned.74

Despite the budget of $200 million, recreating the basic services that sustain human life proved to be both a daunting feat and a failed experiment.75 Given the troubles and expenses that the Biosphere project encountered in its attempts to replicate nature’s intricacies, we should not expect that we can simply engineer our way out of ecological problems. Some foresight in protecting nature from misuse is a necessary step towards creating a sustainable and healthy future for humankind.

The knowledge that nature is interconnected and that ecosystem services are valuable to human existence is not new. Lamenting soil erosion caused by deforestation, Plato once wrote:

[w]hat now remains of the formerly rich land is like the skeleton of a sick man with all the fat and soft earth having wasted away and only the bare framework remaining . . . The soil [used to be] deep, it absorbed and kept the water . . . , and the water that soaked into the hills fed springs and running streams everywhere.76

In the more modern era, George Perkins Marsh heralded the role of microorganisms is sustaining life when he explained that “[c]arth, water, the ducts and fluids of vegetable and of animal life, the very air we breathe, are peopled by

75 Id.
76 Daily, supra note 64, at 5–6 (quoting Plato in DANIEL HILLEL, OUT OF THE EARTH: CIVILIZATION AND THE LIFE OF THE SOIL 104 (1991)).
minute organisms which perform most important functions in both the living and inanimate kingdoms of nature.”

Aldo Leopold and Rachel Carson have similarly trumpeted the importance of understanding and preserving the valuable services that intact ecosystems can provide. Leopold eloquently wrote in 1948:

If the land mechanism as a whole is good, then every part is good, whether we understand it or not. If the biota, in the course of aeons, has built something we like but do not understand, then who but a fool would discard seemingly useless parts? To keep every cog and wheel is the first precaution of intelligent tinkering.

Some years later, in the famed Silent Spring, Rachel Carson captured the interconnectedness of humans and living systems through the simple words “in nature nothing exists alone.”

Though our understanding of ecosystem services has been percolating in the minds of some for many years, scholars trace growing mainstream appreciation for these services to the publication of Nature’s Services in 1997. This was the mark of a new era in which ecologists, economists, and lawyers began to systematically examine the contributions of ecological services to social welfare. Written by well-respected scientists and economists, Nature’s Services explained in great detail the services that ecosystems provide to society, with separate chapters devoted to climate, biodiversity, soil, pollinators, pest control, and major biomes (for example, oceans, freshwater, forests, grasslands). Nature’s Services was also the first serious attempt to put a dollar figure on the value of ecosystem services—pollinators, for example, were estimated to contribute a $4–7 billion to the United States agricultural economy each year.

---

77 GEORGE P. MARSH, MAN AND NATURE; OR, PHYSICAL GEOGRAPHY AS MODIFIED BY HUMAN ACTION 123 (1864).
79 RACHEL CARSON, SILENT SPRING 51 (1962). Carson similarly emphasized the interconnectedness of living systems when she stressed that the soil community “consists of a web of interwoven lives, each in some way related to the others—the living creatures depending on the soil, but the soil in turn a vital element of the earth only so long as this community within it flourishes.” Id. at 56.
80 See, e.g., Ruhl & Salzman, supra note 43, at 158.
82 See Daily, supra note 64, at 6–10.
83 Gary Paul Nabhan & Stephen L. Buchmann, Services Provided by Pollinators, in NATURE’S SERVICES: SOCIETAL DEPENDENCE ON NATURAL ECOSYSTEMS, supra note 64, at 133, 141.
Since the release of Nature’s Services, there have been numerous scientific studies describing and valuing ecosystems services around the world.\textsuperscript{84} One salient example is the Millennium Ecosystem Assessment (“the Assessment”), which consisted of a series of reports published between 2001 and 2005 assessing how changes in ecosystem services impact human well-being.\textsuperscript{85} The Assessment was modeled on the Intergovernmental Panel on Climate Change and involved approximately 1,360 experts from over 95 different countries.\textsuperscript{86} It starts with the premise that “[e]veryone in the world depends completely on the Earth’s ecosystems and the services they provide, such as food, water, disease management, climate regulation, spiritual fulfillment, and aesthetic enjoyment.”\textsuperscript{87} One of the Assessment’s major findings is that sixty percent of the ecosystem services studied “are being degraded or used unsustainably, including fresh water, capture fisheries, air and water purification, and the regulation of regional and local climate, natural hazards, and pests.”\textsuperscript{88} Stop and think about those two statements in concert. Ecosystem services are our lifelines. We are destroying over half of those lifelines.

2. False Dichotomy Between Natural Resources Protection on Public and Private Land

The importance of ecosystem services to society at large, and the fact that many of these vital services happen to fall on private lands, demonstrate the need for a legal regime that protects those services from unfettered private control. Protecting nature’s services, therefore, requires that we embrace property law’s legitimate role in serving public welfare as well as private interests.

The idea that public ownership rights exist on private lands is neither new nor radical. Many property scholars, having traced the historical foundations of property law in much detail, argue for a more public-rights based model of property.\textsuperscript{89} Based on a more mature understanding of the interconnectedness of nature and ecosystem services, scholars like Professors Sax and Freyfogle have concluded that reviving the voice of public ownership and communal needs to the

\textsuperscript{84} For citations to works published on ecosystem services, see Ruhl & Salzman, supra note 43, at 161.
\textsuperscript{86} Id. at viii.
\textsuperscript{87} Id. at 1.
\textsuperscript{88} Id.
institution of property is both correct and necessary.\textsuperscript{90} Reviving that voice is necessary to prevent undue and irreversible damage to ecosystem services. It is also correct in that the institution of property both protects societal needs and preserves individual rights.

To understand why a public rights view of property is consistent with natural resource protection and property law, consider how property was conceived at the birth of this nation. As a historical matter, property rights were created to serve public welfare and not just individual interests.\textsuperscript{91} Indeed, the reigning republican view of property recognized that "property is held by the individual in trust for the benefit of society as a whole."\textsuperscript{92} Implicit in that statement is the understanding that private property exists only because of society’s agreement to protect the interests of the individual.\textsuperscript{93} In exchange for that protection, private interests must in some cases give way to the greater needs of society.\textsuperscript{94}

Professor Freyfogle has been a prominent voice in this important discourse. In \textit{The Land We Share}, he reminds us that limits on land uses are deeply engrained in the institution of private property. Those limits prevent harm to other landowners or society as a whole. He explains that at the beginning of this nation’s history, "[c]ourts agreed that legislatures possessed broad powers to control how private land was used. Even for Chief Justice Roger Taney, a Southern conservative and author of the proslavery Dred Scott decision, private desires were properly subordinated to public need."\textsuperscript{95}

Early on, and throughout much of the eighteenth century, property law continued to yield significant respect for societal welfare.\textsuperscript{96} For example, "[o]wners of attractive sites for water mills could have their lands seized if they failed to use them in the public interest."\textsuperscript{97} Indeed, many colonial and early federal-era laws "went well beyond the avoidance of harm to impose affirmative duties on

\textsuperscript{90} \textit{Freyfogle, supra} note 12, at 9 ("It is possible, I believe, for the Constitution to give lawmakers substantial power to redefine the rights of landowners, bringing them up to date and promoting conservation, while at the same time heightening the protections landowners enjoy against unfair government treatment."); \textit{Sax, supra} note 26, at 1451–52; \textit{see also Goldstein, supra} note 12, at 429; \textit{Wilgus, supra} note 11, at 100 ("[T]raditional notions of property law must contend with new scientific discoveries about our environment as well as popular notions of an environmental ethic.").

\textsuperscript{91} \textit{Jerry L. Anderson, Takings and Expectations: Toward A “Broader Vision” of Property Rights,” 37 U. Kan. L. Rev. 529, 530 (1988) ("[T]he history of views of private property suggests that this country was founded on the ideal of protecting not only private property but also the social good.").

\textsuperscript{92} \textit{Id.} at 532.

\textsuperscript{93} \textit{Id.} at 533.

\textsuperscript{94} \textit{Id.}

\textsuperscript{95} \textit{Freyfogle, supra} note 12, at 5.

\textsuperscript{96} \textit{Id.} at 58–63.

\textsuperscript{97} \textit{Id.} at 60.
private owners to help achieve social aims.\(^{98}\) Evidence of community-minded views of property rights found its way into the Supreme Court’s 1887 decision \textit{Mugler v. Kansas}.\(^{99}\) Writing for the majority, Justice Harlan reiterated that “all property in this country is held under the implied obligation that the owner’s use of it shall not be injurious to the community.”\(^{100}\)

Though many individual laws regulated land use in the public interest throughout the eighteenth century,\(^{101}\) the recognition of public interests in private lands was expressed most directly and uniformly through the common law doctrine \textit{sic utere tuo ut alienum laedas} (“\textit{sic utere tuo}”). This doctrine translates to mean, “use your own so as to cause no harm”; it is referred to as the “no harm” rule for short.\(^{102}\) Under this doctrine, otherwise reasonable land uses were restricted if they caused harm to other landowners or to the public at large.\(^{103}\)

Over time, society’s perception of what constitutes harm (and benefits) to the public at large has changed. In the beginning, the “no harm” rule gave rise to applications like the natural-flow rule, where landowners had the right to use water only in ways that left downstream users unaffected.\(^{104}\) By contrast, during the industrial revolution, society began to elevate the importance of economic development such that developing natural resources was unequivocally seen as a public benefit.\(^{105}\) The result is a conception of property that would—when viewed purely from a common law perspective—permit landowners to “use land for maximum gain,” even when such use “severely disrupted neighbors.”\(^{106}\)

The common law picture of private property is not complete, however. The role of protecting natural resources from harm has shifted principally from common law to public, or statutory, law. Aside from the common law doctrine of nuisance, environmental statutes are now primarily responsible for regulating land uses and protecting nature from harm.\(^{107}\) Therefore, when environmental regulations are criticized for undermining individual freedom on private lands, we would be wise to remember that those regulations help fill in the complete picture of private property—one that has always embodied a healthy respect for public interest (whether in the common law itself or in the public law).

Though the public-rights based view of private property has a strong legacy, another voice could be heard in the property debate, one that would grow louder and more forceful over time—the voice of Federalists like James Madison who stressed the importance of individual rights and sought freedom from

\(^{98}\) Id. at 62.

\(^{99}\) 123 U.S. 623 (1887).

\(^{100}\) Id. at 665.

\(^{101}\) FREYFOGLE, supra note 12, at 60.

\(^{102}\) Id. at 67.

\(^{103}\) See id.

\(^{104}\) Id.

\(^{105}\) See id. at 65–99.

\(^{106}\) Id. at 72–73.

\(^{107}\) See id. at 83–84.
governmental interference as much as possible. Indeed, somewhere around the turn of the nineteenth century, with the coming of the Industrial Revolution, pressures for development and economic advances began to overshadow detrimental consequences of destroying public resources for private gain. Soon these pressures changed perceptions of property rights and gave rise to a divisive vision of property, pitting public welfare against private interests.

The 1922 Supreme Court decision Pennsylvania Coal Co. v. Mahon aptly illustrates and foreshadows the dichotomous views that continue today. Justice Holmes, writing for the majority, concluded that to steer clear of the Fifth Amendment, statutes could limit uses of private property only to a minor degree. On the other hand, Justice Brandeis echoed public-rights based views of private property in his dissent when he urged that no regulatory takings should be found where the statute “merely prevents the owner from making a use that interferes with the paramount rights of the public.”

These two voices have continued their uneasy duet in the property debate throughout the settlement and development of this nation. Today, the voice of those who would elevate private property rights and financial interests above the public good can most readily be found in leaders of the “wise use” movement. More specifically, under the banner of regulatory takings, wise use leaders have rounded up a broad range of economic and political interests, including developers, small property owners, and timber companies into the so-called “property rights” movement.

108 Anderson, supra note 91, at 533–34.
110 Id.; see also Freyfogle, supra note 7, at 20–21 (describing as a “universal benchmark” the idea that property can be held in absolute ownership and that any government regulation therefore curtails private property rights).
111 260 U.S. 393 (1922).
112 See id. at 415–16. For a discussion about how the property rights debate has evolved and continued throughout the twentieth century, see Anderson, supra note 91, at 537–62.
113 See Anderson, supra note 91, at 538 (summarizing divisive views in Pennsylvania Coal).
114 See id. (quoting Pennsylvania Coal, 260 U.S. at 417 (Brandeis, J. dissenting)).
In their newsletters, journals, books, and presentations, leaders of the wise use movement routinely argue that environmental regulations are destroying private property rights. This rhetoric asserts that environmentalists have generated an avalanche of regulatory red tape that threatens to suffocate small property owners and destroy industrial civilization altogether. “Supporters of both privatization and strong individual property rights distrust—and at times, even scorn—government regulation conducted in the name of the public interest.”

Although extreme, the property rights movement is the product of divisive rhetoric that pits the public interest against the private one. This world—in which regulation of private property for the public good is characterized as an intrusion of rights—is premised on a false dichotomy between public interest and private rights. It is false not because divisive viewpoints have been absent from our history, but because the views that pit private and public rights against one another (1) do not reflect the physical realities of nature and (2) do not appreciate property law’s capacity for simultaneously protecting private rights and public resources.

On the first point, natural resources do not differentiate between private and public lands. Resources that are important to the public interest, such as privately owned wetlands, cannot be segregated from private ownership any more than valuable topsoil can be dug up and set aside for cultivating society’s food. So long as land is privately held, certain resources that are invaluable to society’s healthy functioning inevitably will be in private control. In other words, it would have taken place over the last 11 years correspond quite closely to a blueprint for takings doctrine proposed by Professor Richard Epstein in his now-famous book called *Takings, Private Property and the Power of Eminent Domain.* Epstein “advanced a notion of property rights under which individuals should not be forced to bear community burdens.” Klein, *supra* at 1186. Epstein called for a broad application of the Fifth Amendment’s Takings Clause, admitting that his expansive interpretation would call into question “many of the heralded reforms and institutions of the twentieth century: zoning, rent control, workers’ compensation laws, transfer payments, progressive taxation.” RICHARD A. EPSTEIN, *TAKINGS: PRIVATE PROPERTY AND THE POWER OF EMINENT DOMAIN*, at x (1985).


See, e.g., Klein, *supra* note 116, at 1166.

Anderson, *supra* note 91, at 536 n.38 (“[W]hatever the state of its title, one parcel of land is inextricably intertwined with other parcels, and . . . causes and effects flow across artificially imposed divisions in the land without regard for legal boundaries. The land simply cannot be neatly divided into mine and yours.” (quoting Donald W. Large, *This Land Is Whose Land? Changing Concepts of Land as Property*, 1973 WIS. L. REV. 1039, 1045)).
be impossible to set aside or buy up every piece of private land that served some public function. Not only that, but ecosystem health will often turn on more than the health of an individual parcel of land; the value of some resources manifest as part of a larger landscape that spans multiple parcels and multiple landowners. So while some land uses are harmful on their own because of soil erosion or critical habitat destruction, other land uses are harmful in the aggregate. “When too many fields in a watershed are plowed, or too many fields are drained, or too much wildlife habitat is altered, or too many homes are built in an area, or too much impervious pavement distorts hydrologic patterns, the ecological status of entire landscapes can be degraded.”

The issue of aggregate harm is worth probing because it helps to explain the need for a coordinated legal system that identifies and remedies the misuse of nature. In general, the idea that private land practices can be individually insignificant but collectively destructive is analogous to Garrett Hardin’s “tragedy of the commons.” As Hardin explains, unregulated and limited resources that are shared communally will eventually be depleted. In this case—where ecosystem health is depleted from unregulated and uncoordinated land use—the problem is similar in that ecosystem services are common resources benefitting society as a whole. But the problem is also different in that ecosystems services cross many land parcels, some of which are publicly managed and other of which are privately held.

The problem with ensuring ecosystem health is more like a “tragedy of fragmentation.” Our current predicament arises from uncoordinated land use practices despite the interconnectedness of nature. As such, Hardin’s suggestion that land be divided and placed into private ownership is not a satisfactory solution for resolving current land use problems. In fact, one might consider taking Hardin up on his other, less-discussed, suggestion that would mitigate aggregate harm by placing restraints on land use practices: “mutual coercion, mutually agreed upon.” Consistent with Hardin’s analysis, the solution to ecosystem health will require considering land use practices in the aggregate and some measure of collective restraint.

In light of our greater understanding about what constitutes harmful land use and what comprises the public welfare, revitalization of common law doctrines

---

120 [Freyfogle, supra note 12, at 221.]
122 To illustrate his theory, Hardin supposes that several herders share a field for grazing cows. *Id.* at 1244–45. Even though overgrazing will eventually destroy the resource, Hardin suggests that it is in each herder’s interest to put additional cows into the field given that all the benefits of grazing and additional cow will be reaped by the individual but all the damage will be shared by the group as a whole. *Id.* Central to Hardin’s example is that the commons is unregulated. *Id.*
123 [For an excellent discussion and analysis of Hardin’s work and connection to modern day natural resource protection, see Freyfogle, supra note 12, at 158–74.]
124 *Id.* at 160, 171–72.
such as *sic utero tuo* is needed for property law to keep pace with new ecological understanding. These fundamental doctrines and their specific applications need to reappear in property law and regain proper respect alongside other, well-funded property rights such as the right to exclude. Indeed, as part of this effort at revitalization we would be wise to recall the corollary common law principle of *salus populi suprema lex est*—the welfare of the people is the supreme law. 125

The Wisconsin Supreme Court’s 1972 decision *Just v. Marinette County*126 is an example of how property law can be more mindful of ecological realities and begin to shift the balance of property rights more toward an equilibrium between private and public welfare. 127 In *Just*, the court upheld the constitutionality of a shoreline protection law and prevented the Justs from filling in a wetland on the south shore of Lake Noquebay in Marinette County. In so holding, the court explained:

> An owner of land has no absolute and unlimited right to change the essential natural character of his land so as to use it for a purpose for which it was unsuited in its natural state and which injures the rights of others. The exercise of the police power in zoning must be reasonable and we think it is not an unreasonable exercise of that power to prevent harm to public rights by limiting the use of private property to its natural uses. 128

Advocates of staunch private property rights would predictably be aghast at the suggestion that all private lands are inherently limited by public ownership and must give way to overarching societal interests. 129 Again, such individualistic views of property are born from a divisive rhetoric that fails to appreciate that private and public interests are inextricably intertwined; 130 private interests are still served when property law imposes reasonable limitations on land use.

In all instances, property law resolves land use conflicts, either by allowing or restricting an activity. Private interests are simultaneously expanded and curtailed. The law simply decides whose private interests are given preference—those that would use land intensely or those that would use their land in a less-disruptive manner. On this point, Freyfogle concludes, “there is no ‘pro-private property’

125 The need for the property law to keep pace with changing perceptions of public welfare is not to argue that the common law can—or should—carry the weight of protecting nature from misuse. This may well be a role best served by legislation. Nonetheless, property law and the balance between private and public rights in the ownership of nature must be restored in order for legislated protection to have legitimacy.
126 201 N.W.2d 761 (Wis. 1972).
127 See id. at 771 (discussing land value in terms of balancing private and public use).
128 Id. at 768.
129 See supra notes 118–120 and accompanying text.
130 See supra notes 118–120 and accompanying text.
position that the law can take.” 131 This is not to say that the policy choices have no material consequences. Rather, elevating intensive land uses above all else will harm other landowners, the public at large, and nature itself. In this way, intensive land use weakens the property rights held by others, such as the right to enjoy land free from unreasonable interference. Unfettered rights to use land without regard to others, therefore, threaten the very institution of property. In other words, the security of private property rights depends on limits. 132

In the end, although divisive rhetoric is simple and effective for drumming up controversy, private interests and public welfare cannot be separated out and pitted against one another. Like nature itself, they are interconnected. Achieving ecosystem health will require a more complete understanding of property. In doing so, the voice of public welfare must return to any conversation about proper uses of nature.

Fortunately, property is not a static institution. Instead, an important and well-recognized aspect of property law is its flexibility to accommodate new assumptions of what the public interest contains: “As a strictly legal matter, landowners possess only such rights to develop as property law allows at any given time. Property law, like other law, evolves to keep in line with shifting communal needs.” 133 It is because property law is capable of evolving that Professor Sax has urged a new definition of property in light of what he termed the “economy of nature.” 134 This new definition would respect the interconnectedness of nature and the ability of undeveloped lands to serve valuable ecosystem functions. 135 More specifically, it would focus less on individual dominion, approach rights in land from an ecosystem’s perspective, recognize that different lands play different kinds of roles, and impose affirmative obligations on landowners to “protect natural services, with owners functioning as custodians as well as self-benefitting entrepreneurs.” 136 Sax recognizes the mismatch between viewing private property as an individualistic institution, a more sophisticated understanding of the natural world, and the rich history of property law in serving communal needs. 137

If indeed there is a role for public ownership in private lands, one might object to such an approach on the basis that achieving a balance between protection of public resources and respect for private lands would be legally or logistically difficult. It seems much easier, albeit detrimental, to accept a world with bright line divisions between “yours” and “mine.” Fortunately, there are historical examples that shed light on how a workable relationship between public rights and private ownership can be achieved. The area of water law and water

131 FREYFOGLE, supra note 12, at 20.
132 See id. at 18.
133 Id. at 123; see also id. at 208–09 (describing how perceptions of the communal good have changed over time and how property law has responded accordingly).
134 Sax, supra note 26, at 1442.
135 See id. at 1451.
136 Id.
137 Id. at 1442–46.
rights is a good example. Given the continuous, interconnected characteristics of water, not to mention its importance to public welfare, water has never been considered a good candidate for total privatization.\(^{138}\) As a result, water has long belonged to the people collectively, even when it runs underneath or through private property—private owners possess only conditional rights to use the water resource.\(^{139}\) Their rights are limited to “reasonable” uses that are “beneficial.”\(^{140}\) And, when community interests require, private property interests in water are curtailed to serve public needs.\(^{141}\)

Likewise, the public trust doctrine, which historically safeguarded public rights in navigable waterways, embodies the notion that the public possesses inviolable rights in certain natural resources.\(^{142}\) Wildlife has enjoyed similar protection from unfettered private ownership because of its importance to public welfare and natural movement in and out of private lands.\(^{143}\) For these historically elite resources, the subordination of private ownership to public rights continues to this day.\(^{144}\)

Like water, many ecosystem functions serve the public welfare and are manifestations of the interconnectedness of nature.\(^{145}\) Just as laws have recognized the importance of recognizing and protecting public rights for resources like water and wildlife, the time is ripe to expand our vision and develop a legal framework that recognizes public ownership in ecosystems services. As discussed above, a legal framework that gives the public a voice in the misuse of nature on private lands need not run afoul of private property rights.\(^{146}\) Rather, such a framework would merely restore a long-recognized but recently forgotten element of property

\(^{138}\) FREYFOGLE, supra note 12, at 231; Sax, supra note 26, at 1452–53.
\(^{139}\) See, e.g., Sax, supra note 26, at 1452–53.
\(^{140}\) FREYFOGLE, supra note 12, at 231.
\(^{141}\) Sax, supra note 26, at 1453.
\(^{143}\) FREYFOGLE, supra note 12, at 230–38 (describing unique property relationships for water and wildlife).
\(^{144}\) For example, several courts have upheld the power of government to prohibit fences that would encumber movement of wildlife. See, e.g., Dep’t of Cmty. Affairs v. Moorman, 664 So. 2d 930 (Fla. 1995) (upholding the validity of a law prohibiting fences that interfered with an endangered species of deer); New York v. Sour Mountain Realty, Inc., 714 N.Y.S.2d 78 (App. Div. 2000) (upholding an order directing an landowner to remove a 3500 foot long snake-proof fence that blocked the migration of an endangered snake species). For a further illustration of the subordination of private interests in wildlife management, see Clajon Prod. Corp. v. Petera, 70 F.3d 1566, 1575 (10th Cir. 1995) (rejecting private landowner’s argument that they had the exclusive “right to hunt the ‘harvestable surplus’ from their land—i.e., the excess animals available for hunting which were produced on their land”).
\(^{145}\) See supra notes 88–110 and accompanying text.
\(^{146}\) See supra notes 93–117 and accompanying text.
law: public ownership. What such a framework might look like and how it might provide remedies for the misuse of nature is taken up in Part III of this Article.

II. THE NEED FOR A STATUTORY CIVIL REMEDY

If we accept the two foundational claims set forth in Part I—that we ought not misuse nature and that the public has an interest in private lands—then we have begun to redefine what it means to own and harm nature. With that new conception of harm in place, we can begin to think about remedies. It might be useful, however, to first consider why a new remedy is necessary within natural resource and environmental law. In other words, why are lawmakers, practitioners, and scholars still talking about a broken system of environmental protection when we have “ten thousand commandments” and hundreds of thousands of pages of environmental statutes and regulations on the books?

For the purposes of appreciating the need for a new statutory civil remedy to protect nature from misuse, there are three inherent limitations within the existing legal framework that are important to explore. First, most environmental and natural resources laws protect discrete elements of nature, not nature as an interconnected whole. Second, and related to the first, there is a significant gap in the law’s protection of ecosystem services. Third, common law doctrines, such as nuisance, are not sufficient to remedy harms to nature, especially when those harms manifest as ills to society at large rather than discrete landowners.

A. Discrete Character of Environmental and Natural Resources Law

Environmental laws can be roughly divided into two categories: traditional pollution control laws and natural resource laws. Traditional pollution control laws regulate discharge, emission, disposal, and cleanup of industrial sources of pollution. These laws segregate the environment into various media—air, water, or waste. Classic examples include the Clean Air Act, the Clean Water Act, the Resource Conservation and Recovery Act, the Toxics Substances Control Act, and the Comprehensive Environmental Response Compensation and Liability Act.
These federal statutes form the basis of state pollution control programs as well.\textsuperscript{154} Natural resource laws similarly divide nature into discrete elements, focusing on extractable resources such as wildlife, minerals, timber, oil, and gas.\textsuperscript{155} Natural resources law is largely “dominated by [a] ‘resource-ist,’ utilitarian approach rather than by a naturalist intrinsic value approach.”\textsuperscript{156} In that spirit, Merriam-Webster’s dictionary defines “natural resources” as “industrial materials and capacities (as mineral deposits and waterpower) supplied by nature.”\textsuperscript{157} Federal examples of natural resource laws include the Endangered Species Act,\textsuperscript{158} the National Forest Management Act,\textsuperscript{159} the Mineral Leasing Act,\textsuperscript{160} and the Federal Land Policy Management Act.\textsuperscript{161} At the state level, natural resource laws include game programs that regulate fishing and hunting,\textsuperscript{162} forestry statutes that regulate timber on state lands,\textsuperscript{163} and water codes that allocate water use and instream flows.\textsuperscript{164} Overlapping both pollution control regulation and natural resource laws are environmental statutes that are more ubiquitous in their application—namely the National Environmental Policy Act (NEPA)\textsuperscript{165} and their state equivalents.\textsuperscript{166} 

\begin{thebibliography}{9}
\bibitem{note154} Many of the major federal pollution control statutes—the Clean Air Act, Clean Water Act, and Resource Conservation and Recovery Act—are based on a concept of cooperative federalism. They allow the EPA to delegate the programs to state agencies, so long as certain minimum criteria are met. For a breakdown of the delegated programs operated in each state. \textit{See} ENVT. COUNCIL OF THE STATES, http://www.ecos.org (last visited Feb. 19, 2012).
\bibitem{note155} Lazarus, \textit{supra} note 142, at 631 (describing natural resource law as “historically concerned with the maintenance and orderly exploitation of basic natural resources such as water, fossil fuels, oil, natural gas, mineral deposits, and timber . . . .”).
\bibitem{note156} Fischman, \textit{supra} note 148, at 733; see also James Peck, Comment, \textit{Measuring Justice for Nature: Issues in Evaluating and Litigating Natural Resources Damages}, 14 J. LAND USE & ENVTL. LAW 275, 277 (1999) (“Traditional definitions of natural resources were limited to resources providing quantifiable economic products such as industrial minerals, energy sources, timber, and agricultural land.”).
\bibitem{note157} \textit{Natural Resource Definition}, MERRIAM-WEBSTER, http://www.merriam-webster.com/dictionary/natural%20resource (last visited Feb. 19, 2012); \textit{see also} BLACK’S LAW DICTIONARY 1027 (6th ed. 1991) (“Any material in its native state which when extracted has economic value . . . [and also] features which supply a human need and contribute to the health, welfare, and benefit of a community, and are essential to the well-being thereof . . . .”).
\bibitem{note162} See, e.g., Game and Fish, MNN. STAT. ANN. § 97A (West 2009 & Supp. 2010).
\bibitem{note163} See, e.g., Idaho Forestry Act, IDAHO CODE ANN. § 38-102 (2011).
\bibitem{note164} See, e.g., WASH. REV. CODE ANN. § 90.03.247 (West 2004).
\end{thebibliography}
Despite the breadth of environmental media and parts of nature addressed in these numerous federal and state statutes, no major laws focus on the protection of ecosystems as a whole.\textsuperscript{167} By and large, our environmental regulatory regime addresses discrete segments of nature on the overarching assumption that intense management and regulation of the parts will ultimately protect the whole: “We are accustomed to managing environmental and natural resource problems one-at-a-time and in isolation from each other, as if pollution control, water supply, fisheries management, and habitat conservation had nothing to do with each

\textsuperscript{166} In general terms, NEPA requires federal agencies to examine the impacts of major federal actions on the environment before deciding on a course of action. \textit{Id.} § 4332(2)(C). NEPA does not mandate substantive outcomes; rather it requires agencies to consider impacts of the propose action and viable alternatives. See, \textit{e.g.}, Dep’t of Transp. \textit{v.} Pub. Citizen, 541 U.S. 752, 756–57 (2004). Having done so, agencies are then free to choose among its proposed alternatives. See \textit{id.} at 757. Unlike NEPA, some state environmental protection statutes, or “little NEPAs,” do provide substantive mandates. See, \textit{e.g.}, MINN. STAT. ANN. § 116D.04, Subd. 6 (West Supp. 2010) (prohibiting the state agency to approve actions that significantly affect the environment if “there is a feasible and prudent alternative consistent with the reasonable requirements of the public health, safety, and welfare”); \textit{see also} CAL. PUB. RES. CODE § 21061.1 (West 2007) (defining feasible as “capable of being accomplished in a successful manner within a reasonable period of time, taking into account . . . environmental . . . factors.”); N.Y. ENVTL. CONSERV. LAW § 8-0109(1) (2005) (“Agencies shall . . . act and choose alternatives which, . . . to the maximum extent practicable, minimize or avoid adverse environmental effects . . . ”).

\textsuperscript{167} See, \textit{e.g.}, J.B. Ruhl, \textit{Biodiversity Conservation and the Ever-Expanding Web of Federal Laws Regulating Nonfederal Lands: Time for Something Completely Different?}, 66 U. COLO. L. REV. 555, 578 (“No single federal law can reasonably be portrayed as encompassing all the goals of biodiversity conservation and the authorities needed to carry them out.”); Salzman, \textit{supra} note 74, at 137 (“[I]t should come as no surprise that our laws do not explicitly protect ecosystem services.”); \textit{see also} Mary Jane Angelo & Mark T. Brown, \textit{Incorporating Emergy Synthesis into Environmental Law: An Integration of Ecology, Economics, and Law}, 37 ENVTL. L. 963, 967 (2007) (“The ecological shortcomings in current environmental statutes are rooted in the fact that most environmental statutes were enacted in the 1970s and 1980s, prior to many of the recent developments in the ecological sciences, and most of these statutes are media-based rather than ‘system’-based.”); Salzman et al., \textit{supra} note 81, at 309–10 (“In recommending that reduced ecological risk become a primary focus of EPA, its scientists and managers have revealed the single greatest failing of modern environmental law and its greatest challenge today—the inadequate protection of ecosystems and the services they provide.”).
other.”

This is true regardless of the distinctions drawn between pollution control laws and natural resource laws. Of course, there are some aspects of existing environmental and natural resource laws that do provide indirect opportunities for ecosystem protection. The most notable of these opportunities come from NEPA’s requirement that federal agencies broadly examine the indirect and cumulative impacts of their actions, the Endangered Species Act’s protection of habitat as a critical aspect of species recovery, and the Clean Water Act’s Section 404 regulation of wetlands. Each of these statutes, however, contains significant limitations in their ability to fill the gap of ecosystem protection.

For instance, NEPA appears to be a promising tool for ecosystem-level protection because it requires an expansive examination of impacts beyond a particular species, discrete element, or isolated media of nature. But NEPA is

---

168 Karkkainen, supra note 61, at 204; see also Annecoos Wiersema, A Train Without Tracks: Rethinking the Place of Law and Goals in Environmental and Natural Resources Law, 38 ENVTL. L. 1239, 1241 (2008) (noting the growing scholarship advocating “a holistic approach to environmental protection, moving away from a focus on separate media like air, water, and waste.”).

169 Salzman, supra note 74, at 136–37 (explaining that neither pollution control laws, conservation laws, nor resource management laws provide legal standards for conserving natural capital and ecosystem services).

170 For a comprehensive survey and analysis of federal laws with some biodiversity conservation potential, see Ruhl, supra note 167.


173 33 U.S.C. § 1344 (2006). There are also some provisions of federal land management statutes that would appear, to require a broader vision for ecosystem management by prohibiting degradation of certain resources. For example, the Federal Land Policy and Management Act prohibits “unnecessary or undue degradation” of the lands managed by the Bureau of Land Management. 43 U.S.C. § 1732(b) (2006). While promising in theory, such provisions have not historically had much effect: “The reality of natural resources law is that commodity users have overridden the good intentions and discretionary language of the Multiple-Use Sustained-Yield Act of 1960, the Federal Land Policy and Management Act, the National Forest Management Act, and similar statutes without breaking stride.” Oliver A. Houck, On the Law of Biodiversity and Ecosystem Management, 81 MINN. L. REV. 869, 882–83 (1997); see also Nagle & Ruhl, supra note 30, at 72 (citing Michael J. Bean & Melanie J. Rowland, The Evolution of National Wildlife Law 278 (3d ed. 1997)) (“[W]ildlife law is not a commanding force in federal land management. It is one of several management objectives on all major federal land classifications . . . . It is not the exclusive, or even the dominant goal on any lands but the national wildlife refuges.”).

174 In its simplicity, NEPA requires all federal agencies to consider the impacts of major actions having a potentially significant impact on the human environment. 42 U.S.C. § 4332 (2006); see also Robertson v. Methow Valley Citizens Council, 490 U.S. 332, 347–50 (1989). As part of that consideration, agencies must examine not only direct, but also
limited in notable ways. First, it applies only to major federal actions, which means that actions on nonfederal lands are outside NEPA’s reach unless they are otherwise connected to federal approval or funding. Second, even when it does apply, NEPA is procedural in nature; it requires agencies to consider impacts of their actions but it does not require agencies to choose the environmentally responsible alternative. Indeed, it is a common observation that NEPA prevents uninformed decisions, not unwise ones.

Unlike NEPA, the Endangered Species Act (ESA) does impose substantive mandates; once a species is listed as endangered or threatened, the ESA aggressively prohibits takings on both federal and nonfederal lands. Notably, its protections extend to habitat. The United States Fish and Wildlife Service (FWS) regulations, for example, prohibit “significant habitat modification or degradation where it actually kills or injures wildlife by significantly impairing essential behavioral patterns, including breeding, feeding, or sheltering.” And though the Supreme Court endorsed only a strict reading of this harm rule, the FWS has successfully used it to advance ecosystem protection goals by policing habitat destruction. Another important tool for habitat protection is the ESA’s indirect and cumulative impacts of its decision on a myriad of resources including “natural systems.” This includes evaluating all reasonably foreseeable impacts in the larger context of the project. The CEQ regulations implementing NEPA expressly provide that a project’s potential to impact “ecologically critical areas” informs whether the action is deemed “significant” within the meaning of the Act. 175 176


176 See, e.g., Dep’t of Transp. v. Pub. Citizen, 541 U.S. 752, 756–57 (2004) (quoting Robertson, 490 U.S. at 350) (“NEPA itself does not mandate particular results . . . . Rather NEPA, imposes only procedural requirements on federal agencies . . . .”); Lands Council v. Powell, 395 F.3d 1019, 1026 (9th Cir. 2004) (“NEPA imposes procedural requirements, but not substantive outcomes, on agency action.”); see also Ruhl, supra note 167, at 612–13 (discussing NEPA’s shortcomings as an effective biodiversity conservation tool, including its “dearth of substantive effect”).

177 Robertson, 490 U.S. at 351.


179 50 C.F.R. § 17.3(c)(3) (2010); see also Endangered and Threatened Wildlife and Plants; Definition of “Harm,” 64 Fed. Reg. 60,727, 60,727 (Nov. 8, 1999) (to be codified at 50 CFR Part 222) (NFMS final rule adopting a similar definition).

180 In Babbbitt v. Sweet Home Chapter of Communities for a Great Oregon, the United States Supreme Court upheld the rule as consistent with congressional intent but only to the extent that the harm resulting from habitat modification is foreseeable and the proximate cause of actual death or injury to identifiable individuals of the species. 515 U.S. 687, 711 (1995).

requirement that the FWS designate critical habitat for listed species. This tool, however, is underutilized and largely dependent on litigation.

Like NEPA, the tools available under the ESA can be used to advance some broader ecosystem conservation goals. In the end, however, the ESA’s focus has been on protecting species, not ecosystems. This makes the ESA a poor substitute for laws that are more specifically focused on ecosystem protection. Indeed,

This is especially the case when critical habitat designations are made for larger predators with more expansive ranges and habitat needs. See, e.g., Robert B. Keiter, Beyond the Boundary Line: Constructing a Law of Ecosystem Management, 65 U. COLO. L. REV. 293, 308 (1994) (describing the role that the ESA has played in constraining development on large blocks of land in order to ensure grizzly bear recovery in the Northern Rocky Mountains). But see Julie B. Bloch, Preserving Biological Diversity in the United States: The Case for Moving to an Ecosystems Approach to Protect the Nation’s Biological Wealth, 10 PACE ENVTL. L. REV. 175, 202 (1992) (noting one of the shortcomings of the ESA is the fact that it “protects high profile species that do not usually play an important role in overall biodiversity”), Bradley C. Karkkainen, Biodiversity and Land, 83 CORNELL L. REV. 1, 20 (1997) (A “well-understood limitation of the ESA is that it has generally afforded greater protection to high-profile ‘charismatic’ species, especially large vertebrates, at the expense of lesser-known or less popular species.”).

In reality, the critical habitat designation requirement has failed to protect approximately two-thirds of listed species. In particular, in August 2007, the U.S. Fish & Wildlife Service reported that critical habitat had been designated for only 492 of the 1351 listed endangered and threatened species. U.S. FISH & WILDLIFE SERVICE, CRITICAL HABITAT: WHAT IS IT? 2 (2007), available at http://library.fws.gov/Pubs9/critical_habitat07.pdf; see also Patrick Parenteau, An Empirical Assessment of the Impact of Critical Habitat Litigation on the Administration of the Endangered Species Act (Vt. Law School Faculty Papers, Paper No. 1, 2005), available at http://lsr.nellco.org/cgi/viewcontent.cgi?article=1000&context=vermontlaw_fp; Matthew D. Crawford, Note, The Timing of Challenges to Compel Critical Habitat Designation Under the Endangered Species Act: Should Courts Toll the General Federal Statute of Limitations?, 36 B.C. ENVTL. AFF. L. REV. 497 (2009). Moreover, most of those designations have come in direct response to litigation and court-ordered timelines. In fact, between 1990 and 2005, 350 out of 357 critical habitats designated by FWS were the result of litigation. Parenteau, supra at 2 n.7. “Each critical habitat designation made since 1997 has resulted from a court order or a settlement agreement . . . .” U.S. GEN. ACCOUNTING OFFICE, GAO-03-803, ENDANGERED SPECIES: FISH AND WILDLIFE SERVICE USES BEST AVAILABLE SCIENCE TO MAKE LISTING DECISIONS, BUT ADDITIONAL GUIDANCE NEEDED FOR CRITICAL HABITAT DESIGNATIONS 34 (2003), available at http://www.gao.gov/new.items/d03803.pdf. So while the number of listed species afforded critical habitat designations has climbed from roughly nine percent in 1999 to thirty-six percent in 2007, the Fish & Wildlife Service acknowledges that this increase is due, in large part, to litigation.

Houck, supra note 173, at 870 (“One of the more rational conclusions to emerge from American’s experience with the Endangered Species Act is that we need to manage ecosystems and protect biological diversity on a scale larger than individual species on the brink of doom.”); Keiter, supra note 182, at 307–09 (explaining the limitations of the ESA and concluding that “although the ESA breaches conventional boundary lines and protects enumerated species against extinction, it cannot be regarded as a general ecosystem
notwithstanding the ESA’s promise “to provide a means whereby ecosystems upon which endangered and threatened species depend may be conserved,” no other provision of the ESA contains action measures specific to ecosystem protection. Moreover, the ESA’s ability to serve double duty as an ecosystem protection statute is limited by the fundamental reality that species only trigger the protections of the Act when they are endangered or threatened with extinction over a significant portion of their range. Benefits of the ESA to ecosystems therefore come from a reactive posture; they only become effective after habitats and their broader ecosystem functions have already been curtailed or adversely modified.

It is worth examining one last example of an environmental statute that has potential to protect some parts of ecosystems. Section 404 of the Clean Water Act (CWA) authorizes the Army Corps of Engineers to issue permits for the “discharge of dredged or fill material into the navigable waters at specified disposal sites.” Although this section does not actually mention wetlands, its permitting regulations construe the term “navigable waters” broadly to include wetlands. Over time, Section 404 has become synonymous with wetlands regulation—it prohibits wetlands from being drained without some balancing of the expected benefits of the permitted activity and its potential for environmental harm.

Although the CWA plays an important role in protecting at least one ecosystem service, wetlands are but one piece of an ecosystem, and protecting wetlands cannot substitute for broader ecosystem protection. As useful as the CWA is, it is not (and does purport to be) a complete package for ensuring the longevity of ecosystem services as a whole.
B. Gaps in Protection of Ecosystem Services

Assorted provisions—isolated within disconnected statutes, administered by different agencies, and providing only indirect tools—are hardly an appropriate substitute for the broad-based and comprehensive legal regime that is necessary to effectively protect ecosystems. Indeed, most scholars agree that the complex web of environmental statutes has been largely unsuccessful at providing adequate protections to ecosystems on a holistic basis.¹⁹² Fifteen years ago, Professor J.B. Ruhl observed that “[n]o single federal law purports to encompass all that is meant by biodiversity conservation; rather, a handful of different statutes addresses particular facets of biological resource protection on nonfederal lands. Gluing those laws together without any clear, unifying principles has not created an effective, flexible system of biodiversity conservation.”¹⁹³ More recently, Professor Annecoos Wiersema reiterated, “most commentators now seem willing to agree with two propositions. First, environmental law must be responsive to ecological insights about the complexity of natural systems. Second, traditional approaches to environmental law appear insufficiently responsive to science, and further, insufficiently flexible even to develop responsiveness to science.”¹⁹⁴

This conclusion has been similarly expressed by federal agencies, lamenting the inability of the current regulatory web to adequately protect important
daily loads” (TMDLs) for impaired waterways has the potential to elicit a more comprehensive assessment of all the sources of pollution that give rise to unacceptable levels of pollutants. TMDLs, therefore, force evaluation of nonpoint as well as point sources of pollution and could be used to understand pollution issues on a watershed basis. For a discussion of TMDLs as a tool for watershed-based pollution control, see Michael M. Wenig, How “Total” Are “Total Maximum Daily Loads”?—Legal Issues Regarding the Scope of Watershed-Based Pollution Control Under the Clean Water Act, 12 TUL. ENVTL. L.J. 87, 106 (1998).

¹⁹² See Angelo & Brown, supra note 167, at 967 (“Environmental law’s current integration of ecological science is overly simplistic, ad hoc, and outdated.”); Fischman, supra note 148, at 741 (“Over the past thirty-five years, both environmental law and natural resources law have struggled to broaden their scopes to encompass ecological concerns.”); Karkkaninen, supra note 147, at 77 (noting that the current approach to environmental protection has not been “effective in the more complex and integrative tasks of protecting ecosystems”); Karkkainen, supra note 61, at 197 (“[W]e have constructed an architecture of laws and management systems that are poorly matched to the challenge of managing ecosystems as complex dynamic systems.”); Ruhl, supra note 167, at 561–62 (arguing that a more effective federal policy for conserving biodiversity “would come from melding the disorganized system of federal biodiversity conservation regulation into a single law designed principally to promote biodiversity conservation”); Wiersema, supra note 168, at 1249 (“[L]aws that focus on one medium, such as air, water, land, or individual species, will not adequately take account of the multiple connections.”).

¹⁹³ Ruhl, supra note 167, at 565.
¹⁹⁴ Wiersema, supra note 168, at 1245.
ecosystem functions. In 1994, the EPA exclaimed the need for a more place-driven approach to environmental protection in a planning report named the Edgewater Consensus:

To date, the EPA has accomplished a great deal, addressing many major sources of pollution to the nation’s air, water and land. Yet, even as we resolve the more obvious problems, scientists discover other environmental stresses that threaten our ecological resources and general well-being.

. . . Although many federal, state, and local regulations address these problems, past efforts have been as fragmented as our authorizing statutes. Because EPA has concentrated on issuing permits, establishing pollutant limits, and setting national standards, the Agency has not paid enough attention to the overall environmental health of specific ecosystems. In short, EPA has been “program-driven” rather than “place-driven.”

Recently, we have realized that, even if we had perfect compliance with all our authorities, we could not assure the reversal of disturbing environmental trends.195

A report published by the United States General Accounting Office around the same time similarly concluded that “[e]ven though many laws have been enacted to protect individual natural resources—air, water, soils, plants, and animals, including forests, rangelands, threatened and endangered species,

---

195 Ecosystem Prot. Workgroup, U.S. Envtl. Prot. Agency, Toward a Place-Driven Approach: The Edgewater Consensus on an EPA Strategy for Ecosystem Protection (Draft 1994), in THE LAW OF BIODIVERSITY AND ECOSYSTEM MANAGEMENT, supra note 30, at 363, 363; see also, U.S. ENVTL. PROT. AGENCY, 200-B-94-002, THE NEW GENERATION OF ENVIRONMENTAL PROTECTION: A SUMMARY OF EPA’S FIVE-YEAR STRATEGIC PLAN 2 (1994) (“In the past, the Agency’s division into air, water, and land programs led EPA to overlook both the cross-media effects of some pollution problems and the potential for new kinds of cross-media programs.”); SCI. ADVISORY BD., U.S. ENVTL. PROT. AGENCY, SAB-EC-90-021, REDUCING RISK: SETTING PRIORITIES AND STRATEGIES FOR ENVIRONMENTAL PROTECTION 6 (1990) (recommending the EPA reset its priorities to “attach as much importance to reducing ecological risk as it does to reducing human health risk”); Bloch, supra note 182, at 200 (“Many U.S. laws mandate conservation of some aspect of biological diversity, from the broad mandate of the National Forest Management Act to the habitat conservation approach of the Endangered Species Act . . . . Nevertheless, these programs and statutes do not form a coherent comprehensive framework for assessing or ensuring progress toward a common goal . . . .”’ (quoting COUNCIL ON ENVTL. QUALITY, U.S. NATIONAL REPORT PREPARED FOR SUBMISSION TO THE U.N. CONFERENCE ON THE ENVIRONMENT AND DEVELOPMENT 69 (Draft 1991) (on file with CEQ))); Salzman et al., supra note 81, at 309 (“The top managers and scientific advisors in the [EPA] have consistently declared that maintenance of productive natural systems demands more attention and should, in fact, become one of the agency’s highest priorities.”).
wetlands, and wilderness areas—ecological conditions on many federal lands have declined.” 196

In conjunction with these and other recognitions of statutory shortcomings, many federal agencies, including the four major public land management agencies, prepared statements and guidelines in the mid-1990s expressing their commitment to more broad-based ecosystem management. 197 For example, the Department of Interior proposed a national biological survey that would inventory the nation’s natural resources. The purpose was to “get, at least, a first cut of how we relate to these ecosystems.” 198 In December 1993, the Bureau of Land Management (BLM) issued a report stating its intent to incorporate ecosystem management principles into its oversight of public lands. 199 Among other things, the BLM envisioned “an interdisciplinary approach to land management in which program advocacy will yield to ecosystem advocacy.” 200 The FWS issued a report in March 1994 that described how to apply ecosystem management to fish and wildlife conservation. 201 And, finally, the Forest Service released a report of its own in the same year that pledged to follow four fundamental principles in its approach to ecosystem management: (1) the use of an ecological approach to multiple-use management; (2) application of the best scientific knowledge and technologies to decision-making; (3) encouragement of partnerships with state agencies and private landholders; and (4) the promotion of grass-roots participation in the planning process. 202 The Forest Service even announced its intention to revise its forest planning regulations to incorporate these four ecosystem management principles. 203

Encouraging as it is that the major federal agencies understood the need for an ecosystem focus to land management, did these recognitions of shortcomings and statements of good intentions manifest genuine ecosystem management requirements? In other words, was Professor Oliver Houck correct in 1997 when he exclaimed, “tough odds call for precise law” and cautioned that amorphous goals like ecosystem management would require teeth to be successful? 204

Perhaps the best evidence of progress, or lack thereof, lies in the scientific assessment of continuing ecosystem, species, and habitat decline over the past

196 U.S. GEN. ACCOUNTING OFFICE, ECOSYSTEM MANAGEMENT: ADDITIONAL ACTIONS NEEDED TO ADEQUATELY TEST A PROMISING APPROACH 3 (1994).
198 Id., supra note 197, at 45.
199 Id.
200 Id.
201 Id.
202 Id. at 70.
203 Id.
204 Houck, supra note 173, at 883.
decade. In November 2009, the International Union for Conservation of Nature (IUCN) published its 2008 Red List of Threatened Species, reporting that 17,291 species out of the 47,677 assessed species are threatened with extinction. In particular, “the results reveal 21 percent of all known mammals, 30 percent of all known amphibians, 12 percent of all known birds, and 28 percent of reptiles, 37 percent of freshwater fishes, 70 percent of plants, 35 percent of invertebrates assessed so far are under threat.” Jane Smart, Director of IUCN’s Biodiversity Conservation Group remarked that “[t]he scientific evidence of a serious extinction crisis is mounting.” While the ICUN Report analyzes the global patterns of species and habitat loss, the United States features prominently among nations with top-twenty numbers of threatened species for amphibians, birds, mammals, and conifers. This apparent failure to preserve biodiversity is a useful benchmark of the law’s failure to ensure ecosystem health more generally.

The failed effort to utilize existing environmental statutes as hard-nosed tools for protecting ecosystems is also evidenced by the wealth of scholarship urging the need for new approaches. Professors Mary Jane Angelo and Mark T. Brown conclude that “[a]lthough many existing environmental laws pay lip service to ecological science, they do not incorporate scientific understanding of the ecological world in any meaningful way or are not implemented in a manner that significantly incorporates ecological science.” To resolve this issue, Professor Bradley Karkkainen argues for expanded embrace and coordination of collaborative ecosystem governance. Others have similarly argued that new models of governance are necessary to adapt to the ever-changing state of...
ecological understanding. Much of this work has focused on developing procedural laws capable of capturing the flexibility that is fundamental to adaptive management. Professor Wiersema also recognizes the disconnect between fragmented environmental laws and whole ecosystems, but she advocates staying focused on substantive laws as the primary mechanism for achieving needed change: “long-term environmental protection can only be achieved by these models if we can be sure that all of the values that are at stake in environmental protection will be adequately represented by the procedural mechanisms that these institutions envision.” While these scholars, and others, have primarily argued for new systems of governance that are adaptive to evolving information concerning ecosystems, there is a body of scholarship more specifically devoted to the protection of ecosystem services through common law doctrines.

From this wealth of scholarship and our examination of substantive environmental laws, we can properly draw two conclusions. First, regulating parts of nature instead of nature as a whole is well understood to be an incomplete approach to environmental and natural resource regulation. Second, despite the fact that scientists, federal agencies, lawmakers, and scholars have long been aware of the need to manage ecosystems on a more holistic basis, there is still a gap in regulation, sparking a need for substantive laws expressly aimed at protecting ecosystems and the important services that they provide.

C. Limitations of the Common Law

The increased mainstream attention that ecological services have been receiving from ecologists, economists, and government agencies since the publication of Nature’s Services in 1997 has led to a corresponding explosion of legal scholarship exploring how to synthesize environmental laws with ecological understanding. For the most part, this scholarship has grappled with two main issues. First, the ability of existing legal doctrines to utilize the economic value associated with ecological services. Second, the valuation of ecological services and the creation of markets to capture that value.

212 Wiersema, supra note 168, at 1242–44 & nn.7–22 (surveying scholarship that urges the adoption of new governance models rooted in procedural law).
213 Id. For a detailed discussion and critique of adaptive management, see Holly Doremus, Adaptive Management as an Information Problem, 89 N.C. L. REV. 1455 (2011).
214 Wiersema, supra note 168, at 1244.
215 See infra Part II.C.
216 In the last ten years, over 450 law review articles referencing ecological services have been published, 75 with “ecological services” or “ecosystem services” appearing somewhere in the title or summary.
217 See, e.g., infra notes 219–230 and accompanying text.
218 See, e.g., James Salzman, Valuing Ecosystem Services, 24 ECOLOGY L.Q. 887 (1997) (approaching ecological services from the perspective of valuation issues); Salzman, supra note 74, at 133–34 (tackling the issue of how to create ecosystem services
Among the most prominent and prolific writers in this field is Professor J.B. Ruhl. Most recently, Professor Ruhl’s work has focused on the common law nuisance doctrine and its ability to provide a remedy for economic injuries to ecological services. Ruhl argues that the birth of ecological economics and understanding of ecology create a legal system that is ripe for revisiting nuisance doctrine as a tool for remedying harms to ecological services. He explains that using the nuisance doctrine to address these harms was made possible by the Supreme Court’s decision *Lucas v. South Carolina Coastal Council*. In that case, Justice Scalia carved out an exception for regulatory takings when the challenged restrictions merely reflect those already imposed by common law nuisance or other “background principles” of property law. According to Ruhl, this recognition provides the opening for nuisance law to address new knowledge regarding the value of ecological services.

Perhaps because its roots lie in capitalizing on *Lucas*, Ruhl’s argument for pressing the nuisance doctrine into service is limited to cases where ecological injuries contain identifiable economic values. His is a self-described “instrumentalist” approach based on “welfare economics.” He recognizes that there are moral, ethical, or scientific arguments on behalf of ecological integrity, markets such that the valuation of these services can play a role in environmental laws); Barton H. Thompson, Jr., *Markets for Nature*, 25 WM. & MARY ENVTL. L. & POL’Y REV. 261 (2000) (discussing regulatory markets, public goods markets, and ecosystem services markets).


See *Lucas*, 505 U.S. at 1029.

*Ruhl, Making Nuisance Ecological*, supra note 219, at 758–60; Ruhl, *supra* note 221, at 536–40. Though Ruhl contends that his arguments are not motivated by takings avoidance, his arguments are nonetheless premised on the openings left by the Supreme Court in *Lucas*, and his pragmatic approach to the use of nuisance law for addressing harms to ecological services suggests that takings avoidance must play a role in the development of his legal theories. Cf. James L. Huffman, *Beware of Greens in Praise of the Common Law*, 58 CASE W. RES. L. REV. 813, 813–14 (2008) (criticizing Ruhl and other scholars for embracing the common law after *Lucas* as a takings avoidance strategy).

but decides to take a pragmatic approach by trying to protect ecosystem services within the playing fields already established—namely market-valuation and common law doctrines that recognize economic injury:

I am not suggesting that nuisance law take on the whole of ecosystem management law . . . . Rather, what I have in mind looks and feels like a rather conventional nuisance action, the only novel feature being that the plaintiff is linking damage to ecological resources on defendant’s property with injury to use and enjoyment of plaintiff’s property.227

Ruhl also recognizes the limits of this approach and its ability to fully protect ecological services. With respect to private nuisance doctrine, he acknowledges that

it cannot be that all losses of ecosystem services have a remedy in nuisance. Indeed, what I have outlined as an ecosystem service nuisance is intended to fit within the conventional doctrine of private nuisance, not to morph it into a general ecological protection regime. In the absence of a plaintiff whose use and enjoyment of property is substantially injured as a result of another landowner’s degradation of natural capital, no ecosystem service nuisance has been committed.228

Ruhl, in other words, offers nuisance doctrine as a tool for remedying only the most traditional and narrow category of ecological harms.229

Other scholars have more aggressively argued that common law doctrines can remedy broader categories of ecological nuisances. Professor John Copeland Nagle, for example, would take Ruhl’s argument a bit further.230 Using the destruction of wetlands as an example of ecological nuisances, Nagle argues that we should move beyond the traditional role of nuisance law in protecting economic interest; he argues that environmental ethics should inform which actions give rise to ecological nuisance claims.231

By thinking beyond economics, Nagle gets closer to recognizing the breadth and depth of the problem at hand. Ethics and broader considerations of what constitutes good land use are part of any environmental regulatory decision, whether so acknowledged or not.232 In other words, we can hardly decide on a satisfactory point of environmental regulation without making a value judgment

227 Id. at 777.
228 Id. at 774–75.
229 See id.; see also, e.g., Anderson, supra note 92, at 551; Goldstein, supra note 12, at 347; Wilgus, supra note 11, at 99.
231 Id. at 802–11.
232 See FREYFOGLE, supra note 16, at 144–46.
about acceptable tradeoffs between interests of the private and the public—for example, private economic interests and public health, or private property rights and public resources. Given the importance of land health and ecological services to human welfare, economic valuation is not an acceptable proxy for the worthiness of protecting us from ourselves.233

If Nagle picks up where Ruhl left off, this Article picks up where Nagle left off. While both Ruhl and Nagle looked to common law as an avenue for remedying harms to ecological services, both agreed that the common law, though informative and the only tool that appears readily available, does not offer a broad or comprehensive enough solution. In his work on illustrating ecological nuisance through wetlands regulation, Nagle commented, “individualized assessment provided by nuisance law is not a panacea for wetlands regulation. Ethical norms are more readily incorporated into statutory provisions that are crafted with particular goals in mind.”234 Ruhl made a similar observation when he expressed skepticism that the common law could accommodate the underlying ethical and moral considerations in an ecological context.235 In the end, this limitation of the common law appears to have driven his preference for tempting the law’s “instrumentalist core” with economics of natural capital.236

Other scholars have more strongly criticized the common law’s ability to serve as the foundation for ecosystem protection. Professor James Huffman objects to the creation of ecological nuisances through common law because it would allow courts, and not the legislature, to set environmental policy.237 Huffman is also critical of arguments advanced by Ruhl and others because he sees those arguments merely as attempts to avoid regulatory takings through common law doctrines.238

While Huffman’s suggestion that the common law has no useful role in shaping a new law of ecosystem protection is likely overstated, he aptly observes that the challenge before us requires a broader, coordinated approach. The common law provides the foundation for modern environmental pollution control laws and is “profoundly adaptive.”239 But it is also a slowly churning machine, processing one individual dispute at a time, eventually turning out a by-product of

233 See ACKERMAN & HEINZERLING, supra note 46, 8–9.
234 Nagle, supra note 31, at 811.
235 Ruhl, Making Nuisance Ecological, supra note 219 at 784–85.
236 Id. at 785.
237 Huffman, supra note 224, at 813–14.
238 Id. at 813 (“But for the most part the fledgling environmental case for revival of common law remedies is rooted in a belief that a reinvigorated common law will further weaken constitutional protections of property rights that might otherwise stand in the way of command and control regulation.”).
239 Ruhl, Ecosystem Services, supra note 219, at 8 (“Almost a century ago the U.S. Supreme Court decision in Georgia v. Tennessee Copper Co., 206 U.S. 230 (1907), suggested that the common law could play an important and innovative role in pollution control.”).
policy that is many cases and years in the making. Because of this, the common law is widely recognized as a “grossly inadequate” tool for providing a broad and effective policy of pollution control.240

Nearly three decades ago, Professor Richard Lazarus wrote of another common law doctrine—the public trust doctrine—and criticized its use as impeding the necessary evolution of natural resources law.241 Similar to Ruhl’s invocation of the nuisance doctrine to protect ecosystem services, Lazarus explained how the public trust doctrine was invoked to impose limits on the privatization of natural resources.242 Though he agreed that limits were necessary, Lazarus criticized the public trust doctrine’s application as a legal fiction, arguing that natural resource laws needed to reflect modern ecological concerns and that clinging to old doctrines simply resulted in “tortured constructions of the present rather than repudiations of the doctrine’s past.”243 In addition, Lazarus cautioned against an overreliance on the public trust doctrine to address natural resource concerns because it was a product of judge-made law.244 In that way, the doctrine depended on a judiciary with a proenvironment bias to achieve sought-after results of natural resource protection. Though Lazarus spoke of a different time, both in terms of ecological understanding and regulatory takings law, some of Lazarus’s criticisms hold true today as nuisance law is advanced to protect ecosystem services and to set limits on private landowners’ misuse of such services.

Though the common law is a flexible tool that achieves legitimacy through its case-by-case embodiment of evolving public values,245 such a slow-moving, piecemeal approach is not responsive enough to the pace at which landscapes are altered and ecological services are destroyed. This is not to say that common law doctrines are inconsistent with these proposals, or that some of the proposals have not been embraced historically by the common law. But there are many layers of change that would need to take place in current thinking before the common law

240 Id. at 5 (citing ROBERT PERCIVAL ET AL., ENVIRONMENTAL REGULATION 72 (4th ed. 2003); Boomer v. Atlantic Cement Co., 257 N.E.2d 870, 871 (N.Y. 1970) (“A court should not try to do this on its own as a by-product of private litigation and it seems manifest that the judicial establishment is neither equipped in the limited nature of any judgment it can pronounce nor prepared to lay down and implement an effective policy for the elimination of air pollution. This is an area beyond the circumference of one private lawsuit.”)).

241 See Lazarus, supra note 142, at 631–33.

242 Id. at 632–33.

243 Id. at 711.

244 Id. at 712–13.

245 See Ruhl, Ecosystem Services, supra note 219, at 6 (“Most comprehensive treatments of the evolution of environmental law begin with the common law as the first meaningful stage of development. . . . The common law [] provided much-needed legitimacy to the public law agenda for pollution control.”); see also id. at 7 (arguing that ecosystem management law has gained little traction because “ecosystem management legislation tried to leapfrog its common law formative stage”).
could uniformly embrace new ecological understanding and heal the schism between responsible land use on public and private land. Doing so in a timely fashion asks too much of the common law.

Like Lazarus’s criticism of reliance on the public trust doctrine as a panacea, it is dangerous to put too much stock in the ability of nuisance law to reshape remedies for the misuse of private lands that harbor valuable ecosystem services. Such an evolution is not only slow, but it is also dependent on a judiciary being willing to recognize new ecological concerns, and it does not necessarily strike at the heart of some of the more nettlesome issues underlying ecosystem protection.

Given the limitations of the common law, a more explicit and broad-based approach is needed in order to remedy harms to ecosystems. A rethinking and expansion of natural resource damages law is one such approach.

III. THE PROMISE OF A REMEDY IN NATURAL RESOURCE DAMAGES LAW

Before discussing how an expansion of natural resource damages law might be reformed, a recap of the major arguments made so far might be useful at this point. First, we ought not misuse nature, though where we draw the line between use and misuse has yet to be determined.246 Second, our understanding of nature has matured through the study of ecology, and we now have a greater appreciation for the valuable services that ecosystems provide to society.247 Third, we need to broaden our notions of public ownership and assert control over parts and processes of nature that were once deemed worthless but are now understood as valuable.248 Finally, the multitude of environmental laws currently in effect do not protect nature as an interconnected whole.249 It is time to put behind us the language of nature as separable elements and talk about the disruption of nature’s vital functional processes.

Putting all these pieces together, the key claim is that we ought to consider a new, broader, more ecologically and ethically informed understanding of “natural resource damages.” More specifically, if we combine a concept of natural resources that recognizes ecosystem services; the foundation of natural resource damages law as it exists today; the normative claim that we ought not misuse nature; and a more complete acceptance of public ownership of nature, then we can begin to conceive of a new natural resource damages law that provides a remedy for the misuse of nature.

This Part explores the promise and limitations of expanding natural resource damages law. Part IV then sketches the ways in which we might craft causes of action, provide remedies, and empower citizens through a reformed natural resource damages law.

246 See supra Part I.A.
247 See supra Part I.B.1.
248 See supra Part I.B.2.
249 See supra Part II.B–C.
resource damages framework. Enumerating the challenges ahead will set the stage for a longer and more detailed discussion in the future.

A. The Promise of Natural Resource Damages Law

Natural resource damages law is a promising starting point for developing a framework for remediating misuse of nature. Key features of its framework include a broad definition of natural resources, the use of nontraditional economic methods, and treatment of natural resources as part of the public trust.

By way of background, federal statutory authority for natural resource damages originated in the 1977 amendments to the Clean Water Act. Today, authority is contained primarily in five federal statutes, including section 107 of the Comprehensive Environmental Response, Compensation and Liability Act (CERLCA); section 311(f)(4) of the Clean Water Act (CWA); section 1002 of the Oil Pollution Act of 1990 (OPA); the Marine Protection Research and Sanctuary Act; and the National Park System Resource Protection Act.

Technically, the natural resource damages concept was first used in the Trans-Alaska Pipeline Authorization Act of 1973 and was also addressed in the Deepwater Port Act of 1974. The Clean Water Act, however, brought the concept into significant light by applying it beyond particular, designated resources. Manus, supra note 13, at 424.

Of these federal statutes, CERLCA is perhaps the most well known and largest source of natural resource damage recovery actions. The attributes of CERLCA’s natural resource damages provisions will, therefore, serve as the primary stepping stone for discussion of natural resource damages potential and limitations here. Many states have similarly enacted state superfund statutes that include provisions for recovery of natural resource damages. E.g., Lloyd W. Landreth & Kevin M. Ward, Natural Resource Damages: Recovery Under State Law Compared with Federal Laws, 20 ENVTL. L. REP. 10, 134 (1990) (reviewing existing state laws establishing liability for natural resource damages). For example, the New Jersey Spill Act authorizes natural resource damage recovery for discharges of hazardous substances. See N.J. STAT. ANN. § 58:10-23.11(g)(c) (West 2006). A detailed analysis of state natural resource damages laws and their potential to inform a new framework for remediating the misuse of nature is a separate undertaking. This Article examines the promise and limits of natural resource damages as conceived in key federal statutes.

42 U.S.C. § 9607 (2006) (providing that a “trustee” may recover for injury to, destruction, or loss of natural resources, including reasonable cost of assessment, resulting from the release of hazardous substances).

33 U.S.C. § 1321(f)(4) (2006) (providing that “costs of removal of oil or a hazardous substance” recoverable under the statute include any cost or expenses incurred by federal or state government in the restoration or replacement of natural resources damaged or destroyed “as a result of a discharge of oil or a hazardous substance”).

33 U.S.C. § 2702 (2006) (providing recovery of “removal costs and damages” that result from discharges of oil into navigable waters or the adjoining shoreline, including recovery by a trustee for damages for injury to natural resources). OPA was enacted in response to the Exxon Valdez oil spill. See William H. Rodgers, Jr., et al., The Exxon
In general, these statutes create liability for injury to, or destruction or loss of, natural resources.\(^{257}\) CERLCA and OPA define natural resources to include “land, fish, wildlife, biota, air, water, ground water, drinking water supplies, and other such resources” that are managed or otherwise controlled by the United States.\(^{258}\) “Injury” is also broadly defined to include nearly any adverse impact on a resource.\(^{259}\)

CERLCA’s broad definition of natural resources reflects a willingness to move beyond the traditional treatment of natural resources as discrete segments of nature and recognize that ecosystems and their services are part of natural resources. To that end, the Department of Interior’s (DOI) damage assessment regulations advise trustees to quantify injuries to natural resources as well as “lost services.”\(^{260}\) Similarly, in the event that trustees must consider acquisition of resource equivalents in lieu of restoration, the DOI requires trustees to consider the “services” provided by the resources. Trustees are instructed not to treat services as commodities that can be restored independently of the resource:

[DOI] does not believe that Congress intended to allow trustee agencies to simply restore the abstract services provided by a resource, which could conceivably be done through an artificial mechanism. For example, nothing in . . . CERLCA suggests that replacement of a spring with a water pipeline would constitute ‘restoration, rehabilitation, replacement, and/or acquisition of equivalent resources.’\(^{261}\)

The National Oceanic and Atmospheric Administration’s (NOAA) regulations similarly recognize the need to rehabilitate ecological services. Those regulations provide that, where the injury has resulted in the loss of ecological services,


\(^{256}\) 16 U.S.C. § 19jj-1 (2006) (“[A]ny person who destroys, causes the loss of, or injures any park system resource is liable to the United States for response costs and damages resulting from such destruction, loss, or injury.”).


\(^{259}\) 43 C.F.R. § 11.14(v) (2010) (DOI regulations defining injury as “a measurable adverse change, either long- or short-term, in the chemical or physical quality or the viability of a natural resource . . .”); 15 C.F.R. § 990.30 (2011) (NOAA regulations providing a similarly broad definition).

\(^{260}\) 43 C.F.R. § 11.83(c)(1) (“The compensable value can include the economic value of lost services provided by the injured resources, including both public use and nonuse values such as existence and bequest values.”).

trustees must consider using Habitat Equivalency Analysis and similar “resource to resource” and “service to service” approaches to restoration. Because natural resource damages law already takes a more holistic view of natural resources, as opposed to considering them as discrete and separable elements of nature, it could provide the foundation that is needed to remedy misuses of nature more generally.

The second key feature of natural resource damages is the willingness of Congress and courts to accept nontraditional methods of valuing natural resource injuries. Instead of holding natural resource damages captive to traditional market valuations, CERLCA requires damage assessment regulations to “identify the best available procedures to determine such damages, including both direct and indirect injury, destruction, and loss. . . .” CERCLA further provides that the relevant factors for assessing natural resource damages will include “replacement value, use value, and [the] ability of the ecosystem to recover. . . .”

In Ohio v. U.S. Department of Interior, the leading case interpreting natural resource damage valuation under CERCLA, the D.C. Circuit held that “Congress intended the damage assessment regulations to capture fully all aspects of loss.” This means that damages are measured, at a minimum, by the cost of restoring the damaged natural resources to their preinjured condition. In some cases, where restoration is not feasible, replacing the injured resource becomes the focal point of the damages assessment.

Even if full restoration is feasible, the public may suffer a loss of enjoyment of the resource while it is injured. In these situations, damages may also include the lost use and nonuse values. Lost use values are measured by the benefits derived from current or expected future uses of the resource by identifiable persons; “use values are the more market-translatable values of a natural resource, encompassing both consumptive uses, such as hunting and fishing, in which

---

262 Habitat Equivalency Analysis is a method for assessing resource injuries. Itzchak E. Kornfeld, Of Dead Pelicans, Turtles, and Marshes: Natural Resource Damages in the Wake of the BP Deepwater Horizon Spill, 38 B.C. ENVTL. AFF. L. REV. 317, 330 (2011) (“The principal concept underlying [Habitat Equivalency Analysis] is that the public can be compensated for past losses of habitat resources through habitat replacement projects providing additional resources of the same type.”).

263 15 C.F.R. § 990.53(d).

264 42 U.S.C. § 9651(c)(2)(B) (2006). Indirect injuries include those that result from the implementation of remediation, such as the destruction of a wetland through the dredging of contaminated sediment. See 43 C.F.R. § 11.15(a)(3)(ii).

265 42 U.S.C. § 9651(c)(2)(B); see also 43 C.F.R. § 11.80(b) (DOI regulations setting forth two components of natural resource damage claims, including (a) a claim for monies required to restore, replace, or acquire an equivalent of the injured natural resource; and (b) a claim for the lost compensable value of the resource).

266 880 F.2d 432 (D.C. Cir. 1989).

267 Id. at 463 (emphasis added).

268 Id. at 463; see id. at 454 n.34, 464, 476–78; see also id. at 462–63 (“From the bald eagle to the blue whale and the snail darter, natural resources have values that are not fully captured by the market system.”).
resources are harvested, and nonconsumptive uses, such as hiking and bird watching, in which the activity does not reduce the stock or resources.269 Nonuse, or passive use, values are benefits that people derive from knowing that certain resources exist and could be used in the future.270

It is the inclusion of nonuse values in the damage assessment that underscores the desire to ensure the full measure of harm is remedied through natural resource damages law. For example, in Utah v. Kennecott Corporation,271 the district court rejected the state trustee’s damage assessment for groundwater contamination because it was based solely on the market value of the volume of water lost to the public.272 The court concluded that the trustee “adopted a too narrow interpretation of use value by equating such with market value only. . . . [I]t failed to assess the non-consumptive use values of the aquifer, i.e., option and existence values.”273

Other courts have similarly recognized the need to include nonuse values into the natural resource damage assessment equation and upheld nontraditional methods of valuation. In Ohio v. U.S. Department of the Interior,274 the D.C. Circuit explained that “[o]ption and existence values may represent ‘passive’ use, but they nonetheless reflect utility derived by humans from a resource, and thus, prima facie, ought to be included in a damage assessment.”275 Contingent valuation,276 a particularly controversial method for assessing nonuse values, has even been upheld as a valid method for valuing natural resource damages.277 In addition, at least one court has upheld Habitat Equivalency Analysis278 as an appropriate method for valuing ecosystem damages.279

269 Manus, supra note 13, at 447.
272 Id. at 571.
273 Id.
274 880 F.2d 432.
275 Id. at 464; see also Idaho v. S. Refrigerated Transp., No. 88-1279, 1991 WL 22479, at *18 (D. Idaho Jan. 24, 1991) (holding that commercial, existence, and recreation values all “exist and would be appropriate items of damage if proved at trial”).
276 For an explanation of the contingent valuation methodology and why it is controversial, see Miriam Montesinos, Comment, It May Be Silly, But It’s an Answer: The Need to Accept Contingent Valuation Methodology in Natural Resource Damage Assessments, 26 Ecology L.Q. 48 (1999).
277 Dep’t of the Interior, 880 F.2d at 476–78 (upholding as consistent with congressional intent DOI’s endorsement of the use of contingent valuation methodology to measure use, existence, and option values); see also Manus, supra note 13, at 446–49 (providing an overview of contingent valuation method and observing that much of the debate surrounding natural resource damage valuation has centered on that method).
278 For a description of Habitat Equivalency Analysis, see supra note 266.
279 United States v. Fisher, 977 F. Supp. 1193, 1198–1200 (S.D. Fla. 1997), aff’d, 174 F.3d 201 (11th Cir. 1999). Similarly, the natural resource damage regulations adopted by
The effort of natural resource damages to compensate losses that range from purely economic to nonutilitarian underscores its broad-based conception of natural resources and recognition of resources beyond their strict market valuation. In fact, the effort to reflect nonutilitarian losses to humans has been described as “the most poetic and profound element” of natural resource damages law.280

In addition to its expansive definition of natural resources and willingness to accept nontraditional economic methods of valuing harms, another critical attribute of natural resource damages law is its inherent recognition that natural resources are part of the public trust. There are several basic natural resource damage provisions that frame the public trust character of natural resources. First, both CERCLA and OPA limit the pursuit of natural resource damages to federal, state, or tribal trustees.281 These trustees are charged with acting “on behalf of the public . . . as trustees of natural resources. . . . ”282 Second, natural resources are defined to include “other such resources belonging to, managed by, held in trust by, appertaining to, or otherwise controlled by” federal, state, and tribal governments.283 Finally, trustees are allowed to use natural resource damages recoveries “only to restore, replace, or acquire the equivalent of” the affected resources.284 Recovered monies, in other words, must actually be used to directly benefit the public; trustees cannot simply place natural resource damage recoveries into their general revenue funds.

Several scholars have recognized the benefits of incorporating the public trust model into natural resource damages law. Professor Manus, for instance, explains that

the incorporation of the public trust model into the [natural resource damage] provisions of several major statutes may represent a positive

---

280 Manus, supra note 13, at 446.
282 Oil Pollution Act § 2706(b); Comprehensive Environmental Response Compensation, and Liability Act § 9607(f)(1).
reminder to government officials, particularly those in the agriculture, forestry, and other environmentally related agencies, that overriding all of their individual actions and decisions is an obligation to present and future generations to prevent an irrevocable imbalance in nature.\textsuperscript{285}

In this spirit, The National Association of Attorneys General has emphasized the importance of the trust relationship in natural resource damages law:

\begin{quote}
[t]he states and the Federal Governments [sic] are trustees for the people, and . . . their trust corpus includes this nation’s glorious natural resources. We, as trustees, have an obligation to protect these often irreplaceable resources from harm, and those that harm them have the obligation to restore them for all the people.\textsuperscript{286}
\end{quote}

The public trust aspect of natural resource damages law underscores the promise that the law holds in serving as the foundation for developing remedies for misuses of nature more generally. If we recognize that a broad range of natural resources and ecosystem services are part of the public trust, we can better accept more fundamental principles such as the notion that we should not misuse nature and that maintaining strict boundaries between public and private property is unrealistic.

\textbf{B. The Limits of Natural Resource Damages Law}

Given these three attributes—broad definition of natural resources, nontraditional means of valuing injuries to natural resources, and recognition of natural resources as public trust—natural resource damages seem to be a logical starting place from which to build a more comprehensive approach to remedying misuses of nature. To that end, it is useful to understand its current dimensions and limitations. In general, natural resource damages law suffers from limitations of scope, accessibility, and process.

Instead of being universally available to redress harms to natural resources, they are only available in particular situations that serve the purpose of the statutes

in which they are housed. In this way, natural resource damages are limited in scope and embody the same fragmentation that plagues environmental law more generally. For example, natural resource damages are only available under CERLCA when there has been a release of hazardous substances into the environment. Likewise, under OPA and the CWA, a release of oil into navigable waters is required to trigger liability for natural resource damages. Other, less prominent statutes containing natural resource damage provisions—for example, the Marine Protection Research and Sanctuary Act—suffer from similarly narrow applications by limiting natural resource damages to very specific categories of federal lands.

Not only are natural resource damages limited to certain types of releases or harms, but they are also limited to resources “belonging to, managed by, held in trust by, appertaining to, or otherwise controlled by” federal, state, or tribal governments. While this definition does not actually require federal or state governments to own the natural resources, it does exclude purely private property from the scope of natural resource damages. Given that private property constituted over sixty percent of the land in the United States, excluding this land from the purview of natural resource damages is a significant limitation to any framework seeking a comprehensive remedy for the misuse of nature.

In addition to their limited scope, natural resource damages are also limited in their accessibility. As described above, only authorized trustees are permitted to pursue natural resource damage recoveries. This means that private plaintiffs seeking to recover damages for injuries caused by harm to the environment—for example, the private landowners or commercial fishermen plaintiffs in the Exxon Valdez oil spill litigation—could not unilaterally avail themselves of CERCLA or

287 See supra Part II.B.
288 42 U.S.C. § 9607(a), (f).
291 Comprehensive Environmental Response Compensation, and Liability Act § 9601(16) (definition of natural resources).
292 Ohio v. U.S. Dep’t of the Interior, 880 F.2d 432, 460 (D.C. Cir. 1989) (concluding that Congress deliberately excluded purely private property from the definition of natural resources).
294 See supra Part I.B.
OPA’s natural resource damage provisions. Local governments and private parties have to rely on the relief provided under state or common law.  

Foreclosing private citizens from pursuing natural resource damages places the burden of pursuing natural resource damages solely in the hands of agencies with limited time and funding. Without suggesting that natural resource damage recoveries be awarded to private citizens, it is possible that individuals could step into the shoes of trustees through citizen suit-type provisions. In environmental laws generally, citizen suit provisions have been a significant force behind the successful enforcement and interpretation of important statutes such as the Clean Air Act, the Clean Water Act, and the Endangered Species Act. A similar concept could make natural resource damages a more effective remedial tool as well.

The third major limitation of natural resource damages law is its limitation of process; even if federal natural resource damages were more widely available in scope and accessibility, the costs, controversy, and delay associated with assessing and valuing natural resource injuries have limited their usefulness in practice.

In order to support a claim for natural resource damages, for instance, trustees typically prepare a natural resource damage assessment. While not mandated by regulation, a natural resource damages assessment is a key component of proving the extent of damages to which trustees are entitled. To that end, CERLCA creates a rebuttable presumption in favor of damages assessed according to

---

296 42 U.S.C. § 9652(d) (CERLCA does not modify remedies available under the common law).

297 Elizabeth Magill, Standing for the Public: A Lost History, 95 Va. L. Rev. 1131, 1187–90 (2009) (discussing the significance of citizen suit provisions in federal environmental law and explaining that the first such provision, contained in the Clean Air Act of 1970, “imagined citizens as both private enforcers of existing EPA dictates as well as direct watchdogs on EPA activities”).

298 “The measurement of damages, however, is one of the most controversial aspects of a natural resources damage action.” Gerald F. George, Litigation of Claims For Natural Resource Damages, in ENVIRONMENTAL LITIGATION 399, 410 (A.L.I.-A.B.A. Court of Study Materials, June 26–30, 2000). “In most instances, these determinations as well as the scaling of the restoration or compensation activities, will require data and assumptions about ecological services, injury, recovery and productivity rates that are much easier to describe than to develop in practice.” Id. at 412.


300 See generally Craig R. O’Connor, Natural Resource Damages Under the Comprehensive Environmental Response, Compensation, and Liability Act, in ENVIRONMENTAL AND TOXIC TORT MATTERS: ADVANCED CIVIL LITIGATION (A.L.I.-A.B.A. Course of Study Materials February, 18, 1999) (noting that the defensibility of a natural resource damages assessment in court is “vital to the ability of the trustees’ legal representatives to recover the funds necessary to restore injured resources”).
regulatory procedures. Without that presumption, the trustees face the unenviable task of attempting to prove that a particular person caused the harms identified in the natural resource damage claim and that the cost of those harms is properly valued; without the benefit of a rebuttable presumption, CERCLA places the burden of proving and valuing ecological injuries on the trustee. Given the complexity, dynamism, and interconnectedness of natural systems, the task of identifying, let alone valuing, natural resource injuries can be a substantial and formidable undertaking. Because trustees face a “significant proof obstacle” in the identification and valuation of natural resource damages, these assessments are, as a practical matter, indispensable for trustees seeking to avoid challenge and costly litigation by potentially responsible parties.

But natural resource damage assessments are “costly propositions,” and the prospect of undertaking these costs puts pursuit of natural resource damages out

302 Even with the presumption, the task of proving and valuing damages is difficult. In fact, the D.C. Court of Appeals has observed that the presumption does not appear to have created the “powerful advantage” as originally envisioned, but rather has amounted to “nothing more than a burden shifting exercise.” Gen. Elec. v. U.S. Dept. of Commerce, 128 F.3d 767, 772 (D.C. Cir. 1997).
303 Professor Manus posits that Congress’s allocation of the burden of proving liability for natural resource damages on trustees is “[p]erhaps the most telling sign of congressional reticence to allow ready implementation of the ultimate goals of [natural resource damages].” Manus, supra note 13, at 440.
305 Manus, supra note 13, at 442–43 (“[T]he requirement that trustees prove injury could eviscerate the [natural resource damages] claim from within, because in the web of life no ecological change occurs in the sterile isolation required for proof of causation.”).
306 New Mexico v. Gen. Elec. Co., 467 F.3d 1223, 1242 n.28 (10th Cir. 2006) (“[T]he court is well aware that NRD assessment is a costly proposition.”). In this case, the Tenth Circuit went on to explain that

[a]ccording to two commentators, after its 1986 amendments, CERCLA ‘cast trustees adrift to finance their own damage assessment before filing claims against polluters—a costly proposition, given that damage assessments typically cost millions of dollars. This lack of funding has created a virtually insurmountable obstacle considering that agency budgets have historically authorized little or no funding for NRD assessments.’

of reach for some state and federal trustees. Although polluters are ultimately responsible for reimbursing the trustees for the cost of the assessment for successful claims, the time and expense of conducting such an assessment is more than the budgets of many trustees can bear in the first instance.

Moreover, the inherently complex nature of delineating natural resource injuries turns the preparation of natural resource damage assessments into a long and arduous process. As NOAA explains:

Although the concept of assessing injuries may sound simple, understanding complex ecosystems, the services these ecosystems provide, and the injuries caused by oil and hazardous substances takes time—often years. The season the resource was injured, the type of oil or hazardous substance, and the amount and duration of the release are among the factors that affect how quickly resources are assessed and restoration and recovery occurs. The rigorous scientific studies that are necessary to prove injury to resources and services—and withstand scrutiny in a court of law—may also take years to implement and complete.

This means that the time between the event that gives rise to the injury and the settlement of the natural resource damages can take decades in some cases. A 2004 study of state trustees involved in natural resource damage cases found that the average length of time between event and settlement is eleven years.

______________________________


308 Id. Some states, like New Jersey and California, have devised relatively successful natural resource damage programs because they have made funding available for damage assessments up front. Id. at 438–44 (discussing the New Jersey approach to natural resource damages). In addition, these states have also been successful at reducing the cost of assessments by enacting regulations that approve certain simplified methodologies. Id. at 445. Those simplified methodologies, however, are nonetheless vulnerable to legal challenge absent any presumption in favor of the trustee if methodologies are followed. Id.


311 Id.
finding is consistent with the conclusions of a federal advisory committee report released in 2007 that found the federal natural resource damages process needs to be revised in order to make restoration of natural resources “faster, more efficient, and more effective.”

To that end, the advisory committee’s recommendations were aimed, in part, at establishing cooperative relationships with potentially responsible parties in order to: (1) encourage responsible parties to fund damage assessments in the first instance; and (2) avoid valuation issues by encouraging responsible parties to conduct the restoration activities.

Until natural resource damage processes and methodologies can be revised to facilitate effective recovery of public costs to natural resources, they will most likely continue to be underutilized and only be successful in the unusual case.

---


313 On this latter point, one of the most hotly contested issues in natural resource damage claims is valuation. See, e.g., Allan Kanner & Tibor Nagy, Measuring Loss of Use Damages in Natural Resource Damage Actions, 30 COLUM. J. ENVTL. L. 417, 448 (2005) (citing Richard Stewart et al., Evaluating the Present Natural Resource Damages Regime: The Lawyer’s Perspective, in NATURAL RESOURCE DAMAGES: A LEGAL, ECONOMIC, AND POLICY ANALYSIS 163 (Richard Stewart ed., 1995) (“[M]easuring natural resource damages is ‘the most daunting task facing trustees.’”)); James L. Nicoll, Environmental Restoration: Challenges for the New Millennium: The Irrationality of Economic Rationality in the Restoration of Natural Resources, 42 ARIZ. L. REV. 463, 464 (2000) (challenging traditional economic theory in the valuation of natural resources); Thompson, supra note 312, at 60 (“Natural resource damages present a significant challenge for the legal system because in most cases they are non-market commodities.”). How do we effectively measure the loss of, or injury to, certain resources? One approach is by measuring the use and existence value of the resource from a utilitarian perspective, for example, the worth of the resource measured by its value to individuals or society. See Peck, supra note 156, at 279–82. Another approach—the biocentric approach—would measure the intrinsic value of the resource independent of human satisfactions. Id. Not surprisingly, the preferred method of valuing natural resources is to quantify utilitarian values of use and existence through some method of cost-benefit analysis. Id. at 281–82. Three of the most common methods for measuring the value of natural resource damages are market valuation, restoration and replacement cost, and contingent valuation. Id. at 282–85. Regardless of the chosen method, however, controversies will certainly arise given that natural resource damages are unique in many instances and their uses not readily subject to valuation.
IV. SKETCHES OF A REFORMED LAW

Unlike existing laws that tend to protect private owners against damage by others, the reforms proposed in this Article would draw upon and expand on laws that protect parts of nature from damage by the owner itself. Anticruelty laws represent one area where society has deemed it necessary and ethical to protect the thing that is owned from abuse by the owner. For example, the federal Animal Welfare Act requires minimum standards of care and treatment for certain animals bred for commercial sale or transported commercially. At the state level, anticruelty laws protect privately owned animals against cruelty and neglect; the majority of states even provide for the seizure of animals that are being mistreated.

Other laws, though not as obvious in their delivery, also protect elements of nature from abuse by their owners. At the federal level, the Endangered Species Act protects certain listed species and their habitat from harm even if they exist on private lands. The ESA prohibits all people, whether private landowners or

---

314 For example, mining laws like the General Mining Law of 1872 govern individual rights to locate and acquire mineral rights; the law is focused on resolving competing claims between individuals or between individuals and the government. 30 U.S.C. § 22 (2006) (giving “free and open” rights to exploration and purchase of hard rock minerals); id. § 28 (explaining how claims are located and recorded so as to protect against competing claims); see also GEORGE CAMERON COGGINS ET AL., FEDERAL PUBLIC LAND AND RESOURCES LAW 558–84 (6th ed. 2007) (addressing disputes under the General Mining Law). Similarly, the Taylor Grazing Act implements a regime for allocating private grazing permits on public lands. 43 U.S.C § 315 (2006) (setting forth the purpose of the Act as promoting the highest use of public lands); Pub. Lands Council v. Babbitt, 529 U.S. 728, 728, 733–36 (2000) (describing the Act as authorizing the Department of the Interior to “divide the public rangelands into grazing districts” and grant permits to landowners engaged in the livestock business). Water law doctrines like prior appropriation and reasonable use resolve water allocation disputes between individual water users. See COGGINS ET AL., supra at 486–88 (providing a brief overview of the prior appropriation and reasonable use doctrines). Even frequently invoked common law doctrines in the area of natural resources, like nuisance, are aimed at resolving disputes between individual landowners. See, e.g., Bormann v. Bd. of Supervisors in and for Kossuth Cnty., 584 N.W.2d 309, 311 (Iowa 1998) (landowner challenge to county decision to designate certain land as an “agricultural area,” making Confined Area Feeding Operations (CAFOs) immune from nuisance suits). In each of these examples, natural resource laws decide which landowner, speculator, or right holder is entitled to use nature; they do not necessarily guarantee nature itself will be protected from overuse.


commercial businesses, from taking listed endangered and threatened species.\textsuperscript{318} Under some circumstances, this taking prohibition can also protect against habitat modification.\textsuperscript{319} Some state endangered species acts contain similar protections.\textsuperscript{320}

Even nonwildlife elements of nature receive some protection from misuse under various federal, state, and common laws. Wetlands are protected by section 404 of the Clean Water Act.\textsuperscript{321} Instream water flows—which provide benefits such as flood mitigation, groundwater recharge, biological productivity, and recreational opportunities—are also often protected. For example, Washington State enacted a Water Resources Act that requires rivers and streams within the state to “be retained with base flows necessary to provide for preservation of wildlife, fish, scenic, aesthetic and other environmental values, and navigational values.”\textsuperscript{322} Likewise, under the Washington Water Code, permits for water use rights are conditioned upon the protection of instream flows.\textsuperscript{323}

As a final example of current efforts to protect nature from misuse by its owners, farm programs like the Conservation Reserve Program expressly limit farm practices that cause soil erosion as a condition of government funding.\textsuperscript{324} In particular, the program provides that program participants must “[e]stablish and maintain the required vegetative or water cover and the required practices on the land . . . to achieve the desired environmental benefits and to maintain the productive capability of the soil . . . .”\textsuperscript{325}

Though present, these examples are limited to narrow instances, as when a species is in danger of extinction or when a landowner voluntarily signs up for a conservation program. To protect ecosystem services, these specific rules need to be expanded to a more generalized prohibition against misuse of nature. This scheme would not be intended to replace existing common law or statutory remedies, but rather work in tandem with them to more appropriately protect ecosystem services instead of just selected parts of the patchwork.

For natural resource damages to serve as a realistic foundation for developing a comprehensive framework for remedying misuses of nature, it must protect nature while respecting the right of private property owners to use those resources

\textsuperscript{318} Id. § 1538(a).

\textsuperscript{319} See supra notes 178–182 and accompanying text.

\textsuperscript{320} See, e.g., COLO. REV. STAT. § 33-2-105(3), (4) (2011) (prohibiting any person from taking endangered or threatened species); 520 ILL. COMP. STAT. ANN. 10/1 to /11 (West 2004) (prohibiting any person from taking listed species); MINN. STAT. ANN. § 84.0895 (West 2004) (protection threatened and endangered species).


\textsuperscript{322} Water Resources Act of 1971, WASH. REV. CODE ANN. § 90.54.020(3)(a) (West 2004).


\textsuperscript{324} 7 C.F.R. § 1410.20(a) (2011) (setting forth the participant’s obligations and requiring practices that reduce soil erosion as a condition of participation).

\textsuperscript{325} Id. § 1410.20(a)(6).
in legitimate ways. It must also accurately incorporate evolving scientific understanding and reflect the realities of the land itself. In particular, there are at least eight key issues that a new natural resources damages framework will have to resolve:

(1) Accommodating the Uniqueness of Land

There are natural variations among landscapes, with different lands embodying varying levels of resilience to similar uses. What is sustainable or acceptable land use in one area might not be sustainable in the other. A reformed natural resource damages law should recognize those differences and avoid land use restrictions simply because certain practices are detrimental to substantially different types of land. Professor Freyfogle has made a similar point in suggesting that a new, ecological approach to landownership should tailor private rights “to take into account the natural variations among land parcels.”

Incorporating a scientifically based understanding of the landscape would be useful in this regard. For example, ecoregions might prove a useful tool for building a catalog of unacceptable land use practices for various ecotypes. Ecoregions denote areas of general similarity in ecosystems and in the type, quality, and quantity of environmental resources. They are designed to serve as a spatial framework for the research, assessment, management, and monitoring of ecosystems and ecosystem components. By recognizing the spatial differences in the capacities and potentials of ecosystems, ecoregions stratify the environment by its probable response to disturbance.

The difficulty in applying science (such as knowledge of ecoregions) will be in providing enough differentiation to accommodate lands’ unique character, and yet enough bright line rules to provide property owners with clarity as to what legitimate uses of property include. Only by providing such clarity do we preserve the stability of property rights. It is also worth recognizing that this challenge of matching acceptable land use and resource management with varying types of land is made even more difficult in the face of climate change, where weather patterns

326 FREYFOGLE, supra note 12, at 229; see also Freyfogle, Particulars of Owning, supra note 21, at 585 (“Slowly, painfully, people are coming to think that landowner rights should somehow depend on the natural features of the parcel owned.”).
will inevitably alter the landscape and call into question whether uses once deemed sustainable can still qualify as such.\(^{328}\)

(2) Drawing the Line between Use and Misuse

Not all land use and natural resource consumption is bad or unacceptable.\(^ {329}\) Consumption is necessary for survival. As a result, because we cannot simply set land aside and preserve it in perpetuity, a reformed natural resource damages law will have to undertake the challenge of identifying workable rules that balance use against misuse. Naturally, such line drawing will have to accommodate our constantly evolving understanding of ecology. Moreover, because line drawing is a reflection of values in one capacity or another, balancing use against misuse should be rooted in a well-defined and articulated ethical framework.

(3) Creating a New Cause of Action

A central task in shaping a reformed natural resource damages law will be the creation of a new cause of action setting forth the elements of proof. This cause of action will need to be flexible enough to accommodate the uniqueness of land, the ever-changing ecological understanding, and the many types of resource use that this law would ideally protect. On the other hand, the cause of action must not permit so much flexibility that it lacks clarity, for in issues dealing with private property, clarity provides stability.\(^ {330}\) Additionally, determining how this cause of action will intersect with existing state property, tort, and natural resource laws will also need to be worked out.

In striking this balance between flexibility and predictability, select elements of the common law of nuisance and current natural resource damages law are important. Nuisance law is a model of simplicity and flexibility, providing a remedy for injuries ranging from interstate disputes over air pollution\(^ {331}\) to private property disputes over the location of cattle-feeding operations.\(^ {332}\) Federal natural resource damages law, on the other hand, prescribes a framework that is more consciously tailored to some of the unique characteristics of ecological injury. For example, natural resource damages law recognizes the public trust component of

\(^{328}\) See, e.g., Gary W. Yohe et al., Perspectives on Climate Change and Sustainability, in CLIMATE CHANGE 2007: IMPACTS, ADAPTATION AND VULNERABILITY, CONTRIBUTION OF WORKING GROUP II TO THE FOURTH ASSESSMENT REPORT OF THE INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE 811–41 (Martin L. Parry et al. eds., 2007) (“Climate change adds to the list of stressors that challenge our ability to achieve the ecologic, economic and social objectives that define sustainable development.”).

\(^{329}\) See discussion supra Part I.

\(^{330}\) See, e.g., FREYFOGLE, supra note 7, at 87 (2007) (“For private property to produce its many possible benefits landowners need protection for at least certain expectations.”).


natural resources and accordingly permits only designated federal, state, and tribal
trustees to pursue natural resource damages.\textsuperscript{333} Moreover, natural resource
damages law also tailors causal elements to accommodate the difficulty in proving
causation for ecological injuries.\textsuperscript{334}

In the end, the new cause of action will likely be a unique blend of elements
borrowed from traditional causes of action and ones that have yet to be devised.
Along these lines, Professor Bill Rodgers suggests the creation of a “new tort of
contamination,” which he would define as the “interference with the use and
enjoyment of ecosystem functions.”\textsuperscript{335} In his loose sketch of what this new tort
might look like, Rodgers proposes “a strict liability tort with elements being (1)
contamination, (2) causation, and (3) prospects of remediation.”\textsuperscript{336} In this regard,
Professor Rodgers seems to start with common law of nuisance and weave in
elements of current natural resource damages law. Though the details are left for
another day, what is important at this point is the recognition that a legal tool is
missing from the environmental protection toolbox. As Professor Rodgers aptly
observed:

Contamination of natural capital—decline in the baseline—might be
explained on the grounds of ownership failure. The commons may lack
the vigorous champion of private entitlement. But the fish, the drinking
water, the shellfish beds, and the body burdens do not appear to lack
necessary plaintiffs. What is missing is a legal tool to correct the
situation.\textsuperscript{337}

(4) Proving Causation for Ecological Injuries

Although there are many possible forms that a new cause of action for natural
resource damages might take, proving causation is sure to be one of the central
elements. Because proving causation for ecological injuries is not an easy task, it
deserves separate consideration.

Ecological injuries are characterized by five defining characteristics—
complexity, scientific uncertainty, dynamism, precaution, and controversy.\textsuperscript{338}
Because of these characteristics, ecological injuries give rise to difficult problems
in identifying the extent and duration of harm, in proving causal relationships
between manifested harm and the allegedly environmentally detrimental act, and in
providing such identification and proof within the time frame usually required

\textsuperscript{333} See supra Part I.B.
\textsuperscript{334} See infra notes 338–347 and accompanying text.
\textsuperscript{335} Rodgers, supra note 11, at 1259.
\textsuperscript{336} Id.
\textsuperscript{337} Id. at 1257.
\textsuperscript{338} LAZARUS, supra note 304, at 16.
from response plans and redress paradigms. A reformed natural resource damages law will have to overcome these obstacles in identifying ecological injury and proving causation for those injuries.

In his call for a “tort of contamination,” Professor Rodgers offered CERLCA as an example of an environmental statute that incorporates a nontraditional causation standard. CERCLA is a strict liability scheme that imposes joint and several liability on parties found to be responsible for contaminating a site. But because CERCLA defines responsibility broadly, a plaintiff does not have to show that a defendant actually caused the release that resulted in the incurrence of response costs. Rather than employing a traditional causation test that requires tracing the alleged harm to a particular actor or set of actors, CERCLA adopts a more flexible standard that simply requires that the release was “likely to have been a causative factor” giving rise to the alleged harm. Thus, once the plaintiff has proven that the defendant is a potentially responsible party (PRP), a rebuttable presumption of liability arises, and the burden of proof shifts to the defendant, who then may assert an affirmative defense by disproving causation.

By eliminating traditional tort causation from the plaintiff’s prima facie case, the CERLCA standard recognizes that harms to nature can rarely be proven as a straight-line correlation between action and reaction. Whether under this test or

---

339 See Knudsen, supra note 304, at 99.
340 Rodgers, supra note 11, at 1259.
341 Id.
343 Steve C. Gold, Dis-Jointed? Several Approaches to Divisibility After Burlington Northern, 11 VT. J. ENVTL. L. 307, 328 n.126 (2009) (citing to several cases where courts have recognized CERCLA’s truncated causation requirement).
344 Rodgers, supra note 11, at 1259 (citing 4 WILLIAM H. RODGERS, JR., ENVIRONMENTAL LAW: HAZARDOUS WASTE & SUBSTANCES § 8.11, at 660–66 (West 1992)); see also United States v. Alcan Aluminum Corp., 990 F.2d 711, 721 (2d Cir. 1993); Amoco Oil Co. v. Borden, Inc., 889 F.2d 664, 668 (5th Cir. 1989) (citing Ascon Props., Inc. v. Mobil Oil Co., 866 F.2d 1149, 1152–53 (9th Cir. 1989)); Patricia E. Lin & Tom Starnes, Establishing Liability Under CERCLA: The Causal Nexus and the Alternative Liability Theory 1 (June 5, 2000), available at http://www.andrewskurth.com/assets/pdf/article_48.pdf (noting that under 42 U.S.C. § 9607(a) (2006), a plaintiff must prove only (1) that there has been a release or threatened release covered by CERCLA, (2) that the plaintiff has incurred response costs, (3) that the response costs were necessary and consistent with the national contingency plan, and (4) that the defendant is a PRP (potentially responsible party) as defined by CERCLA).
345 See Amoco, 889 F.2d at 668; United States v. Monsanto Co., 858 F.2d 160, 170–71 (4th Cir. 1988).
346 See John Copeland Nagle, CERCLA, Causation, and Responsibility, 78 MINN. L. REV. 1493, 1506–08, 1511 n.86 (1994). To prove causation under traditional tort law, a
a different one, the new natural resource damages scheme would similarly benefit from recognizing the unique character of environmental harm and tailor the causation standard accordingly.347

(5) Addressing the Likely Aggregate Nature of Harm

Ecosystems cross ownership boundaries.348 And ecological harms, particularly those stemming from land use and natural resource consumption, often result from compounding stresses to the land.349 Many landowners might, therefore, cause harm to nature through collective, if uncoordinated, action.350

For the purposes of reforming natural resource damages law, the aggregate nature of some ecological harm makes determination of liability difficult but essential. If harm is caused by several actors who are not necessarily operating in concert, who should be held accountable? One could, for instance, choose to hold accountable the last actor who stresses the resource beyond its natural capacity for alteration. This would make it incumbent on any actor to consider the likely cumulative impact of his or her proposed actions on the ecosystem. If that action would trigger an unacceptably negative ecological response, the actor would be held accountable for resulting injuries, notwithstanding the fact that the injury may have been the product of several collective actions. NEPA imposes similar requirements on federal agencies to consider the cumulative impacts of their proposed actions before deciding on an alternative.351 Last-in-time liability seems to be a rather simple system for assessing liability, assuming one could identify which actor was last in time. But what if the last actor is the one whose land use provides the greatest utility to the community? Should that actor be prohibited plaintiff would have to trace particular molecules of contamination back to the defendant, who may have deposited the materials at the release site decades ago.

347 It is important to note that the relaxed test for causation under CERCLA applies only to actions seeking recovery for response costs. For natural resource damages, CERLCA requires a causal link between the responsible party and the injury to the resource. See 43 C.F.R. § 11.62 (2011) (providing a procedure for establishing causation for water resources, geological resources, and biological resources injury). This regulation was upheld in Ohio v. U.S. Dep’t of the Interior, 880 F.2d 432, 468–73 (D.C. Cir. 1989). To demonstrate injury and causation, the resource must be characterized, samples collected and statistically compared to measure injury, and the discharge modeled through various possible pathways. See 43 C.F.R. § 11.61 (2011).

348 See supra Part I.B.1.


350 See supra Part I.B.2 for a discussion of aggregate harm and the tragedy of fragmentation.

351 See 40 C.F.R. §§ 1508.7, 1508.25(c) (2011).
from taking action simply because others have collectively made inefficient or reckless resource decisions in the past?

Another option for assessing liability would be holding all contributing actors jointly and severally liable for injuries caused by their aggregate activities. Such a system could look to CERLCA’s model of assessing joint and several liability in the environmental contamination context. However, the injuries targeted by CERLCA are more focused than the more comprehensive set of harms that reformed natural resource damages law would be designed to address. CERLCA deals with injuries arising from releases of toxic substances. And although such releases might be an inevitable byproduct of industrial activities, most people would probably agree that all toxic releases are undesirable and should be avoidable given enough care or incentive. In that sense, holding all contributing actors jointly and severally liable for toxic releases does not offend the equitable senses. But injuries arising from resource use generally—aggregate resource use in particular—are not so easily judged. From an equity standpoint, not everyone who contributes to the aggregate harm had equal reason to foresee that use of a given resource would unacceptably stress the ecosystem. Why then should all actors be held joint and severally liable? Moreover, what if all of the actors cannot be identified because of the multiple and complex pathways in which ecological harm can manifest?

The area of wildlife management provides a general example of a situation in which multiple stressors contribute to ecological harm. Wildlife population declines can result from the combination of many negative ecological stressors including “alteration of habitat caused by patterns of agricultural and urban land use, introduced invasive and exotic species, nutrient enrichment, direct human disturbance, and toxic chemicals.” The presence of multiple stressors complicates the ability of wildlife managers to identify which policy changes would most effectively curtail the population decline.

A similar situation is one where there is a single stressor from multiple sources, unwittingly joining forces to overcome an ecosystem’s natural assimilative capacity. Climate change is one prominent area where the issue of liability for collective harms proves exceptionally difficult and important. Because the Earth’s natural systems are capable of assimilating some amounts of

---

352 See 42 U.S.C. § 9607(a) (2006) (the civil liability provision of CERLCA). While the final version of CERLCA deleted explicit references to joint and several liability, courts have held that potentially responsible party liability is joint and several if no basis exists for dividing the harm of the contamination and the response costs. See Steven Ferrey, *Converting Brownfield Environmental Negatives into Energy Positives*, 34 B.C. Envtl. Aff. L. Rev. 417, 460–64 (2007) (citing Amoco Oil Co. v. Borden, Inc., 889 F.2d 664, 672 (5th Cir. 1989), and United States v. Monsanto Co., 858 F.2d 160, 169–70 (4th Cir. 1988)).


355 *Id.*
greenhouse gases without identifiable disruption, one actor emitting carbon dioxide into the atmosphere causes no discernible injury. But multiple actors responding to our combined unyielding thirst for development, industrialization, and consumption has stressed the earth’s natural systems beyond capacity.\footnote{See supra notes 1–5 and accompanying text.} This stress has manifested in very real and identifiable injuries. Given the collective actions contributing to this stress, can any single actor or category of actors be held accountable?

There have only been a handful of climate change litigation cases framed as common law nuisance actions.\footnote{David Markell & J.B. Ruhl, An Empirical Survey of Climate Change Litigation in the United States, [2010] 40 ENVTL. L. REP. (Envtl. Law Inst.) 10,644, 10,647.} In general, those actions have sought to impose property damage or personal injury liability on sources of greenhouse gas emissions for failing to properly mitigate climate change. Most notably, in \textit{Connecticut v. American Electric Power Co.}\footnote{406 F. Supp. 2d 265 (S.D.N.Y. 2005), vacated and remanded, 582 F.3d 309 (2d Cir. 2009), rev’d, 131 S. Ct. 2527 (2011).} plaintiffs invoked the common law public nuisance doctrine to file a lawsuit against six electric power companies for contributing to global warming.\footnote{Id. at 267.} Together, the defendants operated approximately 174 fossil fuel-fired power plants in twenty states that allegedly constituted ten percent of all carbon dioxide emissions in the United States.\footnote{Id. at 268.} The federal district court dismissed the lawsuit on political question grounds.\footnote{Id. at 274.} On appeal, the Second Circuit reversed and allowed the public nuisance claim to proceed.\footnote{Connecticut v. Am. Elec. Power Co., 582 F.3d 309, 315 (2d Cir. 2009), rev’d, 131 S. Ct. 2527 (2011).} In particular, the Court suggested that joint and several liability is appropriate in the climate change context.\footnote{Id. at 274.} To that end, in the course of addressing whether plaintiffs satisfied the redressability element of standing, the court noted that in federal common law of nuisance cases involving air pollution, “ambient air contains pollution from multiple sources” and “liability is joint and several.”\footnote{Id. at 349.} The U.S. Supreme Court reversed, concluding that the Clean Air Act has displaced federal common law of nuisance in the area of greenhouse gas emissions.\footnote{Am. Elec. Power Co. v. Connecticut, 131 S. Ct. 2527, 2537 (2011).} The Court did, however, leave open the possibility that climate nuisance litigation could be viable under a state common law of nuisance theory.\footnote{Id. at 2540.}

Scholars have begun to tackle this issue as well. Professor David Grossman has examined joint and several liability in the global warming context and has concluded that “one could reasonably argue that it is possible to identify
defendants who have contributed substantially to climate change and its resulting effects.\footnote{David A. Grossman, \textit{Warming Up to a Not-So-Radical Idea: Tort-Based Climate Change Litigation}, 28 COLUM. J. ENVTL L. 1, 28 (2003).} Given equity concerns of holding a few actors liable for harm as cumulative and disperse as climate change, Grossman suggests, “apportioning damages (appropriately reduced to account for past emissions) based on a combination of defendants’ market-shares and the greenhouse gas emissions of their products.”\footnote{Id. at 32–33. Professor J.B. Ruhl described this liability conundrum in the context of enforcing the Endangered Species Act’s take prohibition for harms caused by climate change impacts. J.B. Ruhl, \textit{Climate Change and the Endangered Species Act: Building Bridges to the No Analog Future}, 88 B.U. L. REV. 1, 40–42 (2008); see also James R. Rasband, \textit{Priority, Probability, and Proximate Cause: Lessons from Tort Law About Imposing ESA Responsibility for Wildlife Harm on Water Users and Other Joint Habitat Modifiers}, 33 ENVTL. L. 595, 618–23, 628–30 (2003) (criticizing the approach of prosecuting only major contributors for takings violations in the western water diversion context, another area in which there are several and dispersed causal agents giving rise to harm).}

As additional cases and scholarship develop in the context of climate change litigation, fresh perspectives on imposing liability for ecological injuries from aggregate and disperse activities might also prove useful in developing a new natural resource damages law.

(6) Choosing Appropriate Plaintiffs

Part of creating a new cause of action is deciding who might be appropriate to bring claims for harms to nature. Given the public nature of the resources at issue, the plaintiffs raising such claims would be acting in the public interest and must therefore be selected with that important caveat in mind.\footnote{In his discussion of a new “tort of contamination,” Rodgers similarly observed that “the tort of contamination protects public interests, and appropriate plaintiffs should be selected with that goal in mind.” Rodgers, \textit{supra} note 11, at 1260.}

A ready-made solution would be to follow the current natural resource damages law that permits only federal, state, or tribal trustees to initiate claims for natural resource damages. While such a limitation might ensure that the public interest is served, current natural resource damages law has demonstrated that reliance on select trustees to litigate natural resource injuries can stifle the effectiveness of the remedy.\footnote{See \textit{supra} Part II.B.} Relying on designated trustees to identify and pursue remedies for harms to nature would be like relying on federal agencies alone to enforce all environmental statutes without the benefit of citizen suit...
provisions. Perhaps this is why citizen suits have played a central role in the enforcement, clarification, and effectiveness of environmental statutes.\footnote{371}

Just as citizen suit provisions in many environmental statutes have increased the enforcement and effectiveness of environmental regulations, such provisions might also be instrumental in effectively remediying harms to nature. Given the goal of serving the public interest, a modified citizen suit provision might be considered; rather than permitting “any person” to assert a claim for harms to nature, a modified citizen suit provision could limit plaintiffs to designated public trustees as well as public interest organizations who are able to make a nominal showing of competency and experience in litigating environmental issues. And, similar to the sixty-day notice requirement contained in most environmental citizen suit provisions,\footnote{372} a modified citizen suit provision for the new natural resource damages law could require notice to public trustees. In this way, trustees, like federal or state agencies, could elect to pursue certain claims of harm deemed particularly sensitive or important to the public trust resources. Other attributes of a modified citizen suit provision might include the right to intervene in cases brought by public trustees upon showing of standing and good cause, or limits to public participation in cases seeking criminal penalties.

In general, the key to selecting appropriate plaintiffs would be affording public access to the courts without sacrificing competent and genuine pursuit of the public interest.

(7) Defining Scope of Appropriate Remedies

It goes without saying that any scheme providing a remedy for the misuse of nature will have to decide what remedies are appropriate once liability is established. Many environmental statutes contain provisions allowing civil penalties, criminal penalties, and injunctive relief.\footnote{373} In light of these examples of penalty provisions, one might be most easily inclined toward providing a similar suite of remedies in the new natural resource damages law. One could also choose to borrow from existing natural resource damages statutes, adopting CERCLA and OPA’s requirement that monies recovered be put towards restoration.\footnote{374}


The type of remedies that are ultimately deemed appropriate under this reformed natural resources damage law depends in large part on the goal of the law itself. Professor Rodgers argues, for example, that a “tort of contamination” should strive to restore and rehabilitate.\textsuperscript{375} Similarly, the goal of a reformed natural resource damages law might be restoration to the extent conceivable. If the goal is to restore, then penalties would not necessarily be appropriate, unless they are required to be put toward restoration.\textsuperscript{376} On the other hand, if the goal is merely to punish wrongful acts, then we might view the question of traditional penalties differently and encourage penalties and criminalization.

Given how difficult recreating nature’s services can be,\textsuperscript{377} a suite of remedies that maximally deter misuse in the first instance would seem necessary and appropriate. At the same time, the remedies should ensure that once the damage is done, the effectiveness and efficiency of restoration efforts are maximized. Simply borrowing from current natural resource damages law on that issue would not be wise, as the federal natural resource damages process needs to be revised in order to make restoration of natural resources “faster, more efficient, and more effective.”\textsuperscript{378} Given that current natural resource damages law struggles to achieve effectiveness and efficiency of restoration,\textsuperscript{379} a better solution must be designed for a new scheme.

\textbf{(8) Measuring Natural Resource Damages}

Related to the issue of appropriate remedies is the narrower issue of how to value resource damages. Though relatively narrow, valuation is not a simple inquiry. Rather, it is one of the most hotly contested issues in natural resource damage claims.\textsuperscript{380} One approach suggests measuring the use and existence value of

\textsuperscript{375} Rodgers, 	extit{supra} note 11, at 1260.
\textsuperscript{376} Id. (“The history of monetary compensation for loss of sustainable resources is not a happy one. These ‘cash-outs’ can create momentary winners but with a poor distribution and sadly skewed (and sometimes opportunistic) calculation of what has been lost.”).
\textsuperscript{377} See discussion of the biosphere \textit{supra} notes 74–75 and accompanying text.
\textsuperscript{378} See NAT. RESOURCE DAMAGE ASSESSMENT & RESTORATION FED. ADVISORY COMM., \textit{supra} note 312, at 4. The Federal Advisory Committee’s recommendations were aimed, in part, at establishing cooperative relationships with potentially responsible parties in order to: (1) encourage responsible parties to fund damage assessments in the first instance; and (2) avoid valuation issues by encouraging responsible parties to conduct the restoration activities. \textit{Id.} at 13. Until natural resource damage processes and methodologies can be revised to facilitate effective recovery of public costs to natural resources, they will most likely continue to be underutilized and met with success only in unusual cases.
\textsuperscript{379} \textit{Id.} at 1.
\textsuperscript{380} See, e.g., Kanner & Nagy, \textit{supra} note 313, at 488 (citing Stewart, \textit{supra} note 313, at 163) (“[M]easuring natural resource damages is ‘the most daunting task facing trustees.’”); Nicoll, \textit{supra} note 313, at 464 (challenging traditional economic theory in the valuation of natural resources); Thompson, \textit{supra} note 312, at 60 (“Natural resource
the resource from a utilitarian perspective, namely the worth of the resource measured by its value to individuals or society. 381 Another approach—the biocentric approach—would measure the intrinsic value of the resource independent of human satisfactions. 382 Not surprisingly, the preferred method of valuing natural resources is to quantify utilitarian values of use and existence through some method of cost-benefit analysis. 383 Three of the most common methods for measuring the value of natural resource damages are market valuation, restoration or replacement cost, and contingent valuation. 384 Market valuation uses definable markets to measure the worth of a resource. 385 Restoration or replacement costs, as the name implies, measure the resource’s worth by asking how much it would cost to restore the damage. 386 Given that restoration projects can be labor intensive and complicated, these costs can greatly exceed market valuation. 387 Contingent valuation is a controversial method by which value is measured by surveying members of the public to assess how much they would be willing to pay to replace or restore the resource. 388 Regardless of the chosen method, however, there are certain to be controversies given that natural resource damages are unique in many instances and their uses not readily subject to valuation.

V. CONCLUSION

This Article continues the work of scholars who have urged a mended view of private property and others who have described the failings of environmental laws to protect ecosystems as interconnected wholes. Joining the existing dialogue, this Article tackles the uncomfortable and controversial issues of ethical obligations, private property rights, and public ownership: It describes the shared responsibility to avoid the misuse of nature, finding support for this responsibility in literature, religion, culture, science, and law. 389 It explains how the study of ecology has led to more mature views on nature—views that recognize the utility of nature when allowed to function as an interconnected whole. 390 It surveys the scholarship that calls for a similarly mature understanding property—an understanding that

381 See Peck, supra note 156, at 279–86.
382 Id.
383 Id.
384 Id. at 279–85.
385 Id. at 282–83.
386 Id. at 283–84.
387 Id.
388 Id. at 284–85; see also supra note 276 and accompanying text.
389 See supra Part I.B.1.
390 See supra notes 64–66 and accompanying text.
recognizes the legitimate role of public interest in property law. Finally, it highlights how the multitude of environmental laws currently in effect do not protect nature as an interconnected whole.

By bringing these conservations into the same space, this Article sets the stage for a broader vision of natural resource law reform and picks up where others have left off. Turning to natural resource damages law as a touchstone for reform, this Article suggests that the bedrock principles underlying natural damages law are a promising foundation for a reformed legal system that respects broader public interests in how nature is altered and provides a remedy for misuse. The aim of the reformed law is necessarily broad, concerned with misuses of nature on public and private lands alike, as well as misuses that arise from multiple stressors. Though the details of what the reformed law would look like and how it would operate are necessarily left for another day, the challenges enumerated are fundamental considerations for any comprehensive law that seeks to protect ecosystem health across ownership boundaries.

391 See supra Part I.B.2.
392 See supra Part II.
DUE PROCESS DENIED: THE FORGOTTEN CONSTITUTIONAL LIMITS
ON CHOICE OF LAW IN THE ENFORCEMENT OF EMPLOYEE
COVENANTS NOT TO COMPETE

David A. Linehan*

I. INTRODUCTION

Imagine this common workplace scenario: A talented Salesman working for Company A is recruited by Company B to help launch a new product line in direct competition with Company A. When Salesman tenders his resignation, Company A tries to prevent him from going to work for Company B, relying on promises made by Salesman in his employment contract not to work for competitors of Company A in a similar capacity for at least one year after terminating employment. Agreements of this type, and efforts to enforce them, have grown commonplace in many different industries, but are particularly widespread in the financial, high-tech, and life-sciences sectors. And they affect not just salesmen,

* © 2012 David A. Linehan, J.D., University of Texas, 1998; B.B.A., B.A., University of Texas, 1993. David Linehan is a Member of the Seattle law firm of McNaul Ebel Nawrot & Helgren PLLC, where he practices commercial and intellectual property litigation.

1 See generally 3 BRIAN M. MALSBERGER, COVENANTS NOT TO COMPETE: A STATE-BY-STATE SURVEY tbl. C, at C3–C701 (7th ed. 2010) (collecting cases from all United States jurisdictions); see also Katherine V.W. Stone, In the Shadow of Globalization: Changing Firm-Level Employment Practices and Shifting Employment Risks in the United States 17 (UCLA Sch. of Law, Law & Econ. Research Paper Series, Research Paper No. 07-13, 2007), available at ssrn.com/abstract=1023696 (observing that the increasing importance of human capital to firms has made enforcement of covenants not to compete “probably the most frequently litigated issue in the employment area”).


but many types of “knowledge workers,” including engineers, scientists, computer programmers, and management professionals.5

In many states, Company A would have a good chance of enforcing Salesman’s covenant not to compete (“noncompete”).6 So long as the post-termination restrictions are reasonable in duration and geographic scope, and so long as the restrictions appear reasonably necessary to protect Company A’s legitimate business interests, courts in these states would enjoin Salesman from working for Company B if Company A demonstrates that Salesman’s employment with Company B would violate the terms of his noncompete.7 Of course, what constitutes a “reasonable” scope or duration may be interpreted more narrowly in some states than in others.8 But employers can vary the terms of their employment agreements to respect those state-specific differences and thus reasonably ensure that the parties’ expectations at the time of contracting will be enforced if the employee later accepts a job with a cross-town rival.


7 See, e.g., Wood, 438 P.2d at 590–92 (finding under Washington law that noncompetes are a lawful means of protecting legitimate business interests, so long as reasonable in time and geographic scope, but remanding for trial court determination of reasonableness).

8 Georgia and Alabama offer a good comparison of states’ differing views as to what kinds of post-termination restrictions on employment are reasonable and what kinds are not. Compare Corson v. Universal Door Sys., Inc., 596 So. 2d 565, 567–68 (Ala. 1991) (finding a noncompete was reasonable when it limited a former employee’s ability to compete in eight states, three of which the former employer did not do business in), and Fameex, Inc. v. Century Ins. Servs., Inc., 425 So. 2d 1053, 1054 (Ala. 1982) (approving provision that had no territorial limit and precluded solicitation of any of the former employer’s policyholders for three years), with Dent Wizard Int’l Corp. v. Brown, 612 S.E.2d 873, 876 (Ga. Ct. App. 2005) (holding that a noncompete limited to four counties was unreasonable when the former employee had only worked in two or three of them), and Firearms Training Sys. v. Sharp, 445 S.E.2d 538, 540 (Ga. Ct. App. 1994) (noncompete is unreasonably broad and unenforceable when it precludes working in any capacity for the former employer’s competitor).
Now imagine the same scenario, except this time Company A is located in a state that generally recognizes and enforces noncompetes, whereas Company B is located in one of the several states that automatically reject them as unlawful and against public policy. Who prevails? Outcomes can vary widely and, often, unpredictably. But readers may be surprised to learn that under current law, the employee has a decided advantage. As this Article explains below, the outcome of such cases is often determined by which party—the employee or her former employer—wins the proverbial “race to the courthouse.” And because the employee is almost always armed with superior knowledge about the facts and timing of her decision to accept employment with a competitor, the employee nearly always has an outcome-determinative head start in that race. Conventional wisdom holds that employers invariably use their superior leverage and the threat of expensive and protracted noncompete litigation to prevent their employees from seeking more favorable employment opportunities elsewhere. But real-world litigation results call into question this conventional thinking, at least in some states. In fact, when the former employer wins one of these interstate enforcement contests in a state like California, it is often only because the employee or her lawyer (or her new employer’s lawyer) made tactical missteps that effectively relinquished her initial advantage.

This Article explores how and why cases involving interstate enforcement of noncompetes are so often reduced to little more than contests to see which party arrives at the courthouse first. This state of the law becomes inevitable when well-established due process limits on choice of law are overlooked, because states then feel free to act expansively in applying their own law even when the parties have specifically agreed by contract that the law of some other state should govern their

---

9 E.g., California (CAL. BUS. & PROF. CODE § 16600 (West 2008) (declaring all noncompetes void except in narrowly defined circumstances)); North Dakota (N.D. CENT. CODE § 9-08-06 (2006) (similar to California)); and Oklahoma (OKLA. STAT. ANN., tit. 15, §§ 217, 219A (West Supp. 2011) (declaring all contractual, post-employment restrictions to be “void and unenforceable,” although allowing very narrow limitations on direct solicitation of established customers of the former employer)).

10 Compare Application Grp., Inc. v. Hunter Grp., Inc., 72 Cal. Rptr. 2d 73 (Cal. Ct. App. 1998) (finding that California law may apply to enforce a covenant not to compete between an employee who is not a resident of California and an employer based outside of California), with Estee Lauder Cos. v. Batra, 430 F. Supp. 2d 158 (S.D.N.Y. 2006) (finding that New York law applies as stated in the agreement even though it is contrary to California public policy).


relationship. Lax attention by courts to employers’ due process rights, combined with a robust application of full faith and credit to final state-court judgments, effectively enables states with highly restrictive rules against noncompetes—for example, California and Georgia—to export their employee-favoring policies nationwide.

This Article presumes that covenants not to compete are neither desirable nor undesirable. It takes no position on whether noncompetes should, as a matter of public policy, be broadly permitted, tightly restricted, or banned altogether. Different states have reached different conclusions on this question, as have many commentators, and the purpose of this Article is not to convince the reader which side has the better argument. Rather, this Article assumes that states have valid justifications for their policy preferences and that, unless Congress enacts preemptive federal legislation, those preferences deserve respect.

Instead, this Article focuses on distortions in the balance of power among the states. These distortions result from courts failing to respect due process constraints on their power to prefer their own laws to those of sister states. Accordingly, this Article (i) revisits controlling Supreme Court cases establishing due process limits on choice of law; (ii) illustrates how a lack of fidelity to these precedents by courts in some states has tended to unfairly favor the departing employee and the interests of states with the most restrictive views on noncompetes; and (iii) explains how a renewed emphasis on due process considerations would allow enforcement actions to be decided fairly on their merits. Unless we assume that the interests of states like Massachusetts, New York, and Ohio (which generally permit reasonable noncompetes) are illegitimate or inferior to the employee-mobility-favoring interests of states like California and Georgia, then it is in our national interest to ensure that disputes between employees and their former employers are determined on their merits and in accordance with longstanding constitutional principles, rather than by a race to the top of the courthouse steps.

This is not the first article to point out difficulties that arise when an employer tries to enforce a noncompete against an employee who relocates to California or any other state that holds restrictive views on such agreements. However, it is the

---

13 E.g., Rachel S. Arnow-Richman, Bargaining for Loyalty in the Information Age: A Reconsideration of the Role of Substantive Fairness in Enforcing Employee Noncompetes, 80 OR. L. REV. 1163 (2001) (“propos[ing] an alternative theory of enforcement focusing on substantive and procedural issues that arise at the time noncompetes are formed”); Estlund, supra note 11 (providing a critical assessment of current law governing noncompete agreements and arbitration agreements); Ronald J. Gilson, The Legal Infrastructure of High Technology Industrial Districts: Silicon Valley, Route 128, and Covenants Not to Compete, 74 N.Y.U. L. REV. 575, 597–600 (1999) (examining the role of the Uniform Trade Secrets Act (UTA) in noncompete disputes); Brandon S. Long, Note, Protecting Employer Investment in Training: Noncompetes vs. Repayment Agreements, 54 DUKE L.J. 1295 (2005) (arguing “that repayment agreements are a more sensible alternative to traditional noncompetes”).

14 See Sheri Wardwell, Invalidity of Covenants Not to Compete in California Affects Employers Nationwide, 5 SHIDLER J.L. COM. & TECH. 22, ¶¶ 10–16 (2009), available at
first article to reveal that this problem—and hence the solution—relates to courts’
inattention to due process limits on choice of law. It is also the first article to
demonstrate that, due to informational asymmetries, employees and their new
employers have, and will continue to have, a decided advantage in such
proceedings unless courts reinstate proper due process review.

Part II of this Article, describes the most common types of employee
covenants not to compete and the interests they serve. It also summarizes the
differing attitudes that various states have adopted toward the validity and
enforceability of such covenants. Extreme variability in state laws governing
noncompetition covenants underscores the importance that choice-of-law
determinations take on when courts are presented with cases involving parties from
(and restrictions on conduct in) more than one state.

Part III examines key Supreme Court precedents defining the constitutional
boundaries of state courts’ authority to choose the law to govern a particular case. My
analysis includes a fresh look at the seminal due process cases of Allstate
their holdings might affect choice of law in multi-state noncompete disputes.
Hague and Shutts, as many scholars have recognized, provide very modest
restrictions on choice of law.17 Yet both cases make it clear that due process
remains a constraint on choice of law, and that respect for parties’ legitimate
expectations as to what law will apply to their conduct is of paramount importance.
Part III also demonstrates that over time the Full Faith and Credit Clause has lost
its significance as an independent constraint on choice of law, effectively
becoming redundant to the protections of due process. This diminution in the
respect that one state owes to the substantive laws of a sister state contrasts with
the absolute duty of states to recognize the valid judgments of sister states. Thus,
under current doctrine, the Due Process Clause of the Constitution gives states
broad latitude in choosing whether or not to apply the law of another state, yet they
are constitutionally compelled to recognize a valid sister-state judgment even if
they would regard the judgment as repugnant to their own laws.

Part IV shows that even the minimal due process limitations on choice of law
that survive today have largely been ignored in states that are hostile to
noncompetes. Such courts tend to expansively apply their own restrictive rules
against noncompetes to virtually any dispute tried within their borders, regardless
of party expectations to the contrary. Here, I examine recent examples of litigation

http://lctjournal.washington.edu/vol5/a22wardwell.html; Christina L. Wu, Comment,
Noncompete Agreements in California: Should California Courts Uphold Choice of Law
17 RUSSELL J. WEINTRAUB, COMMENTARY ON THE CONFLICT OF LAWS 669, 669–72
(6th ed. 2010) (positing that a choice of law is likely to run afoul of due process standards
in only two “relatively narrow” circumstances); cf. Louise Weinberg, Choice of Law and
law decisions ought to receive only minimal constitutional scrutiny after Hague).
that ensued when employees quit their jobs to accept positions with competitors located in states with restrictive laws against noncompetes. These cases vividly illustrate the unfairness that can result when courts improperly discount the parties’ reasonable expectations regarding the enforceability of their agreements. Time and again, courts have invalidated noncompetes that were valid and enforceable in the state where the employee lived and worked until the day the lawsuit was filed. These courts have justified such results by relying on the employee’s newly acquired residence in the forum state—even though the Supreme Court has repeatedly stated that a plaintiff’s move to the forum state before filing suit cannot support application of forum law. To the extent these courts have even thought about due process, they have mistakenly viewed *Hague* as a license to apply forum law any time they can articulate some nondisingenuous basis for doing so. Even contractual choice-of-law clauses have been no obstacle to courts intent on applying strong forum-state policies against noncompete agreements.\(^1\)

Transforming interstate noncompete litigation from a race to the courthouse into a fair adjudication on the merits that comports with constitutional principles requires a renewed recognition of the limitations that due process imposes on a state’s choice of law. To that end, Part V of this Article offers a new, two-step approach to choice-of-law decisions in multistate noncompete cases. In step one, courts would apply due process scrutiny as a filter to exclude application of any state’s law that would be fundamentally unfair to either party. Having thus excluded all constitutionally impermissible choices, the court would then proceed in step two to apply the forum’s preferred choice-of-law methodology, whatever it may be. It may seem odd to describe this as a “new” approach, given that it results from nothing more than a straightforward application of *Hague* and *Shatts*. But some courts have strayed so far from what the Supreme Court envisioned in those cases that my two-step approach may well appear to be novel to them. Part V also explains how this two-step approach could incentivize desirable behavior in employers and employees, lead to better-drafted noncompetes, and generally inspire greater confidence in the fairness of the judicial process in multistate noncompete cases.

II. NONCOMPETES AND VARYING STATE POLICIES TOWARD THEM

A. Types of Noncompetes and the Interests They Serve

Post-employment covenants not to compete, or “noncompetes,” typically prevent the employee from working for certain types of firms or holding certain types of positions (or both) for some definite period of time after leaving employment.\(^2\) A noncompete might preclude the employee from working for competitors or soliciting customers of the former employer within a specified

---

18 *E.g.*, Convergys Corp. v. Keener, 582 S.E.2d 84, 85–86 (Ga. 2003).

geographic region. Depending on the nature of the employer’s business or the type of position the employee held, the employee might be precluded from working for any competitor nationwide, or even worldwide, for a specified length of time. Another common variant of the noncompete is one where the employee is precluded only from holding certain types of positions with the employer’s competitors or from soliciting business from a certain class of customers or clients. But in all cases, the employer seeks to protect its investment by forbidding the employee from taking the proprietary knowledge or specialized training that have been imparted to him and putting it to use unfairly for the benefit of competitors.

Noncompetes that appear designed merely to insulate the employer from “ordinary” competition in the marketplace are generally unenforceable. By contrast, noncompetes that prevent “unfair competition” are enforced in most states. In this context, “unfair competition” often means that a company is seeking to take unfair advantage of a competitor’s investments—such as investments in product development or promotion, employee training, scientific research, or customer relationships. For example, rather than developing its own

21 E.g., Nordson Corp. v. Plasschaert, 674 F.2d 1371, 1377 (11th Cir. 1982) (enforcing noncompete that extended to all competitors in the United States, Canada, and Western Europe, where the employee had already shared his former employer’s confidential information to further his European operations).
23 See, e.g., United States v. Addyston Pipe & Steel Co., 85 F. 271, 281 (6th Cir. 1898), aff’d, 175 U.S. 211 (1899) (Judge Taft, noting that covenants not to compete were enforceable under common law if reasonably necessary to protect the employer “from the danger of loss to the employer’s business caused by the unjust use on the part of the employ[ee] of the confidential knowledge acquired in such business”); see also Reed, Roberts Assocs., Inc. v. Strauman, 353 N.E.2d 590, 593 (N.Y. 1976) (acknowledging an employer’s “legitimate interest . . . in safeguarding that which has made his business successful” and protecting against “deliberate surreptitious commercial piracy”).
25 E.g., Hasty, 671 S.W.2d at 473–74 (distinguishing “ordinary” from “unfair” competition).
successful products or marketing strategies, a company who competes unfairly might instead recruit one of its competitor’s key employees—who may be deeply familiar with that competitor’s confidential business plans and strategies or who may possess close relationships with that competitor’s best customers—to help it accelerate a new product launch. Likewise, a pharmaceutical company might seek to speed the development of a new drug treatment by recruiting a competitor’s research scientists and thereby obtain access to the competitor’s proprietary research. By hiring a high-value employee away from a rival firm, a company theoretically can “free ride” on the knowledge, training, and investments of its rival, thereby allowing the company to compete more effectively without assuming the risk and expense of independent research and without making its own investments in product development and promotion.27

Business interests that have been recognized by courts as adequate to support a noncompete are varied but limited.28 They include: proprietary scientific or technical information,29 specialized training programs for employees,30 specific customer relationships or customer lists,31 confidential plans for new products or markets,32 promotional strategies and investments,33 and confidential marketing data.34

---


28 Collin, supra note 22, at 417.


33 E.g., Beckman v. Cox Broad. Corp. 296 S.E.2d 566, 569 (Ga. 1982) (television station’s investment in the promotion of a television personality’s on-air image justified enforcement of noncompete to prevent him from taking that personality to competing
Employers have successfully sued many different types of employees for breach of their noncompetition covenants, but it is possible to lump most such employees into the broad category of “knowledge workers.” After all, it takes time and costs money for firms to develop useful business knowledge, and it likewise takes time and costs money to impart that knowledge to its employees. It should therefore come as no surprise that employers use noncompetition covenants most often with employees who are tasked with developing, learning, safeguarding, and exploiting the employer’s valuable business information. In our twenty-first century economy, many companies’ most valuable assets go down the elevator and exit the building every night. It is little wonder that noncompete agreements are so prevalent and so commonly the subject of litigation.

Disputes concerning enforcement of noncompetes have increasingly graced the headlines of business newspapers and websites in recent years, as high profile executives move from one company to another. These events, in turn, have sparked debate over the legitimacy of agreements purporting to limit an employee’s post-termination employment options and, on the other hand, the danger that the former employer’s proprietary information will be unfairly exploited by the employee and his new employer.


36 See supra notes 29, 32, 34.

Some skeptics of noncompetes emphasize the draconian consequences to the employee that may result if he is forced to adhere to a broad noncompete. But many courts are sensitive to this concern and will sometimes modify the terms of the noncompete to reach a more palatable balance between the interests of the employee in earning a living in his chosen profession and the interests of his former employer in protecting its investments. Other commentators opine that noncompetes are inconsistent with our information-age economy, asserting that noncompetes slow or inhibit the transfer of knowledge among innovative firms and thereby pose risks to innovation.

Defenders of noncompetes contend they are necessary to protect businesses from unfair competition and to enable them to recoup their investments in research and human capital. This protection, they argue, fosters an environment where companies feel incentivized to innovate. As noted, this Article does not attempt to establish which faction has the better policy view. Instead, it argues that a court may not always have the constitutional authority to dismiss either viewpoint out of hand merely because it is contrary to the preferred viewpoint of the state in which it sits.

B. State Regulation of Noncompetes

States have adopted policies toward noncompetes that can be classified as (1) permissive, (2) skeptical, or (3) openly hostile.

1. Permissive States

A majority of states generally approve the use of noncompetes, subject to certain restrictions to ensure reasonableness. Washington law, for example, permits a wide range of post-termination employment restrictions, as does the law

---


40 Garrison & Wendt, supra note 38, at 112; Gilson, supra note 13, at 597–600.


42 See, e.g., Callahan, supra note 41, at 715 (“To the extent that inventors are prevented from reaping the benefits of the information they develop, they are discouraged from engaging in costly research and development, and competition will suffer because fewer products will be produced.”).

43 See supra note 13 and accompanying text.

44 See Wardwell, supra note 14, ¶ 1 & n.2.

of Massachusetts, \(^{46}\) Minnesota, \(^{47}\) and many other states. \(^{48}\) Courts in such states will scrutinize noncompetes only to ensure that they address some legitimate business interest and are not overly severe in terms of duration or scope. \(^{49}\) Generally speaking, any legitimate business interest will support a noncompete, so long as it is supported by adequate consideration and so long as the restrictions are not draconian. \(^{50}\) As a general rule, the longer the duration of the noncompete and the broader the geographic scope, the more skeptical a court will be. \(^{51}\) Conversely, the narrower the covenant is drawn, the more likely a court will be to enforce it. But the exact boundaries of what is or is not reasonable are quite often determined by the realities of the particular business in question. \(^{52}\)

2. Skeptical States

Other states are more skeptical of noncompetes, enforcing them only begrudgingly when the evidence of their need is compelling and the restrictions imposed on the employee are no broader than truly necessary. Georgia, for example, employs a “strict scrutiny” approach to enforcement of noncompetes. \(^{53}\) Texas follows a similar (though now somewhat less restrictive) approach, requiring that the post-employment restrictions be narrowly tailored to meet a specific business interest. \(^{54}\) Indeed, a Texas employer cannot enforce a noncompete unless


\(^{48}\) E.g., Borg-Warner Protective Servs. Corp. v. Guardsmark, Inc., 946 F. Supp. 495, 501 n.6 (E.D. Ky. 1996) (finding covenants reasonable under both Kentucky and Tennessee law); Raimonde v. Van Vlerah, 325 N.E.2d 544, 547 (Ohio 1975) (noncompetes are to be enforced [under Ohio law] to the extent necessary to protect the former employer’s legitimate interests); see also Wardwell, supra note 14, ¶ 1 n.2.


\(^{50}\) See Allen v. Rose Park Pharmacy, 237 P.2d 823, 826 (Utah 1951) (“as long as the restrictions as to time and space are reasonably necessary to the protection of the business and the hardship features of the case do not constitute equitable grounds for rescission, or call for the intervention of other rules of equitable relief, then the court is powerless to relieve a party from the effects of his contract.”).

\(^{51}\) E.g., Wood, 438 P.2d at 591–92.

\(^{52}\) Collin, supra note 22, at 420.


\(^{54}\) See TEX. BUS. & COMM. CODE § 15.50(a) (West 2002); Alex Sheshunoff Mgmt. Servs., L.P. v. Johnson, 209 S.W.3d 644, 648, 656–58 (Tex. 2006); Estlund, supra note 11, at 395 n.46 (explaining that under Texas law, noncompetes were traditionally viewed as presumptively invalid but are now somewhat easier to enforce following the Alex
the employer first demonstrates that it has in fact shared confidential or competitively sensitive business information with the employee so as to justify the post-termination restriction on his employment. Until the employer actually entrusts the employee with such information, the noncompete remains an unenforceable, unilateral promise with no consideration. In other words, the furnishing of confidential business information or training creates the consideration that transforms the unilateral promise into a binding contract not to work for a competitor.

3. **Openly Hostile States**

Finally, a small minority of states are openly hostile to noncompetes and have declared them invalid by statute. California, Colorado, North Dakota, and

---

_Shesunoff_ decision, provided that the employer demonstrates “compelling” interests and a good “fit” between the noncompete’s restrictions and those interests.

55 _Alex Sheshunoff_, 209 S.W.3d at 650–51.
56 _Id._ at 655.
57 _Id._
58 CAL. BUS. & PROF. CODE § 16600 (West 2008); _see also_ Edwards v. Arthur Anderson LLP, 189 P.3d 285, 298 (Cal. 2008) (“The majority agrees with the Court of Appeal that the noncompetition agreement was invalid under Business and Professions Code section 16600”); Bosley Med. Grp. v. Abramson, 207 Cal. Rptr. 477, 480 (Ct. App. 1984) (admonishing that California courts are to interpret the statutory prohibition on noncompetes as broadly as its language reads; “[a]t least since 1872, a non-competition agreement has been void unless specifically authorized by [other sections]”). California, like all states, permits restrictive covenants when necessary to safeguard the former employer’s trade secrets. Readylink Healthcare v. Cotton, 24 Cal. Rptr. 3d 720, 727–28 (Ct. App. 2005).
59 Colorado law declares that “any covenant not to compete which restricts the right of any person to receive compensation for performance of skilled or unskilled labor for any employer shall be void . . . .” COL. REV. STAT. § 8-2-113(2) (2010). This statute provides an exception for recovery of expenses incurred for educating or training an employee who serves for less than two years. _Id._ § 8-2-113(2)(c). The statute also permits noncompetes for use with “[e]xecutive and management personnel and officers and employees who constitute professional staff to executive and management personnel.” _Id._ § 8-2-113(2)(d). But Colorado courts have taken a very narrow view of what kinds of employees qualify as “executive and management personnel” or “professional staff to executive and management personnel.” _See_ Gold Messenger, Inc. v. McGuay, 937 P.2d 907, 910 (Colo. Ct. App. 1997) (exceptions to general rule of voidness are narrowly construed); Mgmt. Recruiters of Boulder v. Miller, 762 P.2d 763, 764 (Colo. Ct. App. 1988) (holding that headhunter account executive was not within statutory exception); Atmel Corp. v. Vitesse Semiconductor Corp., 30 P.3d 789, 795 (Colo. Ct. App. 2001) (technical liaison did not fit statutory exception where he supervised no employees and had three levels of management above him). Colorado courts, like those of all states, will enforce reasonable post-employment restrictions that protect actual trade secrets. _COLO. REV. STAT._ § 8-2-113(2)(b).
Oklahoma categorically reject most or all employment-related noncompetes as contrary to public policy and unenforceable except in very narrow circumstances defined by statute. These openly hostile states certainly present obvious concerns for employers who seek to enforce noncompetes. But as discussed in Part IV, even a state like Georgia that does not categorically prohibit noncompetes can still pose a major enforcement challenge if it applies its law expansively to govern employment relationships that emanate from more permissive states.

III. THE EVOLUTION OF CONSTITUTIONAL CONSTRAINTS ON CHOICE OF LAW

To understand why the interstate enforcement of noncompetes often devolves into a mere race to the courthouse, we must first look back on key historical developments in the related fields of conflict of laws and enforcement of judgments. Four related factors explain the difficulties that employers now face in enforcing noncompetes when their former employees relocate across state lines: (a) the establishment of an “ironclad rule of full faith and credit” owed to valid state court judgments; (b) the collapse of full faith and credit as a stand-alone limitation on a court’s choice of law; (c) the blending of full-faith-and-credit principles with due process to create the modern constitutional standard for choice of law; and (d) the underappreciation of due process concerns (and overreliance on local policy preferences) by many courts in determining whether it is permissible to apply forum-state law. The first three of these factors are discussed here in Part III, which leads to a discussion of the fourth factor in Part IV.

---


61 Under statutory provisions added in 2001, Oklahoma law now declares void all contracts that restrain anyone from exercising a lawful profession, trade, or business, except in three narrow circumstances. See OKLA. STAT. ANN. tit. 15, §§ 217–219A (West 2011 & Supp. 2012). Those exceptions are for (1) covenants made in connection with the sale of a business, § 218; (2) covenants made by partners in connection with dissolution of the partnership, § 219; and (3) covenants by an employee not to “directly solicit the sale of goods, services or a combination of goods and services from the established customers of the former employer,” § 219A. Thus, Oklahoma law is quite similar to California and North Dakota, except Oklahoma law permits employment contracts that preclude the employee from “directly” soliciting the “established” customers of his former employer.
A. The Development of Ironclad Full Faith and Credit for Valid Judgments

The Full Faith and Credit Clause\(^62\) is one of several constitutional provisions intended to forge a federal nation out of individual sovereign states.\(^63\) As Justice Stone described it, the purpose of the Full Faith and Credit Clause was to

alter the status of the several states as independent foreign sovereignties, each free to ignore the obligations created under the laws or by the judicial proceedings of the others, and to make them integral parts of a single nation throughout which a remedy upon a just obligation might be demanded as of right, irrespective of the state of its origin.\(^64\)

Thus, the Full Faith and Credit Clause prevents one state from ignoring the public acts, records, or judicial proceedings of another state, or subjecting them to the “gantlet of local ‘public policy,’” as it may freely do when dealing with a foreign sovereign’s public acts, records, or judicial proceedings.\(^65\)

A major impetus for adopting the Full Faith and Credit Clause at the Constitutional Convention was to overcome the problem of migratory debtors, who moved from one colony to another to avoid collection of debts.\(^66\) Creditors wanted to be certain that a judgment obtained against a debtor in one jurisdiction could be enforced in other jurisdictions if the debtor relocated in an effort to avoid payment.\(^67\) Variations in state law governing gambling or usury had made it highly unpredictable whether a debt that was valid in the state where it was created would

---

\(^62\) U.S. CONST. art. IV, § 1. The clause provides: “Full Faith and Credit shall be given in each State to the public Acts, Records, and judicial Proceedings of every other State. And the Congress may by general Laws prescribe the Manner in which such Acts, Records, and Proceedings shall be proved, and the Effect thereof.” Id. Acting upon the authority granted by the Full Faith and Credit Clause, Congress passed an implementing statute in 1790. This statute provides that the acts, records, and judicial proceedings of one state “shall have such faith and credit given to them in every court of the United States, as they have by law or usage in the courts of the State from when the said records are, or shall be taken.” 28 U.S.C. § 687 (1940) (current version at 28 U.S.C. § 1738 (2006)). Originally, the quoted portion of the implementing statute expressly mentioned only “records” and “judicial proceedings,” but Congress amended the language in 1948 to include the “acts” of a state as among the enumerated items entitled to full faith and credit in every other state. Id. For further debate on whether the omission and subsequent addition of “acts” to the language of the statute should be viewed as significant for its interpretation, see Carroll v. Lanza, 349 U.S. 408, 422 (1955) (Frankfurter, J., dissenting); Hughes v. Fetter, 341 U.S. 609, 613–14 n.16 (1951); see also WEINTRAUB, supra note 17, at 715 & n.280.

\(^63\) See WEINTRAUB, supra note 17, at 700.


\(^65\) WEINTRAUB, supra note 17, at 700.


\(^67\) See id. at 1–3.
be enforced in a state that regarded such debts as offensive to local law or policy.\(^{68}\) In some respects, this concern with migratory debtors mirrors the concern of employers today when enforcing noncompetes: Employers seek to ensure that highly mobile employees do not evade their contractual commitments by taking a forbidden job in another state where noncompetes are either subject to very strict limitations or are void entirely as against public policy.

The three prongs of the Full Faith and Credit Clause—“public Acts,” “Records,” and “judicial Proceedings”—each have their own body of case law interpreting them. This Article discusses only the first and third prongs of the Clause.\(^{69}\)

Some have argued that early case law interpreting the Full Faith and Credit Clause does not support the notion that its reference to “public Acts” was intended to affect choice-of-law outcomes.\(^{70}\) As discussed further below, courts still have not completely resolved the question of how much deference the Full Faith and Credit Clause requires one state to give to the “public Acts”—that is, statutes or laws—of another state, particularly when a state court is asked to apply the law of a state with very different laws than its own.\(^{71}\) One example of this uncertainty is seen in Hughes v. Fetter,\(^{72}\) in which the Supreme Court described the tension between, on the one hand, “the strong unifying principle embodied in the Full Faith and Credit Clause looking toward maximum enforcement in each state of the obligations or rights created or recognized by the statutes of sister states,” and on the other hand, “the policy of Wisconsin, as interpreted by its highest court, against permitting Wisconsin courts to entertain” claims that are not recognized under Wisconsin law, even if cognizable under the law of the state where the injury occurred.\(^{73}\)

By contrast, it has long been settled that the “judicial proceedings” of one state are entitled to virtually unwavering enforcement in the courts of every other state, even if the enforcing state regards the judgment as offensive to its local law or policy.\(^{74}\) Thus, with respect to judicial proceedings, some commentators have

---

\(^{68}\) See id. at 2, 102–03 (discussing a 1908 case in which a Mississippi court refused to give full faith and credit to a Missouri judgment regarding a futures contract because Mississippi law viewed futures contracts as a form of gambling).

\(^{69}\) The “public records” aspect of the Full Faith and Credit Clause has, for the most part, been rendered moot, or at least uncontroversial, by modern rules of civil procedure and evidence, which now generally govern issues of authentication and admissibility of legal records from other states. Id. at 13–18.

\(^{70}\) Id. at 10.

\(^{71}\) See infra notes 76–81, 88–109 and accompanying text.

\(^{72}\) 341 U.S. 609, 612 (1951).

\(^{73}\) Id. at 612 & n.10, 613–14.

\(^{74}\) See Fauntleroy v. Lum, 210 U.S. 230, 237 (1908) (holding that there is no public policy exception to the enforcement of valid judgments); see also Miller v. Kingsley, 230 N.W.2d 472, 474–75 (Neb. 1975) (holding that the court must give full faith and credit to a sister state’s judgment for punitive damages even though a court in Nebraska could not award punitive damages).
described the Full Faith and Credit Clause as establishing an “iron law” of preclusion, meaning that the matters decided in a final judgment from one court cannot again be litigated in another court, except in rare circumstances. 75 This “rigid” rule requires the forum to recognize and enforce the out-of-state judgment regardless of whether the underlying claim contravenes the “basic public policy” of the forum,76 or whether the judgment resulted from the rendering court’s misapplication of its own law.77 This ironclad rule of full faith and credit for state court judgments is one of the factors that motivates disputants to rush to file suit in an effort to be the first to obtain final judgment, as that final judgment will then be entitled to nationwide enforcement.

In theory, merely being the first to file suit should not confer full-faith-and-credit advantages because only final judgments are entitled to nationwide enforcement.78 But in practice, the party who files first will also typically be the party first to obtain final judgment. This is true in both federal and state courts.

In federal courts, the “first-filed” rule creates a strong, common-law presumption that the court “initially seized of a controversy” should be the court that adjudicates the claims.79 Accordingly, the forum where the second-filed suit was brought will often transfer the case to the first-filed forum, or stay the second-filed case pending resolution of the first case.80 Further, in a declaratory judgment action, federal courts retain discretion whether to hear the case at all,81 and a court may exercise that discretion to defer ruling on a matter that is already pending in another forum.82

Similarly in state courts, an employee who files first may be able to obtain a temporary restraining order—possibly even on an ex parte basis—enjoining the employer from seeking to enforce the noncompete in another forum.83 The employer would thus risk contempt-of-court sanctions in the first-filed forum if it filed a second suit elsewhere seeking to establish the validity of the noncompete. Consequently, although full faith and credit does not attach until a court renders a final judgment, as a practical matter the party who reaches the courthouse first has a tremendous leg up on reaching final judgment ahead of any other court where a second suit might be commenced.

75 Reynolds & Richman, supra note 66, at 6, 71.
76 Id. at 44–45.
77 Fauntleroy, 210 U.S. at 237.
79 Merrill Lynch, Pierce, Fenner & Smith, Inc. v. Hayden, 675 F.2d 1169, 1174 (11th Cir. 1982).
80 See Alltrade, Inc. v. Uniweld Prod., Inc., 946 F.2d 622, 623, 628–29 (9th Cir. 1991) (noting that the first-to-file rule “allows a district court to transfer, stay, or dismiss an action when a similar complaint has already been filed in another federal court”).
B. The Erosion of Full Faith and Credit as an Independent Limit on Choice of Law

It is unclear how much, if at all, the Full Faith and Credit Clause and its implementing statute originally were intended to constrain a court’s power to choose which state’s law should govern a particular dispute.\(^\text{84}\) Since passage of the implementing statute in 1790, Congress has rarely exercised its constitutionally granted power to prescribe the effects to be given to state statutes when courts in another state are asked to apply them.\(^\text{85}\) As a result, state courts have been left to their own devices, interpreting the full faith and credit mandate on a case-by-case basis, with choice-of-law results that are often difficult to reconcile with each other.\(^\text{86}\) Some commentators assert that the notion that principles of full faith and credit should apply to choice-of-law decisions was an idea that did not surface until the late nineteenth century.\(^\text{87}\) They argue that this idea had a “few moments in the sun” in the early twentieth century but lingers only in very abbreviated form today.\(^\text{88}\)

This argument is surely correct. With two exceptions—\textit{Bradford Electric Light Co. v. Clapper},\(^\text{89}\) and \textit{Yates v. John Hancock Mutual Life Insurance Co.}\(^\text{90}\)—it seems that courts have viewed the Due Process Clause, not the Full Faith and Credit Clause, as the primary constitutional limitation on choice of law.\(^\text{91}\)

In \textit{Bradford}, the Court held that a New Hampshire court violated full faith and credit by applying New Hampshire’s, rather than Vermont’s, workers compensation law to the claim of a deceased worker.\(^\text{92}\) Although the worker was

\(^{84}\) REYNOLDS & RICHMAN, supra note 66, at 10.

\(^{85}\) One of the few examples is the Defense of Marriage Act, codified as part of the Full Faith & Credit Clause’s implementing statute, preventing any State from being required to “give effect to any public act, record, or judicial proceeding of any other State, territory, possession, or tribe respecting a relationship between persons of the same sex that is treated as a marriage under the laws of such other state, territory, possession, or tribe.” 28 U.S.C. § 1738C (2006). Contrast § 1738B, requiring State courts to “enforce according to its terms a child support order made consistently with this section by the Court of another state.”

\(^{86}\) REYONDS & RICHMAN, supra note 66, at xviii.

\(^{87}\) Id. at 10.

\(^{88}\) Id.

\(^{89}\) 286 U.S. 145 (1932).

\(^{90}\) 299 U.S. 178 (1936).

\(^{91}\) A third exception can be found in a line of Supreme Court cases involving the now-quaint realm of fraternal benefit societies and the death benefits payable under their membership agreements. See WEINTRAUB, supra note 17, at 697–702 (discussing various cases from the early- to mid-twentieth century, where full faith and credit required the forum to apply the law selected in the members’ agreements, rather than forum law, even where the forum had some significant relationship to the plaintiff and the agreement). It is open to debate whether these cases are still good law, but their relevance today certainly seems dubious.

\(^{92}\) Bradford, 286 U.S. at 159–62.
killed while on the job in New Hampshire, he was a Vermont resident working for a Vermont utility company. And his employment was covered by the Vermont workers compensation statute, which provided that statutory compensation was the exclusive remedy against the employer for injuries occurring anywhere in the course of employment. By contrast, New Hampshire’s workers compensation law allowed injured workers to choose between accepting statutory compensation or suing their employer. Justice Brandeis’s majority opinion held that the Vermont statute was a “public act” under the Full Faith and Credit Clause and that New Hampshire had no interest in permitting the worker to sue under its statute when the worker was only in the state temporarily and had no dependents there that required financial support. The Court held that New Hampshire courts would not cause prejudice to the interests of New Hampshire citizens if it recognized the Vermont statute. “[T]he effectiveness of the Vermont act,” however, would have been “gravely impaired” if a Vermont resident had been allowed to sue under New Hampshire law.

Subsequently, in *Yates*, the Court unanimously held that it was unreasonable, and therefore unconstitutional, for a Georgia court to apply Georgia law to preclude a New York insurance company from relying on a New York statute to cancel an insurance policy that had been issued to a New York resident in New York. The mere fact that the beneficiary of the policy had moved to Georgia following the insured’s death did not give Georgia a sufficient interest to justify application of Georgia law because, in Justice Brandeis’s words, “there was no occurrence, nothing done, to which the law of Georgia could apply.” The New York statute permitted insurers to cancel policies for certain misrepresentations during the application process, whereas Georgia law did not. The Court held that allowing Georgia to apply its contrary law would have denied full faith and credit to New York law. Interestingly, *Yates* appears to be a somewhat muddled assertion of full-faith-and-credit principles, as the decision explicitly relies on *Home Insurance Co. v. Dick*, which, as noted below, addressed due process limits on choice of law, not full faith and credit. *Yates* therefore marks the Court’s first significant step toward the eventual consolidation of full faith and credit with due process principles to form a single, constitutional limitation on choice of law.

---

93 *Id.* at 152.
94 *Id.*
95 *Id.* at 162.
96 *Id.* at 161.
97 *Id.* at 159
99 *Id.* at 179.
100 *Id.* at 182.
101 *Id.* at 181–82.
102 *Id.* at 183.
Having enjoyed a brief moment in the sun\footnote{REYNOLDS \& RICHMAN, supra note 66, at 10.} in Bradford, and to a lesser extent Yates, full faith and credit then began to retreat from its status as an independent limitation on choice of law. In a series of workers compensation decisions from the 1930s to 1950s, the Supreme Court consistently upheld state courts’ applications of forum law (usually favoring the injured worker) against full faith and credit challenges.\footnote{See Carroll v. Lanza 349 U.S. 408 (1955) (upholding an Arkansas court’s decision to apply Arkansas’ worker’s compensation law over Missouri’s); Cardillo v. Liberty Mut. Ins. Co., 330 U.S. 469, 475–76 (1947) (upholding a District of Columbia court’s decision to apply the District of Columbia worker’s compensation law over Virginia’s); Alaska Packers Ass’n v. Indus. Accident Comm’n, 294 U.S. 532, 544–50 (1935) (upholding a California court’s decision to apply California’s worker’s compensation law over Alaska’s).} The Supreme Court repeatedly permitted the forum state to apply its own laws regarding compensation for injured workers, even when (a) the employee’s contract specified that the employee’s rights were governed by the law of another state and (b) the injury occurred in a state that had enacted a statute that made recovery under its laws the exclusive means of compensating injured workers.\footnote{See Alaska Packers, 294 U.S. at 538–39.} With each of these cases, the Supreme Court’s full-faith-and-credit analysis focused less and less on weighing the interests of one state against the other, and instead focused on whether the forum state had a sufficient aggregation of contacts with the persons or events in dispute to give the forum state a legitimate interest in applying its own law.\footnote{See Carroll, 349 U.S. at 413 (determining that Arkansas, as the place of injury, likely had an important interest in resolution of the case that allowed it to apply Arkansas law, “even though . . . Carroll’s injury may have cast no burden on [Arkansas’s] institutions”). In Carroll, the plaintiff was a Missouri resident, employed by a Missouri contractor, under a contract entered into in Missouri, who was injured while performing a subcontract in Arkansas, although he was immediately hospitalized in Missouri. Id. at 409.}

In short, the Court got out of the business of reviewing choice-of-law decisions for violations of full faith and credit. Any rational, nondisingenuous basis for claiming a state interest in the outcome of the case would henceforth support a court’s election to apply its local law. As Professor Weintraub has noted, “[i]t was becoming increasingly difficult to distinguish the requirements of full faith and credit from those of due process,” and at least in the field of workers compensation, the distinction had been all but obliterated.\footnote{WEINTRAUB, supra note 17, at 696.} With full faith and credit having been marginalized as an effective, independent restraint on choice of law, due process became the primary consideration in determining the constitutionality of a court’s choice-of-law decision.
C. The Modern Test for Constitutionality: Due Process with Lip Service to Full Faith and Credit

The rise of due process as the primary constitutional mechanism for protecting litigants from unreasonable applications of forum law, is best illustrated by a line of cases beginning with *Home Insurance Co. v. Dick* and ending with *Phillips Petroleum Co. v. Schutts*.

1. Early Precedents Establishing Due Process Limits on Choice of Law

In *Home Insurance Co. v. Dick*, the Supreme Court held that due process prohibited a Texas court from applying Texas’s statute of limitations to an action brought by a Texas citizen against a Mexican insurer. The Texas plaintiff sought to recover insurance proceeds under a policy that was originally issued in Mexico to a Mexican resident and that covered a loss to a tugboat in Mexican waters.109 The policy contained a one-year limitations period for bringing suit on the policy, and plaintiff’s suit was filed outside that timeframe.110 Texas law, however, nullified any contractual provisions that purported to shorten the time to sue on the contract to less than two years.111 The Court held that applying Texas law would violate due process because the only relevant contact Texas had with the dispute was that the owner of the vessel was then a Texas resident.112 Notably, the Court did not deny that Texas had some interest in the outcome; obviously, Texas had a legitimate state interest in providing its residents a judicial forum in which to recover insurance proceeds from foreign insurers. But the Court held that the plaintiff’s residence in Texas was not, in and of itself, sufficient to make it reasonable for Texas to assert the interest that it had.113

Although not explicitly stated, it appears that unfair surprise to the insurer was a primary concern for the *Dick* Court, as there is no indication in the Court’s discussion that the insurer had any reason to believe that ownership of the boat (and thus the policy) would later fall into the hands of a Texas resident.114 From the Mexican insurer’s perspective, it had issued an insurance policy to a Mexican company, covering a Mexican boat that sustained losses while in Mexican waters.115 Given these contacts with Mexico, the subsequent sale of the boat to a

---

109 *Dick*, 281 U.S. at 402–03.
110 *Id.* at 403–04.
111 *Id.* at 404–05.
112 *Id.* at 407–08.
113 *Id.* at 408–10.
114 *See id.* at 408 (acknowledging that “[a]ll acts relating to the making of the policy were done in Mexico,” all those “in relationship to making the contracts of reinsurance were done there or in New York . . . [a]nd, likewise, all things in regard to performance were to be done outside of Texas”).
115 *See id.* at 402–04.
Texas resident apparently failed to satisfy the Court that application of Texas law would be fair to the defendant insurer.\footnote{Id. at 408–09. But see WEINTRAUB, supra note 17, at 666–67 (suggesting that it is “tempting” to explain the Dick result purely as a matter of surprise, but also pointing out other facts in the record that could have put the insurer on notice that the policy might one day be analyzed under Texas law or in Texas courts).}

As noted above, the Court later cited \textit{Dick} as support for its decision in \textit{Yates v. John Hancock Mutual Life}.\footnote{299 U.S. 178 (1936).} Although \textit{Yates} ostensibly was decided under principles of full faith and credit rather than due process, there is more than a hint of concern for unfair surprise in the Court’s opinion. In \textit{Yates}, the only reason that might have supported the Georgia court’s decision to apply Georgia, rather than New York, law was that the plaintiff/beneficiary had recently moved to Georgia following the death of her husband.\footnote{Id. at 179.} However, the Court seemed unwilling to accept that a life-insurance beneficiary’s postdeath move to Georgia could subject an insurer to Georgia law when all other “contacts” prior to the insured’s death pointed to application of New York law.\footnote{Id. Many commentators have questioned whether \textit{Yates} retains any vitality today because the \textit{Yates} court followed a traditional, “vested rights” approach to choice of law, rather than the modern “most significant relationship” or “government interest” approach, as applied in one form or another by the majority of courts today. See, e.g., WEINTRAUB, supra note 17, at 662–63 n.36. The transformation of choice-of-law analysis from a territorial, “vested rights” approach (which relies exclusively on the geographic location of defined events, property, or persons) toward a “government interests” or “most significant relationship” approach, is beyond the scope of this article. See Allstate Ins. Co. v. Hague, 449 U.S. 302, 308 n.11 (1981) (summarizing the switch from geography-based to interest-based considerations in choice of law). The point in mentioning \textit{Yates} here is simply to show that unfair surprise as a constitutional limitation on choice of law “has a substantial history in Supreme Court litigation.” WEINTRAUB, supra note 17, at 674–75.}

2. \textbf{Hague: Full Faith and Credit Merges into Due Process}

Several decades later, in \textit{Allstate Insurance Co. v. Hague},\footnote{449 U.S. 302 (1981).} the Supreme Court squarely addressed the limits imposed by the Constitution on a court’s choice of law. Specifically, the Court confronted the issue of whether the Due Process Clause or the Full Faith and Credit Clause barred the courts of Minnesota from applying Minnesota law to determine the amount of benefits available under an auto insurance policy issued to a Wisconsin resident who was killed in a collision in Wisconsin.\footnote{Id. at 304–05, 320.} \textit{Hague} is a seminal decision, so its facts bear close scrutiny.

Mr. Hague was killed when a car struck the motorcycle on which he was riding as a passenger.\footnote{Id. at 305.} Mr. Hague, the motorcyclist, and the driver of the car were
all Wisconsin residents. Neither the motorcyclist nor the driver of the car had liability insurance. Mr. Hague, however, carried uninsured motorist coverage with Allstate on three vehicles he and his wife owned. The uninsured motorist coverage was limited to $15,000 for each of the three automobiles.

For fifteen years before the accident, Mr. Hague had worked in nearby Red Wing, Minnesota, where he commuted daily from his house in Wisconsin. After his death, Mr. Hague’s wife moved to Red Wing, married a Minnesota resident, and filed suit in Minnesota state court, seeking a declaration (as executor of Mr. Hague’s estate) that the three Allstate policies could be “stacked” under Minnesota law to provide a total recovery of $45,000. The Minnesota trial court concluded that Minnesota’s choice-of-law rules required application of Minnesota law because Wisconsin law, which prohibited stacking, was “inimical to the public policy of Minnesota.” The Minnesota Supreme Court applied Professor Leflar’s “better law” analytical approach to the choice-of-law issue and concluded that Wisconsin’s prohibition of stacking was an inferior, minority rule. It therefore held that Minnesota law applied, over Allstate’s objections.

The Supreme Court accepted certiorari to decide whether the application of Minnesota law was consistent with due process and full faith and credit. Justice Brennan, in writing for a plurality of four Justices, reviewed the Court’s previous decisions addressing constitutional limits on choice of law, including Dick and Yates, and distilled all of them into the following test for constitutionality: “[F]or a State’s substantive law to be selected in a constitutionally permissible manner, that State must have significant contacts or a significant aggregation of contacts, creating state interests, such that the choice of its law is neither arbitrary nor fundamentally unfair.” Even the three dissenters “accepted with few

123 Id.
124 Id.
125 Id.
126 Id.
127 Id. at 305. The record indicated that Mrs. Hague’s move to Minnesota occurred “almost concurrently” with, but was not motivated by, her initiation of the lawsuit. Id. at 319.
128 Id. at 306.
129 Id. at 307. For Professor Leflar’s “better law” approach, see Robert A. Leflar, Choice-Influencing Considerations in Conflicts Law, 41 N.Y.U. L. REV. 267 (1966). Professor Leflar urged courts to consider the following factors in choosing which state’s law to apply: (1) predictability of result, (2) maintenance of interstate order, (3) simplification of the judicial task, (4) advancement of the forum’s governmental interests, and (5) application of the “better” rule of law. Id. at 282.
130 Hague, 449 U.S. at 306–07.
131 Id. at 320.
132 Id. at 304.
133 Id. at 309–11.
134 Id. at 312–13 (emphasis added).
reservations” this test for whether a choice-of-law decision is constitutional. In applying this constitutional test, Justice Brennan proceeded to analyze the nature and extent of Minnesota’s contacts with the parties and the insurance agreements, and whether application of Minnesota law would violate the parties’ legitimate expectations regarding the controlling law.

Significantly for future noncompete cases, Justice Brennan’s plurality opinion specifically acknowledged 

_Yates_ as standing for the proposition that a “change in residence to the forum State” before filing suit is not, by itself, sufficient to justify application of forum law. Nothing in _Hague_ undermines this crucial holding of _Yates_, nor does _Hague_ call into question the similar holding in _Dick_ regarding the limits of due process. And whereas _Yates_ and _Dick_ arguably reflect the Court’s efforts to protect the parties’ justified expectations at the time of contracting, all eight Justices who decided _Hague_ agreed that Allstate had no justified expectations that Wisconsin law would apply to the issue of what benefits were payable under the insurance policies. None of the Justices disputed Justice Stevens’s assessment of the parties’ reasonable expectations at the time of contracting because (i) there was no choice-of-law clause in the Allstate policies, (ii) stacking was permitted in most jurisdictions, and (iii) coverage was not limited to accidents occurring in Wisconsin, Allstate could not reasonably have believed that Wisconsin’s rule against stacking would always apply.

Given that there was nothing to indicate that applying Minnesota law would frustrate Allstate’s expectations and thus be “fundamentally unfair,” the only point of disagreement among the three _Hague_ opinions was whether there was a sufficient aggregation of contacts to create state interests that would make it reasonable to apply Minnesota, rather than Wisconsin, law. Aside from Mrs. Hague’s postaccident relocation to Minnesota, the plurality concluded that there

---

135 Id. at 332 (Powell, J., dissenting); see also Phillips Petroleum Co. v. Shutts, 472 U.S. 797, 819 (1985) (noting the _Hague_ dissenters’ agreement with the test laid out in the plurality opinion).


137 Id. at 311.

138 See Home Ins. Co. v. Dick, 281 U.S. 397, 408 (1930) (holding, for due process purposes, the claimant’s postaccident move to the forum state was “without significance”).

139 Id. at 327–30. Justice Stewart did not participate in the decision, which resulted in a 4-1-3 split: Justice Brennan wrote an opinion for a plurality of four Justices, Justice Stevens concurred in the result, and Justice Powell, joined by two others, dissented.

140 Id. at 328–30 & nn.20–23. Any perceived unfairness to Allstate is surely lessened by the fact that the case involved an automobile insurance policy that provided coverage throughout the continent. _Hague_, 449 U.S. at 318 & n.24. Given that the “place of injury” was at that time (and remains) a significant factor in choosing law to govern tort suits, it would have been well within Allstate’s contemplation that some other law might apply to an accident outside Wisconsin. See id. at 324 (Stevens, J., concurring). Moreover, the policies did not have a choice of law provision that mandated application of Wisconsin law. Id. at 318 n.24. Thus, it likely would be inaccurate to say that Allstate lost the benefit of its bargain by having its policies stacked under Minnesota law. It doesn’t appear that it really bargained for Wisconsin law to apply in all circumstances.
were additional contacts that supported application of Minnesota law—namely, (i) Mr. Hague’s employment in (and commute to and from) Minnesota for the 15 years preceding his death, and (ii) Allstate’s business presence in Minnesota. The plurality held that these contacts “[i]n the aggregate” were sufficient to create state interests such that application of Minnesota law was constitutionally permissible.

These additional Minnesota contacts seem thin, at best. Indeed, Justice Powell dissented on the grounds that Mr. Hague’s employment in Minnesota was a “trivial” contact that failed to give Minnesota any legitimate state interests in applying its own public policy permitting stacking of uninsured motorist benefits. As Justice Powell reasoned, interpretation of Mr. Hague’s insurance contracts had nothing to do with his employment in Minnesota. Further, the mere fact that Allstate was licensed by the state to write insurance policies for Minnesota residents might have supported the assertion of personal jurisdiction there, but it says nothing about whether it was reasonable to apply Minnesota law to determine benefits payable under the Hague insurance policies. If doing business in a state were enough to justify application of that state’s law to any dispute involving that party, then there would truly be no constitutionally impermissible choice-of-law outcomes for companies with national operations, even if the suit does not arise from the company’s in-state business activities. Because neither Mr. Hague’s employment nor Allstate’s general business presence in Minnesota are related to

---

141 Hague, 499 U.S. at 313–18.
142 Id. at 320.
143 Id. at 337–39 (Powell, J., dissenting). Justice Powell noted that if the dispute had involved workers compensation benefits, or employer liability, or any other rights fairly linked to Hague’s status as a Minnesota employee, then perhaps it would have been reasonable to conclude that Minnesota had a legitimate interest in applying its substantive law. See generally id. at 338–39. Minnesota surely has a state interest in ensuring that the rights of employees working in Minnesota are vindicated—and that employers are held accountable for violating those rights—even if they reside in a neighboring state. But it is difficult to see how the mere fact that Mr. Hague commuted to work in Minnesota for fifteen years created any legitimate basis for Minnesota to want to apply its law to determine benefits due under an auto insurance policy between an Illinois insurer and a Wisconsin policyholder for an accident that occurred in Wisconsin involving Wisconsin motorists and injuries treated in Wisconsin. Permitting the stacking of Hague’s insurance policies did not make Minnesota workers or Minnesota roads any safer, nor did Minnesota employers become more accountable for the safety of their employees, nor did healthcare providers in Minnesota receive greater reimbursement, nor was Minnesota’s public fisc any better insulated from payouts to injured persons. There was no evidence that Mrs. Hague was in jeopardy of going on public welfare if stacking had been denied.
144 Id. at 338–39.
145 Id. at 338 & n.4 (Powell, J., dissenting) (discussing John Hancock Mut. Life Ins. Co. v. Yates, 299 U.S. 178 (1936), and noting that the fact that an insurer does business in and is extensively regulated by the forum state is highly relevant to the issue of personal jurisdiction, but not to choice of law).
146 See id. at 337–38 (Powell, J., dissenting) (“[T]his argument proves too much.”).
the interests that Minnesota might legitimately wish to protect by requiring stacking of Hague’s policy proceeds, these “contacts” seem to be nothing more than judicial make-weight, designed to help Justice Brennan’s plurality reach a desired result.\textsuperscript{147}

The plurality’s reliance on such weak affiliating contacts effectively sounded the death knell for the Full Faith and Credit Clause as an independent, constitutional limitation on choice of law. The plurality paid lip service to the notion that a state court may not apply its own law unless it first identifies legitimate state interests.\textsuperscript{148} But after \textit{Hague}, it is hard to see how the Full Faith and Credit Clause (as distinguished from the Due Process Clause) would ever disallow any choice-of-law decision that was not completely arbitrary or manifestly dishonest.\textsuperscript{149} Rather, \textit{Hague} effectively reduces the test for constitutionality to a single due process standard that is violated only when the result is truly arbitrary or fundamentally unfair to one of the parties.\textsuperscript{150}

The \textit{Hague} test implies a possible distinction between “arbitrary” and “fundamentally unfair” choice-of-law decisions. A truly arbitrary choice-of-law determination is unlikely to ever be a concern in noncompete litigation, however. If an employee moves from State A to State B to take a new job, applying State B’s law to determine the enforceability of the noncompete would not truly be arbitrary. State B surely has a legitimate interest in whether one of its citizens (even if newly arrived) is able to work and earn a living in her chosen profession. State B has an interest in attracting a talented workforce and ensuring that its residents have the means to support themselves, contribute to the state’s economy, and pay taxes into the state treasury. State B also has a legitimate interest in ensuring that in-state employers are able to hire and train desirable employees and compete in the world economy, regardless of where those employees formerly resided. Therefore, the real question in such cases is not whether applying the employee-favoring law of State B would be “arbitrary,” but whether it is “fundamentally unfair” to the former employer.

\textsuperscript{147} In his concurring opinion, Justice Stevens seemed equally skeptical that any of these contacts mattered. \textit{See id.} at 331 (Stevens, J., concurring). But Justice Stevens was untroubled by this because, in his unique view, the Full Faith and Credit Clause was satisfied so long as application of forum state law did not threaten our national unity or the sovereignty of another state. \textit{Id.} at 323. Justice Stevens saw no indication that Wisconsin’s sovereignty would be threatened by a Minnesota court applying its own rules permitting stacking of uninsured motorist policies. As such, he saw no full-faith-and-credit problem with Minnesota’s decision to apply its own law. \textit{Id.} at 325.

\textsuperscript{148} \textit{Id.} at 308, 320.

\textsuperscript{149} \textit{See Reynolds & Richman, supra} note 66, at 28 (doubting that any “conscientious” application of any of the now-dominant or traditional choice-of-law theories would ever run afoul of the Full Faith and Credit Clause). Professor Weintraub believes that this has been the de facto reality since the Court’s decision in \textit{Alaska Packers Ass’n}, 294 U.S. 532; \textit{see Weintraub, supra} note 17, at 694.

\textsuperscript{150} 449 U.S. at 313.
In his *Hague* concurrence, Justice Stevens observed that if a court applied the law of a state that the litigants could not reasonably have anticipated would be applied, then fundamental unfairness might result.\(^{151}\) He noted that a “choice-of-law decision that frustrates the justifiable expectations of the parties can be fundamentally unfair.”\(^{152}\) In fact, the three dissenters agreed with Justice Stevens on this point.\(^{153}\) Justice Stevens even characterized the prevention of unfair surprise to litigants as the Court’s “central concern” in the Court’s review of choice-of-law decisions under the Due Process Clause\(^ {154}\)—a concern that would in fact play a “central” role a few terms later in *Shutts*.\(^ {155}\)

Notably, Justice Stevens observed that the Court’s prior decisions striking down applications of forum law can mostly be “explained as attempts to prevent a State with a minimal contact with the litigation from materially enlarging the contractual obligations of one of the parties where that party had no reason to anticipate the possibility of such enlargement.”\(^ {156}\) He was not, of course, referring to any noncompete decisions, but his observations concerning frustration of parties’ reasonable, contractual expectations are highly relevant in that context. In commenting on the “unfairness” prong of the test, Professor Weintraub argues that *Hague* supports the proposition that:

> a person who plans his or her conduct in justifiable reliance on the law of one state, should not have that reliance frustrated by the application of the law of another state if, at the time the person acted, that other state did not have a contact with the parties or with the transaction that would make use of its law reasonable.\(^ {157}\)

3. *Shutts*: Protecting Party Expectations as the Central Concern of Due Process

The Supreme Court got a chance to apply *Hague* just a few terms later in *Phillips Petroleum Company v. Shutts*.\(^ {158}\) In *Shutts*, a Kansas trial court certified a class action consisting of 28,100 individual owners of natural gas leases who

\(^{151}\) *Id.* at 327 (Stevens, J., concurring).

\(^{152}\) *Id.*

\(^{153}\) *Id.* at 333 (Powell, J., dissenting) (citing Russell J. Weintraub, *Due Process and Full Faith and Credit Limitations on a State’s Choice of Law*, 44 Iowa L. Rev. 449, 455–57 (1959)) (with respect to evaluating whether a choice of law is fundamentally unfair, “[t]he touchstone here is the reasonable expectation of the parties”).

\(^{154}\) *Id.* at 327 (Stevens, J., concurring); see also *id.* at 331 (“When the expectations of the parties at the time of contracting are the central due process concern, as they are in this case, an unanticipated postaccident occurrence is clearly irrelevant for due process purposes.”).

\(^{155}\) See infra notes 158–169 and accompanying text.

\(^{156}\) *Hague*, 449 U.S. at 327 n.16 (Stevens, J., concurring) (emphasis added).

\(^{157}\) WEINTRAUB, supra note 17, at 669 (emphasis added).

\(^{158}\) 472 U.S.797 (1985).
claimed that Phillips owed interest on royalty payments that it suspended while awaiting federal approval of price increases.\textsuperscript{159} Of the 28,100 plaintiff class members, less than 1,000 were residents of Kansas, and only one-quarter of one percent of the gas leases were on land located in Kansas.\textsuperscript{160} The class included residents of all fifty states, and the leases were spread out among eleven different states, although predominantly in Texas and Oklahoma.\textsuperscript{161} Phillips was a Delaware corporation with its principal place of business in Oklahoma, although it was a very large company that did a substantial amount of business in Kansas.\textsuperscript{162} The trial court, affirmed by the Kansas Supreme Court, applied Kansas law to all claims in the case\textsuperscript{163} and found Phillips liable for payment of interest on the suspended royalties.\textsuperscript{164} The Kansas courts further set the applicable rate of interest according to Kansas principles of equity.\textsuperscript{165}

The United States Supreme Court granted certiorari to resolve the question of whether it was constitutionally permissible for a Kansas court to apply Kansas law to govern the claims of all class members, even though most of them resided in other states, and nearly all of the gas leases were on land in other states.\textsuperscript{166} The Court, relying on its newly minted “neither arbitrary nor fundamentally unfair” standard, said no. In its 7-1 decision, the Court was highly protective of the contracting parties’ reasonable expectations: “When considering fairness in this context, an important element is the expectation of the parties.”\textsuperscript{167} The Court noted that there was no indication in the record that when the leases—which involved land and royalty owners outside of Kansas—were executed, “the parties had any idea that Kansas law would control.”\textsuperscript{168} Consequently, the Court held “that application of Kansas law to every claim in this case is sufficiently arbitrary and unfair as to exceed constitutional limits.”\textsuperscript{169} Thus \textit{Shutts} firmly supports the proposition that due process will not permit a court to apply the substantive law of the forum to resolve a contractual dispute when the contracting parties would have had no reasonable basis when they entered into the contract to expect that that state’s law would apply.

The Court has never decided a choice-of-law case involving noncompetes. But while the constitutional limits on choice of law were evolving over the course of the last century, covenants not to compete were becoming more commonplace in

\begin{thebibliography}{9}
\bibitem{159} Id. at 801.
\bibitem{160} Id.
\bibitem{161} Id. at 799, 815.
\bibitem{162} Id.
\bibitem{163} Id. at 814.
\bibitem{164} Id. at 816.
\bibitem{165} Id.
\bibitem{166} Id. at 799.
\bibitem{167} Id. at 822.
\bibitem{168} Id.
\bibitem{169} Id.
\end{thebibliography}
employment contracts. And as noncompetes became more widespread, so did noncompete litigation. What some courts seemingly have failed to appreciate—at least in the noncompete area—is that due process limitations post-\textit{Hague}, although generous, are not unlimited. Even if the test for due process is not stringent, it is still a test, and some choice-of-law decisions will not pass it. Under \textit{Shutts}, \textit{Hague}, and earlier cases like \textit{Dick} and \textit{Yates} (which have not been overruled even if their traditional choice-of-law methodology is no longer in favor\footnote{Katherine W. Stone, From Widgets to Digits: Employment Regulation for the Changing Workplace 130 (2004).}), due process may not be ignored, even when applying modern choice-of-law approaches like the approach provided in Sections 187 and 188 of the Second Restatement\footnote{\textit{Restatement (Second) of Conflict of Laws} §§ 6, 188 (1971) (setting forth the “most significant relationship” test for choosing which state’s law to apply to a particular dispute).}. The next Part of this Article looks at specific cases where states have failed to appreciate the limitations on choice of law imposed by the Due Process Clause and, consequently, how noncompete litigation in those states has devolved into an unseemly and inherently unfair sprint to the courthouse.

\section*{IV. Failure to Respect Due Process Limits on Choice of Law Has Transformed Modern Noncompete Enforcement into a Race to the Courthouse}

The constitutional test for choice of law as articulated in \textit{Hague} is not a demanding one, but it does not grant the kind of unfettered choice-of-law authority that many states and commentators seem to have assumed. \textit{Shutts} specifically teaches that in cases involving contractual obligations, it may well violate due process to subject a contracting party to the substantive law of a state that, at the time of contracting, the party had no reason to believe would control. \textit{Hague} itself—by affirming and distinguishing \textit{Yates}—calls into question the ability of a state to apply its own laws to a dispute merely because one of the litigants has recently relocated there\footnote{Allstate Ins. Co. v. Hague, 449 U.S. 302, 319 (1981).}. These cases would suggest, therefore, that a court should be wary of applying its local substantive law to invalidate a noncompete when the employee has only recently moved to that state and is subject to a choice-of-law clause that requires application of the law of the state where she previously lived and worked and where noncompetes are generally permitted.\footnote{472 U.S. at 822.}
Yet nonenforcement of parties’ contractual choice of law is rampant in noncompete cases. There are numerous published and unpublished cases that reflect a troubling lack of concern for due process in choosing the law that will apply to an interstate noncompete dispute. Seven of the nation’s thirty-one largest metropolitan areas are located in two states that are notoriously strict when it comes to enforcing noncompetes in employment contracts: California and Georgia. As noted in Part II, California outlaws them by statute, and Georgia will strike down any noncompete that fails to conform to narrow and very specific requirements. As illustrated below, both of these states (as well as Oklahoma, Colorado, and North Dakota) will unapologetically apply their unforgiving laws to any noncompetes litigated in their courts, without regard to party expectations or fundamental fairness. It is the insistence of states like these to apply their own laws, combined with the iron rule of full faith and credit, that has fostered a race-to-the-courthouse mentality in noncompete litigation.

A. California

California is perhaps the most common battleground for interstate noncompete disputes. California courts are particularly inhospitable for out-of-state employers seeking to enforce employee noncompetes because a California statute explicitly prohibits them except in very narrow circumstances. Employment lawyers know that California courts have a tendency to protect employees from the consequences of a noncompete, even when the employee is new to California and previously worked in another state where noncompetes are allowed. Application Group, Inc. v. Hunter Group, Inc. is one particularly controversial example. In Application Group, the California court of appeals affirmed the trial court’s ruling that California law should be applied to a noncompete entered into between a Maryland employer and an employee residing in Maryland—even though her employment contract contained a Maryland choice-of-law clause and even though her new position would not require her to physically relocate to California. The facts of Application Group perfectly illustrate the remarkable tendency of California courts to apply California law at any opportunity.

176 See, e.g., Larry E. Ribstein, From Efficiency to Politics in Contractual Choice of Law, 37 GA. L. REV. 363, 376 (2003) (observing that the parties’ contractual choice-of-law provisions were not enforced in 29 out of 71 cases involving noncompetes).
177 See infra Part IV.A–E.
178 Los Angeles (No. 2), Atlanta (No. 9), San Francisco (No. 11), Riverside-San Bernardino (No. 13), San Diego (No. 17), Sacramento (No. 24), and San Jose (No. 31).
181 See generally Wardwell, supra note 14; Wu, supra note 14.
183 Id. at 87–91.
Hunter Group was headquartered in Maryland and provided computer-consulting services to clients primarily in the eastern United States but also to customers in California.\textsuperscript{184} Hunter maintained branch offices in multiple states, including California.\textsuperscript{185} At the time of the litigation, only two of Hunter’s ninety employees resided in California, and only five of its customers were located in California.\textsuperscript{186} Application Group was a direct competitor of Hunter.\textsuperscript{187} Application Group was headquartered in San Francisco, California, and 30 of its 106 consultants were stationed in California.\textsuperscript{188}

The employee at the center of the case, Dianne Pike, resided in Maryland, and had done so since 1963.\textsuperscript{189} Pike was hired by Hunter as a consultant in 1991 to work in the Baltimore office and to service customers in Arizona, Colorado, Massachusetts, and New York.\textsuperscript{190} As a Hunter employee, Pike never set foot in California, not even for pleasure.\textsuperscript{191} When Pike resigned from Hunter and accepted a consulting job at Application Group, Hunter sued Pike in Maryland state court for breach of contract, citing Pike’s noncompete.\textsuperscript{192} The noncompete precluded Pike from rendering computer-consulting services to any competitor of Hunter for one year after termination, and it included a Maryland choice-of-law clause.\textsuperscript{193} The Maryland court found the noncompete to be enforceable under Maryland law (without considering any of the choice-of-law questions), but later entered judgment for Pike on the ground that Hunter had not adequately proved damages.\textsuperscript{194}

While the Maryland lawsuit was pending, however, Application Group filed its own suit in state court in San Francisco, seeking a judicial declaration that Pike’s noncompete was unenforceable under California law—specifically, Section 16600 of the California Business and Professions Code.\textsuperscript{195} Application Group and Pike sought a ruling that California law applied and, moreover, that Hunter should be precluded from enforcing any Maryland judgment upholding the noncompete.\textsuperscript{196} The trial court agreed that California law applied to any dispute involving a

\textsuperscript{184} Id. at 75.  
\textsuperscript{185} Id.  
\textsuperscript{186} Id.  
\textsuperscript{187} Id.  
\textsuperscript{188} Id. at 75–76.  
\textsuperscript{189} Id. at 76.  
\textsuperscript{190} Id.  
\textsuperscript{191} Id.  
\textsuperscript{192} Id. Hunter also sued Application Group for unlawful interference with the Hunter-Pike contractual relationship. Id.  
\textsuperscript{193} Id.  
\textsuperscript{194} Id. at 76–77. Hunter’s failure to prove damages is left unexplained in the California court’s opinion. One wonders whether this fact, if known to the trial court, colored its views of the case or its concern for protecting Hunter’s expectations regarding choice of law.  
\textsuperscript{195} Id. at 77.  
\textsuperscript{196} Id.
California employer’s efforts to recruit employees to work for it, and that under California law, Pike’s noncompete with Hunter was invalid. The trial judge held that California’s strong public policy against noncompetes required application of California law, regardless of "where the parties exchanged promises" (Maryland). In the trial court’s view, Section 16600 constituted an addendum to the noncompete that said “VOID IN CALIFORNIA.” Indeed, so strong is the California policy that it voids any noncompete, and are “not dependent on the employee’s domicile nor his contacts with California.” Thus, the court found that Hunter employees like Pike, despite any noncompete or choice-of-law clause to the contrary, were free to perform work in California for a competitor of Hunter, regardless of their state of residence or place of employment, and regardless of whether they worked for Hunter in California.

On appeal, the court of appeals framed the issue as whether California law applies to render a noncompete unenforceable when it pertains to nonresidents recruited by California employers. Because Pike’s employment agreement had a choice-of-law clause, the court stated that it was required to apply the law of the state chosen by the parties in their agreement (Maryland) unless (1) the chosen state had no substantial relationship to the parties or the transaction, or (2) application of that state’s law would be contrary to a fundamental policy of the forum (California). The appeals court went on to state that it was not bound to follow the parties’ choice-of-law clause if California had a materially greater interest than Maryland in determining whether the noncompete was enforceable—that is, if California’s public policies would be more seriously impaired by application of Maryland law than Maryland’s policies would be impaired by application of California law.

Using this “comparative impairment” approach, the court observed that Section 16600 is a strong expression of California public policy that “every citizen shall retain the right to pursue any lawful employment” with the employer of his or her choice. Note the implied expansiveness of the court’s description of this policy: “every citizen” of the United States has this unlimited right of employment, not just every citizen of California. Citing a long line of California precedents, the court noted that an employee’s interests in his own “mobility and betterment” are “paramount” to the competitive interests of employers in preventing employees

---

197 Id. at 77–78.
198 Id. at 78.
199 Id.
200 Id. (emphasis added).
201 Id.
202 Id. at 81.
203 Id. at 82 n.10 (citing, inter alia, RESTATEMENT (SECOND) OF CONFLICT OF LAWS § 187(2) (1971)).
204 Id. at 83–85 (describing the “comparative impairment” analysis used by California courts in resolving conflicts issues).
205 Id.
206 Id. at 85 (emphasis added).
from working for competitors. The court viewed this state policy as also protecting the interests of California employers in being able to recruit and compete for the most talented and skilled employees in their industry, regardless of residence. It therefore followed that “California has a strong interest in protecting the freedom of movement of persons whom California-based employers (such as AGI) wish to employ to provide services in California, regardless of the person’s state of residence or precise degree of involvement in California projects.” Again, note the universal scope of the court’s pronouncements.

By contrast, the court found there was “nothing in the record” to suggest that failure to enforce Pike’s noncompete would “significantly impair” Maryland’s interests in protecting Maryland employers from recruitment of their best or scarcest talent. This finding seems especially dubious. If Maryland’s policy is to enforce reasonable noncompetes, then a refusal to enforce Pike’s noncompete—which involved a Maryland company’s employment of a Maryland resident—seems about as damaging to Maryland’s state interests as can be imagined. One doubts that, if asked, Maryland’s courts or legislature would agree with the Application Group court’s assessment of the degree to which Maryland law would be impaired by the decision to invalidate the Hunter/Pike noncompete. For example, in Estee Lauder Cos. v. Batra, the district court asserted that New York’s interest in protecting New York employers from being raided for their best employees is just as strong as California’s interest in ensuring the mobility of workers in California. One could rightly cite Application Group and Estee Lauder as classic examples of how the outcome of a “comparative impairment” analysis tends to favor the state that is conducting it. Moreover, by couching its holding in terms of what is “in the record,” the Application Group court was disingenuously suggesting that Hunter’s problem was evidentiary, when in reality the court was declaring Maryland’s legal interests to be inferior to California’s.

In a similar vein, the court also found no evidence that Pike provided “unique services” to Hunter such that her departure to work for a competitor would cause harm to Hunter. This holding is a pure straw man; there is no indication anywhere in the court’s opinion or anywhere in the reported decisions of Maryland courts that enforcement of noncompetes is limited to cases where the employee has provided “unique services.” Nor does the court explain why a finding that Pike rendered “unique services” would tip the scales toward application of Maryland law, given that California has a strong interest in ensuring that its employers can compete for the best talent available. If Pike had truly been a provider of unique

---

207 Id.
208 Id.
209 Id. (emphasis added).
210 Id. at 86.
212 See id. at 173 (“Just as California has a strong interest in protecting those employed in California, so too does New York have a strong interest in protecting companies doing business [in New York].”).
213 Application Grp., 72 Cal. Rptr. 2d at 86.
skills, that fact would surely increase California’s interest in assuring that California employers were free to hire her.

In short, the court in Application Group was dismissive of Maryland’s countervailing public policies, preferring to apply California law instead. Maryland may well have strong interests in protecting the investments that Maryland employers make in recruiting, hiring, training, and developing their employees—interests that it expresses through consistent enforcement of reasonable noncompetes, but the Application Group court refused to take them seriously. The Application Group decision does not explain why these interests were deserving of less respect than California’s or why they would suffer less harm by nonenforcement of Pike’s noncompete than California’s interests would suffer if the noncompete were upheld. The court simply declared itself to be “convinced that California has a materially greater interest than does Maryland in the application of its law to the parties’ dispute, and that California’s interests would be more seriously impaired if its policy were subordinated to the policy of Maryland.”

Nowhere to be found in the lengthy Application Group opinion is any reference to whether it might be unconstitutional to apply California law. Hague forbids application of the forum state’s substantive law unless the state has sufficient contacts to create legitimate state interests such that application of its law is “neither arbitrary nor fundamentally unfair.” And Shutts expressly holds that it is fundamentally unfair to apply the law of a state that none of the parties would have thought controlling at the time they entered into their contract. In Application Group, Hunter and Pike’s employment relationship was based entirely in Maryland. When she accepted employment with Hunter in 1991 and entered into the noncompete, there were no affiliating contacts that would have made application of California law reasonable. It is only because she unilaterally elected (years after executing the contract) to take a job with an employer based in California that applying California law to govern her noncompete could be viewed as anything but arbitrary. Dick, Yates (as reaffirmed in Hague), and Shutts all stand for the proposition that a plaintiff cannot expand or change the parties’ contractual rights simply by moving to a more favorable jurisdiction before bringing suit on the contract. Yet the California appellate court permitted Pike to nullify Hunter’s contractual rights merely by taking a job with a California employer—even though she did not move to California. Heedless of Shutts, the Application Group court was thoroughly dismissive of Hunter’s argument that Maryland law should be

214 Id.
217 Application Grp., 72 Cal. Rptr. 2d at 76.
218 Id. (“Pike never set foot in California, even for pleasure, during the time she was employed by Hunter.”).
220 See Application Grp., 72 Cal. Rptr. 2d at 87.
applied because it would fulfill the parties’ expectations. 221 As noted, at the time of contracting, Maryland law was the only apparently applicable law. 222 But the court assumed, based solely on Hunter’s prior efforts to enforce other noncompetes, that Pike’s noncompete was designed to preclude her recruitment by California employers. 223 Perhaps Hunter had California competitors in mind when it executed Pike’s contract, or perhaps not. But there is no indication in the opinion as to what Pike’s intentions or expectations were when she signed the agreement, so it seems particularly troubling that the court just assumed that the parties expected at the time of contracting that California law might one day be relevant to its enforcement. Pike had lived in Maryland for decades and was not hired by Hunter to do any work for California customers. 224 She had not set foot in California for business or pleasure during her employment at Hunter. 225 There was no basis for Hunter to believe when Pike signed the noncompete that California law might one day apply, and thus disregard of the Maryland choice-of-law provision seems fundamentally unfair under Shutts. 226

The willingness of the court in Application Group to apply California law without even pausing to consider due process limitations stands as an open invitation to future litigants to seek refuge from their noncompetes in California. 227 After Application Group, there may as well be a sign posted at the California border saying, “Your noncompete stops here.”

B. Oklahoma

Although far less populous and therefore less burdened with litigation arising from job-hopping employees, Oklahoma is very nearly as restrictive as California when it comes to enforcement of noncompetes. Oklahoma statutes categorically prohibit the use of noncompetes in employment contracts, with one narrow exception: an employer may prohibit its former employees from directly soliciting

---

221 Id.
222 Id.
223 Id. at 87–88.
224 Id. at 76.
225 Id.
226 Despite its many flaws, Application Group is one case where the seemingly unfair result cannot be attributed to the employee’s winning the race to the courthouse. Pike and Application Group could have filed a preemptive suit in California immediately upon her resignation from Hunter (or sooner, in fact); but they did not, which allowed Hunter to file first in Maryland. Hunter lost that tactical advantage by allowing the California court to decide the enforceability of the noncompete even though the issue had already been adjudicated in Hunter’s favor in Maryland. Id. at 76. The opinion does not indicate that Hunter ever raised collateral estoppel to preclude relitigation of that issue in the California courts. It is also unclear from the opinion whether Hunter’s lawyers simply failed to raise due process objections to the choice of California law, or whether the courts ignored their objections.
business from its established customers. Like California, the courts of Oklahoma are vigilant in ensuring that Oklahoma employers are free to recruit employees from other states, regardless of whether those employees are subject to a noncompete.

For example, in *Herchman v. Sun Medical* a medical laser salesman filed suit in Oklahoma seeking a declaratory judgment that his noncompete with a Texas-based employer was invalid. Although living in Oklahoma at the time he filed suit, Herchman resided in Texas when he signed his employment agreement with Sun Medical, which was headquartered in Texas. The agreement contained a Texas choice-of-law provision and a noncompete that prohibited Herchman from selling competing lasers for one year after termination. The noncompete was clearly unenforceable under Oklahoma’s statute but may well have been enforceable under Texas law.

Purporting to employ the most significant relationship test of the Second Restatement, the district court discounted all of the parties’ contacts with Texas. The court declared that Herchman’s residence when he entered into the noncompete had no bearing on the issue of whether it would prevent him from accepting a job selling medical lasers for a competing firm in Oklahoma years later. In the court’s view, Oklahoma law applied because the present effect of the clause was to restrict employment in Oklahoma, regardless of any other contacts with the state of Texas. The district court went so far as to hold that, even if it concluded that Texas law ordinarily would apply under the most-significant-relationship test, it would nonetheless be constrained to find the noncompete unenforceable as contrary to Oklahoma public policy. With no discussion of the interests of Texas in applying its law and no discussion of the legitimate expectations of the parties, the court stated, “Any argument that the contract is enforceable under Texas public policy is irrelevant, as the contract is not sought to be enforced there.” In other words, forget most-significant-relationship analysis, government interests, comity, or fairness to the parties: Oklahoma law

---

230 Id. at 944.
231 Id. at 945.
232 *Herchman* was litigated under a previous version of title 15, section 217 of the Oklahoma Statutes, but Herchman’s noncompete was clearly unenforceable under both the previous and current versions.
233 See supra notes 54–57 and accompanying text.
234 RESTATEMENT (SECOND) OF CONFLICT OF LAWS §§ 6, 188 (1971).
235 *Herchman*, 751 F. Supp. at 945.
236 Id. at 945 n.2 (“[A]n express agreement of the parties that the contract is to be governed by the laws of a particular state will be given effect if the contract bears a reasonable relation to the chosen state and no counter[vening] public policy of forum demands otherwise.”).
237 Id. at 947 n.4.
238 Id.
applies in Oklahoma courts. Period. The due process limits on choice of law as articulated in Hague and applied in Shutts were never mentioned.239 Mr. Herchman strategically filed suit in Oklahoma before Sun Medical could file suit in Texas, and because of the district court’s aggressive application of Oklahoma law, that fact proved outcome-determinative.

It is difficult to reach any firm conclusions about whether the choice of Oklahoma law in Herchman represents an actual violation of due process. The opinion indicates that Mr. Herchman may have moved to Oklahoma and worked there while still employed at Sun Medical,240 although his career at Sun Medical began in Texas and he executed his contract there.241 Therefore, it is at least possible that the parties may have contemplated a move to Oklahoma at the outset of their employer-employee relationship, in which case application of Oklahoma law to Mr. Herchman’s noncompete may not represent the kind of unfair surprise that would offend due process under Shutts, Yates, or Hague. Nonetheless, the breadth of the district court’s dicta indicates that Oklahoma courts have misapprehended the limitations that due process places on their authority to apply Oklahoma law to multi-state disputes. It also means that an employee who can sue first in Oklahoma will have a substantial advantage against his former employer in any action concerning the validity of his noncompete.

C. Colorado

Colorado, like Oklahoma, has developed a strong aversion to noncompetes in employment contracts, and in Dresser Industries, Inc. v. Sandvick242 the Tenth Circuit affirmed a Colorado district court’s decision to apply Colorado law to invalidate the one-year noncompetes of three former employees of a company headquartered in Texas.243 In so doing, the Tenth Circuit acknowledged that it was obligated to consider the public policies of the forum (Colorado) and other states with an interest in applying their law.244 The court noted that several of those states—namely, Colorado, North Dakota, and Montana—all had strong public policies against noncompetes.

The appeals court further acknowledged that the noncompetes were likely enforceable under Texas law and that Dresser likely expected its employees to adhere to the agreements they signed when they accepted employment at Dresser.245 It specifically recognized that the interests of predictability and uniformity in employee treatment would best be served by having the employees’

239 See generally id.
240 Id. at 945.
241 Id. at 944.
242 732 F.2d 783 (10th Cir. 1984).
243 Id. at 784 & n.1, 787–88.
244 Id. at 785–86.
245 Id. at 785.
246 Id. at 785–86.
agreements all governed by Texas law. 247 But while conceding that the parties had “justified expectations” that their noncompetes would be binding, 248 the court nevertheless found that those expectations did not “supersede all other considerations.” 249 The court instead held that Texas’s interest in protecting the parties’ expectations was outweighed by the interests of the states—like Colorado—that have policies against enforcement of noncompetes. 250 Accordingly, the court affirmed application of Colorado law to invalidate the noncompetes, never bothering to explain why due process allowed it to reach a result that defied what it assumed to be the parties’ justified expectations as to enforceability. 251 It declared the price of predictability to be “too high” if it meant disregarding the policy preferences of the forum state 252 but it never explained why circumvention of justified expectations was not an equally high price.

Like the application of Oklahoma law in Herchman, it is not entirely clear that the Dresser Industries court’s application of Colorado law was truly a due process violation. The employees had all worked for Dresser in multiple states, including states where noncompetes are highly restricted. 253 But like the district court in Herchman, the appeals court in Dresser court paid no heed to the constitutional considerations that Hague held were necessary. 254 Instead, Dresser focused exclusively on Colorado’s distaste for noncompetes and justified its holding only in those terms. 255 Any due process limits on choice of law were seemingly overlooked by a court that was preoccupied by parochial interests.

D. North Dakota

Like California, North Dakota’s attitude toward noncompetes in employment contracts is both overtly hostile and liberally applied. Thus, employees tend to do very well when they litigate their noncompetes in North Dakota courts. For example, in Pruco Securities Corp. v. Montgomery, 256 the district court noted the existence of a New Jersey choice-of-law clause in Montgomery’s employment agreement but concluded that North Dakota law should apply because of “public policy considerations.” 257 The court apparently deemed it inconsequential that Prudential was headquartered in New Jersey and that the parties had expressly

247 Id. at 786.
248 Id. at 787.
249 Id.
250 Id.
251 Id. at 787–88.
252 Id. at 786.
253 Id. at 785 (noting the place of performance including a possible ten-state area including North Dakota, Colorado, and Wyoming).
254 See generally Dresser, 732 F.2d 783.
255 Id. at 787–88.
257 Id. at 868.
elected to be bound by New Jersey law. In choosing to apply North Dakota law, the court in *Pruco* did not discuss whether its choice of law comported with due process. Rather, the court viewed “public policy” as carte blanche to invoke forum law.

E. Georgia

Even states that do not absolutely prohibit noncompetes can still present significant barriers to employers wanting to enforce a noncompete against a former employee. Georgia courts, for example, have proved just as vexing to out-of-state employers as California courts have been. Although Georgia courts apply so-called “strict scrutiny” rather than outright prohibition of noncompetes, they take an extraordinarily expansive view of when it is permissible to apply Georgia law to invalidate noncompetes that do not conform to Georgia’s strict and specific requirements. As the following cases vividly illustrate, the fact that parties may have contractually agreed to application of another state’s law is seemingly no obstacle for Georgia courts to apply Georgia law when asked to enforce a noncompete.

In *Enron Capital & Trade Resources Corporation v. Pokalsky*, an employee of Enron in Texas resigned his position as a commodities trader to accept a similar job with Southern Electric in Georgia. Before its spectacular collapse, Enron was primarily in the business of buying and selling energy and trading energy commodities. While working at Enron, Pokalsky helped to develop commodities

---

258 Id. Similarly, in *Forney Industries, Inc. v. Andrus*, 246 F. Supp. 333 (D.N.D. 1965), the employment contracts at issue recited that they were “made and entered into under the laws of the State of Colorado” and included a Colorado choice-of-law and venue provision. Id. at 333. But the North Dakota district court held that deference to the parties’ contractual choice of law was not necessary if it would lead to results that were contrary to the public policy of the forum or place of performance. Id. at 334. The court therefore applied North Dakota law and declared the noncompetes invalid. Id. at 335. In light of Colorado’s statutory prohibition on most noncompetes, one might reasonably ask why Forney Industries provided for application of Colorado law in its employment contracts. I am merely speculating as to the employer’s thinking (if indeed there was any thought at all behind the choice), but it appears from earlier cases that Colorado’s views of noncompetes were not always as clear as they are today, such that application of Colorado law in 1965 would not have yielded consistently pro-employee results. Of course, to the extent that the Colorado choice-of-law clause was mindless boilerplate, it would cut against any due process argument that application of North Dakota law violated the reasonable expectations of the parties. Moreover, due to its vintage, the *Forney* decision was decided without the benefit of *Hague* and *Shutts*.

259 See 264 F. Supp. 2d 862. *Forney* was decided in 1965, more than fifteen years earlier than *Hague* in 1981 and *Shutts* in 1984. So it is not surprising the *Forney* court did not consider due process.


261 Id. at 137.

262 Id.
trading procedures. His employment agreement included a noncompete that precluded Pokalsky from engaging in any business that was competitive with Enron in any geographic area where Enron did business—which effectively meant anywhere in the world—for one year following termination. On the day he resigned his post in Texas, Pokalsky and Southern filed a declaratory judgment action in Georgia state court seeking to have his noncompete declared unenforceable. The Georgia trial court declared his noncompete invalid and issued a temporary restraining order (TRO) enjoining Enron from taking any action to enforce the noncompete.

Before the Georgia court entered the TRO, Enron filed suit in Texas seeking to enforce the noncompete. Southern and Pokalsky responded by asking the Georgia court to enjoin Enron’s Texas suit and any discovery in it. The Georgia trial court granted Southern’s motion and ordered Enron to dismiss its suit in Texas and not to sue elsewhere. The Georgia Court of Appeals affirmed the trial court’s injunction and its order compelling Enron to dismiss the Texas case, reasoning that Enron could bring any counterclaims in the Georgia action rather than maintaining a separate suit in Texas.

The court further held that Georgia law controlled the validity of the noncompete, despite a choice-of-law clause specifying application of Texas law. The court stated that Georgia courts do not have to enforce a choice-of-law clause if it is “particularly distasteful.” Although it did not explain why the choice-of-law clause itself was distasteful, the appeals court concluded that the noncompete provision was “particularly distasteful” because it prohibited competitive employment worldwide, which Georgia law did not permit. Consequently, the court simply refused to give effect to the parties’ contractual choice of Texas law because it viewed the noncompete as contrary to Georgia public policy and therefore “prejudicial to the interest[s] of this state.” It reached this decision without even the pretense of comparing Georgia’s state interests against those of Texas; without examining the various state contacts with the parties and the contract or the state interests created by those contacts, as would be proper under Hague; and without mentioning, much less analyzing, whether the parties had

263 Id.
264 Id. at 138.
265 Id. at 137–38.
266 Id. at 138.
267 Id.
268 Id.
269 Id. at 138–39.
270 Id. at 139.
271 Id.
272 Id.
273 Id.
274 Id.
legitimate expectations that might make application of Georgia law fundamentally unfair, as required by *Hague, Shutts*, and *Dick*.  

Pokalsky thus is a perfect illustration of an employee capitalizing on informational asymmetries to choose a favorable forum and thereby circumvent any serious, choice-of-law analysis. Pokalsky joined forces with his new employer to file suit on the day he gave notice to Enron, thereby guaranteeing that his noncompete would be evaluated by a very skeptical court. And because Georgia is extremely aggressive in applying local law, Pokalsky and Southern could be confident that the Texas choice-of-law clause would be disregarded, even though it would seem eminently reasonable for contracting parties to choose the state where the employee worked and the employer was headquartered as the law to govern their agreements. As was the case in *Shutts*, the mere fact that a state might have some interest in applying its law does not mean that due process permits it to do so. With no consideration of whether Enron might reasonably have expected that Georgia law would govern the enforceability of Pokalsky’s noncompete, the Pokalsky court’s decision thus reflects complete judicial disregard for due process limits on choice of law. And, indeed, by enjoining Enron’s later-filed suit in Texas with no mention of comity concerns, the court arrogantly imposed its view of Georgia public policy on a Texas company and an employee who, even on the day of suit, was still living in Texas and working in the Texas office of a Texas employer. Given the Georgia court’s failure to appreciate the due process implications of its choice-of-law approach, Enron essentially lost the case when Pokalsky won the race to the courthouse.

Aggressively expansionist application of Georgia noncompete law is not a phenomenon observed only in state courts. Federal courts applying Georgia law have been complicit in expanding Georgia’s “strict scrutiny” regime to noncompetes emanating from other states without adequate regard for due process. Indeed, the importance of winning the race to the courthouse is perhaps best exemplified by two recent cases from the Eleventh Circuit, again applying Georgia law: *Keener v. Convergys Corp. (“Keener III”),* and *Manuel v. Convergys Corp.*

---

275 See generally id.
276 Id. at 137–38.
277 Id. at 136–37.
278 342 F.3d 1264 (11th Cir. 2003) (applying the Georgia Supreme Court’s response to certified questions); see also *Keener v. Convergys Corp. (Keener I)*, 312 F.3d 1236 (11th Cir. 2002) (certifying question of “whether a court applying Georgia conflict of laws rules follows the language of Restatement (Second) Conflict of Laws § 187(2) and, therefore, first must ascertain whether Georgia has a ‘materially greater interest’ in applying Georgia law, rather than the contractually selected forum’s law, before it elects to apply Georgia law to invalidate a non-compete agreement as contrary to Georgia policy”); Convergys Corp. v. Keener (*Keener II*), 582 S.E.2d 84 (Ga. 2003) (considering certified question from the 11th Circuit).
279 430 F.3d 1132 (2005).
When *Keener v. Convergys Corp.* ("*Keener I*")\(^{280}\) first reached the Eleventh Circuit, it appeared that due process and fundamental fairness still had a fighting chance. James Keener had been an employee of Convergys and its predecessor entities since 1984, working in the company’s offices in Cincinnati, Ohio, and Chicago, Illinois.\(^{281}\) He entered into a noncompete in 1995 that precluded him from engaging in certain competitive activities for a period of two years following termination of employment.\(^{282}\) Keener’s noncompete contained a choice-of-law provision that selected Ohio law to govern the agreement.\(^{283}\) In 2001, Keener accepted a position with H.O. Systems, which was a competitor of Convergys.\(^{284}\) But when he announced his resignation, Keener told his coworkers that he was going to work in the banking industry, leading them to believe that he would not be working for a competitor.\(^{285}\) Keener later admitted that he was concerned about his noncompete and negotiated with his new employer for assistance if Convergys sought to enforce it.\(^{286}\) A few weeks after Keener resigned and moved to Georgia to begin working for H.O. Systems, a Convergys employee happened to encounter Keener making a business call on a mutual customer.\(^{287}\) Convergys attorneys corresponded with H.O. Systems’ legal department, ultimately resulting in Keener entering into a separation agreement whereby H.O. Systems agreed to pay him $50,000 to resign.\(^{288}\)

Keener then sued in the Southern District of Georgia, seeking a declaration that his noncompete was unenforceable and an injunction to restrain Convergys from trying to enforcing it.\(^{289}\) The district court ruled in Keener’s favor, declaring the noncompete unenforceable under Georgia law and enjoining Convergys from enforcing it.\(^{290}\) On appeal, the Eleventh Circuit seemed concerned that the only connection argued by Keener to support application of Georgia law was that Keener was then living in Georgia.\(^{291}\) The court questioned whether, under section 187(2) of the Restatement (Second) Conflict of Laws, it was permissible to apply Georgia law without first finding as a threshold matter that Georgia had a “materially greater interest” than Ohio in applying its own law to the noncompete, which the district court had not done.\(^{292}\) The panel noted that the Georgia Supreme Court had previously cited section 187(2) in *Nasco, Inc. v. Gimbert*,\(^{293}\) a case

\(^{280}\) 312 F.3d 1236.  
\(^{281}\) *Keener I*, 312 F.3d at 1238.  
\(^{282}\) *Id.*  
\(^{283}\) *Id.*  
\(^{284}\) *Id.*  
\(^{285}\) *Id.*  
\(^{286}\) *Id.*  
\(^{287}\) *Id.*  
\(^{288}\) *Id.* at 1238–39.  
\(^{289}\) *Id.* at 1239.  
\(^{290}\) *Id.*  
\(^{291}\) *Id.*  
\(^{292}\) *Id.* at 1241.  
\(^{293}\) 238 S.E.2d 368 (Ga. 1977).
involving enforcement of a noncompete that included a Tennessee choice-of-law provision. To the Eleventh Circuit in Keener I, this suggested that Georgia law would not apply if another state had a “materially greater interest” in applying its own law. Further, two prior Eleventh Circuit cases had interpreted Nasco as requiring a court to find that Georgia has a “materially greater interest” before deciding to apply Georgia law over the law specified in the parties’ contract. However, the panel noted that the district court had applied Georgia law without first making the “materially greater interest” finding, and that the language of the Nasco opinion seemed to support the notion that Georgia’s conflict-of-laws rules required application of Georgia law whenever the law chosen by the parties yielded a result contrary to Georgia public policy. In the face of these seemingly conflicting authorities, the Eleventh Circuit certified the choice-of-law question to the Georgia Supreme Court for a definitive answer.

In answering the certified question, the Georgia Supreme Court held in Convergys Corp. v. Keener (“Keener II”) that the prior Eleventh Circuit cases had erroneously interpreted Georgia’s conflict-of-laws precedents. Claiming that the issue had already been squarely decided in Nasco, the Georgia Supreme Court reiterated the proposition that:

The law of the jurisdiction chosen by parties to a contract to govern their contractual rights will not be applied by Georgia courts where application of the chosen law would contravene the policy of, or would be prejudicial to the interests of, this state. Covenants against . . . competition . . . affect the interests of this state, namely the flow of information needed for competition among businesses, and hence their validity is determined by the public policy of this state.

There was no requirement that Georgia courts must first find that Georgia has a “materially greater interest” before applying Georgia law because, as the Keener II decision explains, Georgia had never adopted section 187(2) of the Restatement (Second) Conflict of Laws; Nasco merely cited to it with a qualifying “see,” and followed that citation immediately with reference to Georgia authorities standing for the proposition that Georgia courts will not apply law which would contravene

294 Id. at 369.
295 Keener I, 312 F.3d at 1241.
296 Id. at 1240 (discussing Nordson Corp. v. Plasschaert, 674 F.2d 1371 (11th Cir. 1982), and Bryan v. Hall Chem. Co., 993 F.2d 831 (11th Cir. 1993)).
297 Id. at 1240–41 (citing Mktg. & Research Counselors, Inc. v. Booth, 601 F. Supp. 615 (N.D. Ga. 1985)).
298 Id. at 1241.
299 582 S.E.2d 84 (Ga. 2003).
300 Id. at 87.
301 Id. at 85–86 (quoting Nasco, Inc. v. Gimbert, 238 S.E.2d 368, 369 (Ga. 1977)).
Georgia public policy or be prejudicial to the interests of Georgia.\textsuperscript{302} Indeed, the court in Keener II pointed to other Georgia cases that it characterized as expressly rejecting the Restatement’s “center of gravity” approach, calling it “neither less confusing nor more certain than our traditional approach.”\textsuperscript{303} Thus, under Keener II, Georgia courts will determine the enforceability of a noncompete by applying Georgia law, regardless of any contractual choice-of-law provision to the contrary.\textsuperscript{304}

That at least is how the Eleventh Circuit understood and applied Keener II when the case came back from the Georgia Supreme Court. This was unfortunate because, in fact, Keener II (perhaps inadvertently) left an “out” for the Eleventh Circuit in Keener III to avoid the inevitable application of Georgia law, if only the panelists had been paying closer attention. The Georgia Supreme Court’s opinion in Keener II took some liberties with its prior decision in Nasco, including bolstering it with the dubious assertion that Nasco had permitted application of Georgia law only “[a]fter first ascertaining that there were significant contacts with the State of Georgia, such that the choice of [Georgia] law was neither arbitrary nor constitutionally impermissible . . . .”\textsuperscript{305} In reality, the Nasco opinion says no such thing. Nasco did not mention Hague (indeed, it pre-dates Hague) and did not examine what contacts the litigation had with Georgia that might permit the application of Georgia law consistent with due process. But applying Georgia law was likely unproblematic in Nasco because the employee was a resident of Georgia when he entered into the noncompete and continued to work in Georgia.\textsuperscript{306} Under those circumstances, it is doubtful that the former employer in Nasco could credibly have complained of unfair surprise in applying Georgia law, despite the inclusion of a Tennessee choice-of-law clause.

The same cannot be said for the litigants in Keener, however. Keener had always worked for Convergys in Ohio and Illinois.\textsuperscript{307} His noncompete was executed while he was in Ohio and included an Ohio choice-of-law clause.\textsuperscript{308} The

\textsuperscript{302} Id. at 87 (citing GA. CODE ANN. § 1-3-9 and Ulman, Magill, & Jordan Woolen Co. v. Magill, 117 S.E. 657 (Ga. 1923)).

\textsuperscript{303} Id. at 86 (quoting Gen. Tel. Co. of the Se. v. Trimm, 311 S.E.2d 460, 462 (Ga. 1984)).

\textsuperscript{304} Id. at 85–86; Keener III, 342 F.3d 1264, 1267 (11th Cir. 2003). The Georgia Supreme Court in Keener II criticized section 187(2) interest analysis as susceptible to manipulation and obfuscation. The opinion perceptively observes that a state court’s interpretation of its own public policy consideration is “outcome determinative” in choice-of-law questions far more often the complex rules derived from Section 187(2). Id. at 87; see also Katherine Florey, State Extraterritorial Powers: A Reply, 85 NOTRE DAME L. REV. 1157, 1160 (2010) (noting that judges seemingly pick and choose state interests arbitrarily among a “garbled menu” of choices in the context of extraterritorial analysis); RICHARD A. POSNER, THE PROBLEMS OF JURISPRUDENCE 429–30 (1990) (decrying the uncertainty of consequence-based choice-of-law rules).

\textsuperscript{305} 582 S.E.2d at 85.

\textsuperscript{306} Nasco, Inc. v. Gimbert, 238 S.E.2d 368, 369 (Ga. 1977).

\textsuperscript{307} Keener I, 312 F.3d 1236, 1238 (11th Cir. 2002).

\textsuperscript{308} Id.
only contact any party had with Georgia was that Keener moved there to take a job with H.O. Systems after quitting his job at Convergys.\(^{309}\) Had the Eleventh Circuit in *Keener III* been more mindful of due process limitations, it could have held that, despite the broad authority of Georgia courts to apply Georgia law—as established by *Nasco* and confirmed by *Keener II*—application of Georgia law to Keener’s noncompete was nonetheless “constitutionally impermissible” because, unlike in *Nasco*, there were no prior contacts with Georgia that would make it fair to apply Georgia law to the parties’ contract. The court could have held that applying Georgia law was “fundamentally unfair” under *Shutts* (and *Yates* and *Dick*) because neither party had any prior connection to Georgia. Furthermore, the only reason a Georgia court was given the opportunity to apply Georgia law was because Keener unilaterally decided to move there after quitting his job at Convergys.\(^{310}\) The court could have referred to referenced Justice Brennan’s plurality opinion and Justice Stevens’ concurring opinion in *Hague*, both of which rejected the idea that plaintiff’s recently acquired residency in the forum state could, by itself, justify application of forum law.\(^{311}\) In reversing the district court’s application of Georgia law as a violation of due process, the *Keener III* opinion could have simply cut-and-pasted from the *Shutts* opinion: “There is no indication that when the [noncompetes] involving [an employee and employer] outside of [Georgia] were executed, the parties had any idea that [Georgia] law would control.”\(^{312}\)

Instead, the Eleventh Circuit panel in *Keener III* fatalistically condemned the Convergys noncompete to invalidation by rote applying Georgia law and affirming the trial court’s entry of declaratory and injunctive relief.\(^{313}\) With seemingly no appreciation of the significant due process concerns raised by the specific facts of the case, the Eleventh Circuit declared conclusorily that “application of Georgia law is not arbitrary or constitutionally impermissible because Convergys would be attempting to enforce the . . . [noncompete] against

\(^{309}\) *Id.* at 1239 (“The only connection argued is that Keener currently lives in Georgia and was employed by H.O. Systems in Georgia.”).


\(^{311}\) *Allstate Ins. Co. v. Hague*, 449 U.S. 302, 331 (1981) (Stevens, J., concurring);

\(^{312}\) *id.* at 319 (plurality opinion) (noting that while the Court had previously held that “postoccurrence change of residence to the forum State was insufficient in and of itself to confer power on the forum State to choose its [own] law,” but that the Court had not yet held that a change in residence was irrelevant).

\(^{313}\) *Shutts*, 472 U.S. at 822 (“When considering fairness in this context, an important element is the expectation of the parties. There is no indication that when the leases involving land and royalty owners outside of Kansas were executed, the parties had any idea that Kansas law would control.” (citation omitted)).

\(^{313}\) *Keener III*, 342 F.3d 1264, 1266 (11th Cir. 2003).
Keener, who is living and working in Georgia, where the effects would be felt.”314 In so concluding, the court was undoubtedly correct that Georgia had an interest in applying its law to the noncompete because it affected a Georgia company and a (recently relocated) Georgia employee. But due process requires more than an identifiable state interest; it requires that the choice of law not be “fundamentally unfair.”315 Failing to acknowledge this key distinction, Keener III held that the noncompete did not comport with Georgia’s strict requirements for geographic specificity (among other defects), and thus declared it invalid.316

Beyond the fundamentally unfair result for Convergys, the Keener cases establish a mandatory choice-of-law rubric applicable to all future litigants in Georgia courts. This rubric should be enough to raise the eyebrows of any conflict-of-laws scholar. Keener III neatly summarized it this way: “The . . . [noncompete] was contrary to Georgia public policy, Georgia law therefore applied, and the [noncompete] was unenforceable under Georgia law due to its overbreadth.”317 In other words, if the noncompete is objectionable under the forum state’s law, then the forum state’s law applies, which necessarily means the noncompete will not be enforced. It is hard to fathom a more overtly self-aggrandizing (not to mention result-driven) approach to choice of law. Rather than asking first whether it is nonarbitrary and fair to apply Georgia law under the particular facts of the case, Georgia courts now must apply Georgia law if the noncompete would violate Georgia law. This effectively means that Georgia law always applies in Georgia courts, because the only way for a noncompete to survive is if it satisfies Georgia’s strict requirements.

Employees have learned remarkably well from the lessons taught by Keener. Indeed, less than a year after the Georgia Supreme Court decided Keener II, another Convergys employee, William Manuel, successfully followed the blueprint laid out by Mr. Keener. In Manuel v. Convergys Corp.,318 Mr. Manuel executed a two-year noncompete in 2003 when he accepted a promotion to director of sales.319 The noncompete contained a choice-of-law provision requiring application of Ohio law (again, Convergys was an Ohio corporation headquartered in Ohio), although Mr. Manuel resided in Florida throughout his employment with Convergys.320 In March 2004, while still working for Convergys in Florida, Manuel secretly began negotiating with Mellon Financial for a position to be located in either North

314 Id. at 1267–68.
315 Hague, 449 U.S. at 302–13 (plurality opinion); see also Shutts, 472 U.S. at 822 (admonishing courts that when considering fairness, “an important element is the expectations of the parties”).
316 342 F.3d at 1268–69.
317 Id. at 1269.
318 430 F.3d 1132 (11th Cir. 2005).
319 Id. at 1134.
320 Id. The agreement also contained a forum-selection clause requiring any suit on the noncompete to be brought in Ohio. Manuel v. Convergys Corp., No. 1:04-1279-MHS, 2004 WL 5545025, at *1 (N.D. Ga. Oct. 18, 2004). This provision was inexplicably ignored in the Manuel district court and court of appeals opinions.
Carolina or Georgia; he eventually was offered and accepted the job in Georgia.\textsuperscript{321} During these negotiations, Manuel discussed with Mellon the possibility that Convergys might sue to enforce the noncompete.\textsuperscript{322} Manuel, however, told Convergys that he was not going to work for a competitor.\textsuperscript{323}

Without informing Convergys, Manuel accepted Mellon’s offer on April 5, 2004, promptly met with Georgia lawyers to prepare a lawsuit, and then resigned from Convergys on April 8, 2004, although he agreed to stay on until the end of the month.\textsuperscript{324} During that time, Manuel denied that he had accepted employment with another company and that he was going to work for a competitor.\textsuperscript{325} On April 9, he signed an apartment lease in Georgia and obtained a Georgia driver’s license.\textsuperscript{326} Then, on April 20, 2004, he filed suit against Convergys in Georgia state court seeking a declaration that the noncompete was invalid and unenforceable.\textsuperscript{327} Having now realized Manuel’s true intentions, Convergys tried to discuss with Mellon how Manuel’s job could be structured so as not to violate the noncompete, but when those discussion proved fruitless, Convergys filed its own suit in Ohio.\textsuperscript{328}

The district court in Georgia denied Convergys’s request to stay the case pending the outcome in Ohio, citing the “first-filed” rule, which holds that “in the absence of compelling circumstances, the court initially seized of a controversy should be the one to decide the case.”\textsuperscript{329} On appeal from Manuel’s victory in the district court, the Eleventh Circuit first addressed the question of venue. It upheld the district court’s ruling that there was nothing improper about Manuel’s anticipatory filing in Georgia, such that the case should have been dismissed or stayed pending the outcome of the Ohio lawsuit.\textsuperscript{330} The Eleventh Circuit noted that Convergys did not immediately file suit after learning of Manuel’s employment with Mellon, but rather tried to seek assurances that his new job could be structured to comply with the noncompete.\textsuperscript{331} In this respect, the court effectively penalized Convergys for trying to work things out rather than rushing to file suit. In addition to its reliance on the “first-filed”

\textsuperscript{321} Manuel, 430 F.3d at 1134. Significantly, North Carolina has a much more permissive attitude toward noncompetes and does not apply its law as expansively as does Georgia. See Manuel, 2004 WL 5545025, at *3 (noting that the defendants alleged that the employee selected Georgia because its law was more favorable).

\textsuperscript{322} Id.

\textsuperscript{323} Id.

\textsuperscript{324} Id.

\textsuperscript{325} Id.

\textsuperscript{326} Id.


\textsuperscript{328} Manuel, 430 F.3d at 1134.

\textsuperscript{329} Manuel, 2004 WL 5545025, at *2 (citing Merrill Lynch, Pierce, Fenner & Smith, Inc. v. Haydu, 675 F.2d 1169, 1174 (11th Cir. 1982)). In fact, the federal district court in Ohio stayed the Ohio case pending the outcome of the Georgia proceedings. Id. at *1.

\textsuperscript{330} Manuel, 430 F.3d at 1136–37.

\textsuperscript{331} Id. at 1134.
the court defended the propriety of suit in Georgia on standard venue grounds—such as, location of documents and witnesses and convenience to the parties. Based on all these factors, the Eleventh Circuit in *Manuel* found that it was not an abuse of discretion for the Georgia district court to entertain the declaratory judgment action.

If one disregards the forum-selection clause in favor of Ohio courts (which curiously neither the district court nor the appeals court appears to have considered), then *Manuel* is surely correct from a venue perspective. Yes, Mr. Manuel was forum shopping, but courts have always tolerated a certain amount of forum shopping, and there is nothing inherently scandalous or off-putting about a plaintiff filing suit in the judicial district where he resides, even if his residency there is recently acquired (assuming his intent to remain there is not fraudulent). Moreover, the first-filed rule is a sensible rule recognized across the federal circuits. The real trouble with *Manuel* lies in its cavalier approach to choice of law once it determined that the district court in Georgia properly entertained the action.

Convergys argued that Georgia lacked significant contacts with the parties and events giving rise to the dispute to justify application of Georgia law to the noncompete. The Eleventh Circuit, however, treated this issue as a matter of settled law. In holding that Georgia law could constitutionally be applied, the *Manuel* court rubber stamped *Keener* III’s conclusory due process analysis, concluding that Manuel’s residence in Georgia at the time of suit was sufficient contact with the state to permit application of Georgia law. As with *Keener*, there was no mention of the Supreme Court’s decisions expressly holding that a recent move to the forum state is insufficient to justify application of the forum-state’s law. Convergys did not dispute that the noncompete was invalid under

---

332 Id. at 1135–36 (citing U.S. Fire Ins. Co. v. Goodyear Tire & Rubber Co., 920 F.2d 487, 488 (8th Cir. 1990)).
333 Id. at 1137.
334 Id. at 1138.
337 *Manuel*, 430 F.3d at 1138–39.
338 Id. at 1139.
Georgia law. The district court’s decision invalidating the noncompete was therefore affirmed.339

It is troubling that the Eleventh Circuit in Manuel seemed fully cognizant of the race-to-the-courthouse problem and yet was unwilling to address it.340 The court noted that when Manuel sought legal counsel regarding the noncompete, his lawyer warned him that if Convergys obtained an Ohio judgment validating the noncompete, a Georgia court would likely have to enforce it.341 Thus, he recommended—astutely, as it turns out—that “the only thing [Manuel] could do to prevent this scenario would be to file first in Georgia.”342 So it cannot have escaped the Eleventh Circuit’s attention that the only reason the employee prevailed is that he managed to file suit first in Georgia. Indeed, the court recognized in Keener III that Mr. Keener lied to his superiors to gain an advantage in litigation: “Keener, as a consequence of his misrepresentation, avoided enforcement by Convergys in Ohio and benefited in electing a jurisdiction that is hostile to . . . [noncompetes].”343 The Eleventh Circuit has thus explicitly acknowledged that informational asymmetries and lack of candor between the employee and the former employer allow the employee to game the system. It is hard to understand how a federal appeals court could stomach results that incentivize calculated dishonesty and penalize the party who refrains from suing in the hope that an acceptable compromise can be reached without resort to litigation. This fundamentally unfair result was entirely avoidable, if only the court had been mindful of the due process implications of its decision.

By giving short shrift to due process limitations, Manuel, Keener, and Pokalsky have inappropriately invited state and federal judges to view forum-state policy considerations as a choice-of-law trump card. Perhaps even if the Manuel court had paid proper deference to due process as contemplated by Hague and Shutts, it nonetheless would have concluded that Ohio law was not the proper choice because, unlike Mr. Keener, who had always lived and worked in Ohio, Mr. Manuel had lived and worked in Florida before moving to Georgia.344 Rather than fostering a sense of comity among the states or respect for fundamental fairness to the litigants, the Manuel/Keener approach to choice of law is a recipe for rampant provincialism and disregard for due process in noncompete cases.345

339 Id.
340 Id. at 1136 (“In Convergys’s view, Manuel won an illicit race to the courthouse and the district court thereby erred in rewarding him with a declaratory judgment action.”).
341 Id. at 1134.
342 Id. (citation omitted).
343 Keener III, 342 F.3d 1264, 1270 (11th Cir. 2003).
344 See supra notes 278–312, 328–330 and accompanying text.
345 See WEINTRAUB, supra note 17, at 662 (observing that the plurality in Hague probably would have found it “so outrageous as to be unconstitutional” for Minnesota to apply its own law if the plaintiff’s recent move to the forum had been the only justification offered; see also Thomas v. Wash. Gas Light Co., 448 U.S. 261, 272 (1980) (plurality opinion) (noting the Full Faith and Credit Clause was intended to prevent one state’s “parochial entrenchment on the interests of other States”).
Shortly before Manuel was decided, the Eleventh Circuit took the expansionist impulse one step further in Palmer & Cay, Inc. v. Marsh & McLennan Companies. As with Keener and Manuel, the Eleventh Circuit in Palmer & Cay affirmed a district court’s declaration that a two-year noncompete was unenforceable under Georgia law, even though the employee (a senior insurance executive earning $725,000 per year and supervising 2,600 sales employees) moved to Georgia only so he could become president of Palmer & Cay, a direct competitor of his former employer, Marsh & McLennan. Stunningly, the court did not even pause to consider whether application of Georgia law was permissible; it only considered whether the various noncompete provisions in question could survive Georgia’s strict requirements for noncompetes ancillary to employment. Finding that one of the noncompetes failed those requirements, it affirmed the district court’s holding that the covenant was unenforceable.

What makes Palmer & Cay a notable extension of Keener is that the Eleventh Circuit reversed the district court’s holding that the declaratory judgment was limited to Georgia. Rather, it held that declaratory judgments of Georgia courts are enforceable nationwide under the Full Faith & Credit clause, and likewise, under federal common law, the judgment of a federal court sitting in Georgia and exercising its diversity jurisdiction should not try to restrict the effect of its holding to enforcement only in Georgia. In Keener III, the Eleventh Circuit had tried to maintain some semblance of extraterritorial modesty by observing that “Georgia cannot in effect apply its public policy decisions nationwide—the public policy of Georgia is not that everywhere.” It therefore held that the agreement was unenforceable in Georgia only, implicitly leaving open the possibility that Convergys could enforce its noncompete if Mr. Keener later had second thoughts about moving to Georgia and left to work in another state. Palmer & Cay nullifies even this modest effort to reign in the expansionist tendencies of Georgia courts with respect to noncompetes.

Thus, under Keener, Manuel, and Palmer & Cay, the message is clear: due process is no obstacle to application of Georgia law to invalidate a noncompete even if Georgia acquired no interest in the dispute until the day the suit was filed. Georgia courts will apply Georgia law when asked to enforce a noncompete against a Georgia resident (no matter how recently arrived), and moreover, when

---

346 404 F.3d 1297 (11th Cir. 2005).
347 Id. at 1301–02, 1306.
348 Id. at 1306–07.
349 Id.
350 Id. at 1309–10.
351 See id. at 1310.
352 Keener III, 342 F.3d 1264, 1269 (11th Cir. 2003).
353 See id. at 1269–71.
354 Palmer & Cay, 404 F.3d at 1310–11.
that noncompete is invalidated (as it almost inevitably will be), the declaratory judgment of invalidity should be accorded preclusive effect nationwide. 355

F. Employer Turnabout: Is It Fair Play?

Although the employee is almost inevitably capable of winning the race to the courthouse, occasionally the employer gets there first—if, for example, the employee (or his counsel) misjudges the employer’s resolve to enforce the noncompete or is slow to act on his superior knowledge regarding the timing of his move. Just as filing first in a state like California or Georgia can convey an outcome-determinative advantage to the employee, an employer can obtain a similar advantage if it files first in a forum with more permissive attitudes toward noncompetes (for example, its home state). This is often true even when the employee later files a countersuit in an employee-friendly forum.

For example, in *Google, Inc. v. Microsoft Corp.*, 356 a federal district court in California stayed an employee’s second-filed declaratory judgment action, effectively deferring to Microsoft’s first-filed action in Washington state court. Microsoft filed suit in Seattle to enforce the noncompete of Kai-Fu Lee, its Vice President for Research and Development and head of operations in China, when he announced he was leaving to become Vice President of Engineering and the head of Google’s newly opened research center in China. 357 Sensing the urgency of obtaining a favorable forum, Microsoft filed suit in Washington on the same day that Lee resigned. 358 Three days later, Lee and Google countersued in California, arguing that Lee had become a California resident and that his noncompete was unenforceable under California law. 359

Being first to file in Washington was crucial for Microsoft, given California’s hostility to noncompetes. While acknowledging the large gulf between California and Washington law, the district court in San Jose, California declined to hear Lee

---

355 Defenders of the *Palmer & Cay* decision might point out that the appeals court, although holding that the declaratory judgment should not be limited only to Georgia, nonetheless affirmed the district court’s decision to limit the scope of its *injunction* to Georgia. *Id.* at 1309. It is unclear how this limitation could be of any consolation to Convergys. Perhaps it would remain free to seek enforcement of the noncompete against Mr. Keener in another state, but the clear holding of *Palmer & Cay* is that the other state would not be free to second-guess the Eleventh Circuit’s declaratory judgment rendering Keener’s noncompete invalid and unenforceable. *Id.* at 1309–11 (citing the Full Faith and Credit Clause, U.S. CONST. art. IV, § 1; 28 U.S.C. § 1738; Baker v. Gen. Motors Corp., 522 U.S. 222, 233 (1998); Hostetler v. Answerthink, Inc., 599 S.E.2d 271, 275 (Ga. Ct. App. 2004)). This holding thus confirms the wisdom of Mr. Manuel’s lawyer, who warned that if Convergys obtained an Ohio declaratory judgment validating Manuel’s noncompete, a Georgia court would have to enforce it, and thus his best option was to file first in Georgia. *Manuel v. Convergys Corp.*, 430 F.3d 1132, 1134 (11th Cir. 2005).

357 *Id.* at 1019–20, 1025 n.8.
358 *Id.* at 1019.
359 *Id.* at 1020.
and Google’s declaratory judgment action, invoking its discretion to stay the case under *Brillhart v. Excess Insurance Company of America*. The district court held that the Washington court was fully capable of deciding whether to apply Washington or California law, and thus there was no reason for a court in California to entertain a duplicative proceeding. To the district court, it appeared that the only reason Google and Lee filed suit in California was to secure a more favorable forum, and the court saw no need to expend additional judicial resources on a parallel lawsuit. Without resolving the choice-of-law issue, the district court acknowledged that there appeared to be colorable arguments for application of either California or Washington law, but concluded that under *Brillhart*, it was prudent to let the Washington state court resolve the issue without interference from a parallel federal action. With the California case on hold, Microsoft was able to negotiate a settlement with Lee and Google after the Washington court ruled under Washington law that Lee could not engage in certain research and development activities for Google until his one-year noncompete expired.

A similar result obtained in *Swenson v. T-Mobile USA, Inc.*, when a high-ranking T-Mobile executive quit her post at T-Mobile’s headquarters in Bellevue, Washington, and accepted a position with a start-up wireless carrier based in Los Angeles. Three days after Swenson left her job as Chief Operating Officer, T-Mobile sued her in Washington state court to enforce her noncompete. Two days later, Swenson filed a declaratory judgment action in state court in San Diego, which T-Mobile promptly removed to federal court. The federal district court followed the example set in *Google v. Microsoft* and declined to hear Swenson’s second-filed suit, deferring to T-Mobile’s earlier-filed Washington action under *Brillhart*. But rather than merely granting a stay, the district court in *Swenson*

---

360 Id.; see also *Brillhart v. Excess Ins. Co. of Am.*, 316 U.S. 491, 495 (1942) (holding that federal courts may decline to grant declaratory relief when another lawsuit involving the same parties and same issues of state law is pending in a state court).


362 See id. at 1021–22.

363 Id. at 1024–25.

364 Id. at 1025–26.


367 Tricia Duryee, *Decision Favors T-Mobile over Former Top Exec*, SEATTLE TIMES (Nov. 18, 2005), http://community.seattletimes.nwsource.com/archive/?date=20051118&slug=t-mobile18.

368 *Swenson*, 415 F. Supp. 2d at 1103.

369 Id.

370 Id. at 1105–06.
dismissed the California action outright.\footnote{371} Predictably, the Washington court applied Washington law and enforced the noncompete, precluding Swenson from accepting the position she had been offered with the competing wireless carrier.\footnote{372}

As Microsoft and T-Mobile demonstrate, employers can use the same preemptive tactics employees have used to greatly increase their chances of a favorable outcome in noncompete litigation. Significantly, however, the employer’s preemptive suit will almost never present the kind of due process/unfairness problems that the employee’s preemptive suit may. For example, the Microsoft and T-Mobile employees in the cases just discussed could hardly claim unfair surprise at the application of Washington law to uphold their noncompetes. Both employees were working in Washington when they accepted their new jobs in California, and their employment agreements both contained Washington choice-of-law clauses.\footnote{373} Additionally, given that Washington is the location of the companies’ corporate headquarters,\footnote{374} Washington surely had a sufficient interest in enforcement of the noncompetes such that application of Washington law cannot be regarded as arbitrary or fundamentally unfair under Hague. Thus from a constitutional perspective, there is typically a vast qualitative difference between an employer’s preemptive suit in a forum that is receptive to noncompetes and an employee’s preemptive suit in a forum that is hostile to them.

But even putting aside constitutional concerns, the disparity in outcomes between cases like Application Group and Google raises an important policy question: Why should the judicial system effectively reward whichever party is the first to pull the lawsuit trigger?

The answer is, it should not. The present state of affairs is unsound as a matter of constitutional law and undesirable as a matter of public policy.

V. RESTORING DUE PROCESS CONSTRAINTS ON CHOICE OF LAW IN THE ENFORCEMENT OF NONCOMPETES

As shown in Part III, a long line of Supreme Court cases has established constitutional limits on the ability of a state to apply its own laws to disputes in its courts, even if the test for constitutionality is not a stringent one.\footnote{375} Although the precise scope of those constitutional limits is not well defined in all cases, one rule is clearly discernible: The Constitution forbids a state to declare that it will apply its own law to every dispute in its courts, regardless of whether the parties ever anticipated that its law would apply.\footnote{376}

\footnote{371} Id.
\footnote{372} See id. at 1106 (noting that the Washington state court had already issued a preliminary injunction enjoining Swenson from working with the competitor); see also Durree, supra note 365.
\footnote{373} Swenson, 415 F. Supp. 2d at 1102; Google, Inc. v. Microsoft Corp., 415 F. Supp. 2d 1018, 1019 (N.D. Cal. 2005).
\footnote{374} Swenson, 415 F. Supp. 2d at 1102; Vise, supra note 365.
\footnote{375} See supra Part III.
\footnote{376} See supra Part III.
When a state plays the “public policy” card to trump the parties’ legitimate expectations for choice of law, it risks violating the parties’ due process rights. And as shown in Part IV, failure to respect due process limits on choice of law is rampant in states that have highly restrictive policies against noncompetes.

So what can be done to restore due process considerations to their proper place in choice-of-law analysis? This Part explores two potential reforms, both of which represent an improvement over the status quo, but only one of which avoids slanting too far in favor of the employer. Finally, as this discussion reveals, paying respect to constitutional concerns in the context of noncompete litigation can produce desirable consequences not just for the litigants in a particular case, but also for employee-employer relations and our judicial system.

A. Possible Reforms

One approach to reforming the present system of choice-of-law in noncompete litigation is simply to adopt a new, bright-line rule that tells courts what state’s law to apply in specific factual circumstances. Some of the more egregious due process violations described above could be avoided with an appropriate, mandatory rule. Another approach is to revise the process by which courts apply traditional choice-of-law rules in multistate noncompete cases, with constitutional limits explicitly invoked as a choice-of-law filter. The advantages and disadvantages of each of these two approaches are discussed below. Although either approach would lead to an improvement over the status quo, this Article urges adoption of the second approach.

1. A Bright-Line Proposal

One commentator on California law, Christina Wu, has suggested that California courts adopt a new, bright-line choice-of-law rule that would require acceptance of the parties’ contractual choice of law whenever the employee worked in another state and moved to California only to take a job with a competitor. Wu argues that one benefit to her approach is that it would increase predictability in the courts’ decisions and reduce employees’ incentives for forum shopping. As noted above, it is wrong to disparage as “forum shopping” an employee’s decision to file suit in his new state of residence, assuming the claim of residency is genuine and not purely tactical. But Wu’s proposed bright-line rule would undoubtedly increase predictability in litigation outcomes while minimizing parochial infringement of sister-state interests.

Another policy justification that Wu offers in support of her proposed choice-of-law rule is that it would protect the legitimate expectations of the parties in

---

377 See Wu, supra note 14, at 595–96.
378 Id. at 607, 613.
379 See supra note 335 and accompanying text (noting that our judicial system tolerates forum shopping in some instances).
nearly all cases.\textsuperscript{380} Again, Wu’s argument is logical and represents a significant improvement over the current state of affairs. But it seems unrealistic to expect California courts to adopt this sort of rule based purely on policy grounds. After all, the reason that a new choice-of-law rule seems necessary is because decisions like Application Group permit expansive applications of California public policy to strike down noncompetes, regardless of the parties’ domiciles or contacts with California.\textsuperscript{381} With that mindset in place, a policy-based argument for changing the courts’ choice-of-law rules seems destined to fail. Moderating courts’ impulses to expansively apply their local law to multistate noncompete disputes will likely require a return to constitutional principles, not just an appeal to policy concerns such as predictability or ease of implementation.

Another serious problem with Wu’s proposed rule is that it would essentially become an “employer-wins” rule. Operating under a bright-line rule that mandates enforcement of a contractual choice-of-law clause, the employer will almost always negotiate to include a choice-of-law provision that selects a state that broadly approves of noncompetes—and may even choose a state that has little connection to the employee’s specific duties—in an effort to ensure maximum post-termination restrictions on employment. So long as the chosen state’s connection is not so attenuated as to run afoul of Hague, the employer would thus inevitably prevail in an enforcement proceeding. Under Wu’s bright-line rule, courts in California would be required to uphold the noncompete even if, under the specific facts of the case, California clearly has a materially greater interest in applying its law than the state chosen in the parties’ contract. Accordingly, while Wu’s bright-line rule would be an improvement over the status quo, it ultimately remains an unsatisfactory cure for what ails.

2. A Two-Step Filter Proposal

Fortunately, restoring due process to its proper place in choice-of-law decision-making does not require adopting a rule that effectively means employers always win. Instead, this Article proposes a new, two-step approach to choice of law in noncompete cases, wherein due process acts as a filter to eliminate arbitrary or fundamentally unfair choices. Courts would then be free to employ their customary choice-of-law rules (such as those found in the Second Restatement) to select the controlling law from among those states that remain in contention after passing through the due process filter.

\textsuperscript{380} See Wu, supra note 14, at 619–20 (“Upholding the choice of law provision would ensure that the parties’ original expectation at the time of contracting are met.”). One could imagine a scenario in which enforcing the contractual choice of law could become arbitrary or unfair—for example, if the employment agreement was made long ago and in the intervening years the nature of the employer’s business and the employee’s state of residence had changed, such that the state chosen in the contract no longer bears any significant relationship to the parties or their employment relationship. Such scenarios would surely be rare.

\textsuperscript{381} Application Grp., Inc. v. Hunter, Inc., 72 Cal. Rptr. 2d 73, 78 (Cal. Ct. App. 1998).
Under this two-step approach, either party might prevail on the choice-of-law issue so long as the party advocates applying the law of a state that passes Hague’s admittedly low due-process test. For example, an employee like Dianne Pike in Application Group could successfully argue that California law should be applied if the facts revealed that the parties knew or reasonably should have known when they executed the employment agreement that she would likely go to work for a competitor in California if she were ever to resign from her job in Maryland. If the Maryland employer’s only significant competitors were in California, or if the employee previously lived in California or had a large family contingent in California, then the Maryland employer would not necessarily be unconstitutionally blindsided by application of California law. If, on the other hand, at the time of contracting, neither party contemplated relocating to California (that is, there was no more than a mere theoretical possibility of relocation), then applying California law would violate due process under Shutts. The essence of the due process inquiry in all cases would be whether it would have been reasonable at the time the parties entered into the noncompete to anticipate that California would one day have a legitimate interest in applying its own law.

In that respect, this proposed filtering approach to choice of law would somewhat resemble the constitutional test for personal jurisdiction. Due process does not permit a state to assert personal jurisdiction over a non-resident defendant unless the defendant has sufficient contacts with the forum such that maintenance of the suit “does not offend traditional notions of fair play and substantial justice.” Due process forbids maintaining personal jurisdiction over a defendant with whom the forum has “no contacts, ties, or relations.” It should be likewise for choice of law: A court in California, for example, would have to look at the relevant contacts of the parties with the state and decide whether it would have been reasonable at the time of contracting for the employer to have anticipated that California might one day have a legitimate interest in applying California law to the noncompete in question. If so, then applying California’s invalidating statute may comport with due process. But as in the personal jurisdiction context, an employee’s unilateral move into California could not, on its own, support the decision to apply the law of that jurisdiction to the out-of-state employer.

The personal jurisdiction analogy is not perfect, of course, as Justice Powell noted in his dissent in Hague. Arguably, the constitutional threshold for applying

---

382 *Cf.* Florey, *supra* note 304, at 1159 (noting that Hague’s due process test for choice of law provides no more protection to defendants than the minimum contacts test for personal jurisdiction).


384 *Id.* at 319; *World-Wide Volkswagen Corp. v. Woodson*, 444 U.S. 286, 294 (1980).

385 See *Hanson v. Denckla*, 357 U.S. 235, 253 (1958) (concluding that unilateral exercise of power of appointment within a state was not a sufficient contact when the issue was the validity of a trust agreement created in another state).

a state’s substantive laws should be higher than the threshold for determining
the location of the courthouse where the dispute will be tried, as litigants presumably
care far more about their substantive rights than about the venue in which those
rights will be vindicated. And the facts that may lead a court to conclude that a
choice-of-law decision is “not fundamentally unfair” may well be different than
those that determine whether a defendant has sufficient presence in a state to
subject it to personal jurisdiction there. But asking the same basic constitutional
question in both the personal jurisdiction and choice-of-law contexts—“what did
the parties reasonably anticipate?”—means that, unlike Wu’s bright-line rule or
current case law, both the employer and the employee get a fair shot at convincing
the court that due process permits the application of their favored state’s law.

B. Extra-Constitutional Benefits of Renewed Due-Process Scrutiny

Applying a two-step filtering rule for resolving choice-of-law questions in
noncompete cases does more than merely restore constitutional respectability to
court decisions (though that is a worthy goal). There are potential societal benefits
to be gained, as well, including greater transparency in employer-employee
relations, discouraging a sue-first mentality, and increasing public trust in the
neutrality and fairness of the courts as a place to resolve employment disputes.

1. Fostering Transparency

Renewed attention to due-process limitations in choice of law could produce a
number of socially beneficial side effects, beyond “merely” assuring that the
parties’ constitutional rights are vindicated. One possible effect of implementing a
due process filter in multistate noncompete disputes would be to encourage
employees and employers to be more candid with each other at the outset of their
employment relationship, or any other time the employer seeks to impose a
noncompete on an employee. If party expectations must, as a matter of
constitutional principle, feature prominently in a court’s choice of law, then it
would be to both the employer’s and employee’s advantage to disclose any
employee contacts with California—or any other state with restrictive policies
toward noncompetes—that might lead a reasonable employer to believe that the
noncompete may one day be subject to evaluation by a court in that state. The
employer could, for example, include an addendum to the employment agreement
that asks the employee to describe any relevant connections to a specified list of
states where the employer believes the noncompete might be subject to
invalidation.

The employee’s disclosure of such connections would then weaken any future
argument by the employer that application of those states’ laws is fundamentally
unfair. It would also incentivize the employer to think strategically about where its
main competitive interests lie and to craft its noncompetes with greater care to
survive scrutiny in those states.
Furthermore, the employee’s disclosure of contacts with the specified states would be helpful to the employer in evaluating whether it should, in fact, hire the employee or, if the employee has already been hired, whether to entrust the employee with greater responsibilities or access to information. If the employee legitimately sees the potential for relocation to a state like California, then the employee should give the employer fair warning. If the employee fears that full disclosure of his California ties or ambitions might diminish his chances of getting hired (or of getting a promotion), he of course has the right to conceal that information, but he does so knowing that the employer will be able to truthfully claim unfair surprise in a subsequent dispute over enforcement of the noncompete in California. If, having received fair warning of the employee’s California ties, the employer is sufficiently concerned that the employee might leave and take proprietary information or customer relationships with him to a competitor in California, where the noncompete is likely worthless, then the employer can make more informed decisions about the employee’s career path and access to key customers or information.

Restoring due process to at least the minimal respect it is owed under *Hague* would incentivize both employer and employee to be more forthcoming and, if necessary, to specifically negotiate the terms of a noncompete. Both employee and employer would enter into the relationship with a clearer understanding of the risks and rewards involved, rather than doing so with no clear expectations of whether the contract is worth the paper it’s printed on. In that way, the noncompete is more likely to reflect a fair bargaining process. Thus courts may be less likely to dismiss noncompete language as mere boilerplate or view noncompetes suspiciously as the product of overreaching or superior bargaining power by the employer. In short, noncompetes would become more tailored to individual circumstances, more predictably enforced, and fairer to both parties.

2. *Minimizing the Race to the Courthouse*

Another positive side effect of the due-process-as-filter approach is that neither the employer nor the employee would be punished for attempting to negotiate a palatable compromise when the employee announces his departure. As illustrated by *Manuel*, an employer that delays filing a declaratory judgment action to enforce the noncompete and instead seeks to negotiate with the employee or new employer regarding the scope of the employee’s new position—or even pauses to investigate whether it presents any competitive concerns—risks being haled into a hostile forum where the noncompete is all but certain to be invalidated. But if the parties trust that all courts will respect the due process limits on choice of law, then there would be no need to race to the courthouse. Neither side would have incentive at the time of termination to sue first and ask questions later. Realistically, we will never be able to abolish the informational asymmetry between employee and employer regarding changes in the employee’s career plans,

---

387 See supra notes 318–336 and accompanying text.
nor would we want to. But what we can do is reduce the difference it makes when it comes time to enforce a noncompete.

3. Enhancing Perceptions of Court Neutrality

Furthermore, remembering and respecting the role that due process plays in constraining courts’ choice of law would have positive effects on how the public perceives judicial decisions involving multi-state noncompetes. Courts open themselves up to criticism when they reflexively apply their own substantive laws or eccentric public policies to disputes that involve actors and conduct in more than one state. The results of cases like Application Group in California388 or Pokalsky or Keener in Georgia389 may be viewed as so outrageous that courts in other states may begin to employ extreme countermeasures as a means to defend their own residents against perceived overreaching by other states. We have already seen evidence of this in cases like Medtronic, Inc. v Advanced Bionics Corp390 and Advanced Bionics Corp. v. Medtronic, Inc.391

In Medtronic, a Minnesota trial court issued a TRO that not only prohibited a former Medtronic employee, Mark Stultz, from going to work for Advanced Bionics in California, it also prohibited Stultz and Advanced Bionics from pursuing their earlier-filed declaratory judgment action in California.392 Stultz, who had worked for Medtronic in Minnesota for several years, followed the standard employee playbook for evading noncompetes by filing suit in California on the same day he resigned from Medtronic in Minnesota.393 Stultz and Advanced Bionics asked the California court for a declaratory judgment and a TRO enjoining Medtronic from taking any action to enforce Stultz’s two-year noncompete.394 But when Medtronic learned there was a TRO hearing scheduled for the following day, it promptly removed the case to federal court (thereby delaying the hearing) and filed a parallel declaratory judgment action in Minnesota, invoking the Minnesota choice-of-law clause.395 Perhaps wary of how California courts had previously dealt with out-of-state corporations seeking to enforce noncompetes, the Minnesota trial court not only ruled that Minnesota law applied to the noncompete, it also took the extraordinary step of enjoining Stultz and Advanced Bionics from pursuing their California action or otherwise seeking to evade the Minnesota court’s jurisdiction.396 A few weeks later, the California state court, after remand by the federal court, appeared offended by the Minnesota court’s aggressive maneuver. It entered its own countervailing order, declaring that California law

388 See supra notes 182–210 and accompanying text.
389 See supra notes 260–316 and accompanying text.
391 59 P.3d 231 (Cal. 2002).
392 Medtronic, Inc., 630 N.W.2d at 445–46.
393 Advanced Bionics Corp., 59 P.3d at 233.
394 Id.
395 Id.
396 Id. at 234.
controlled the noncompete and enjoining Medtronic from pursuing judgment against Stultz and Advanced Bionics in Minnesota.397

With the parties and the lower courts having dueled to a draw, it was the California Supreme Court that finally blinked and handed a partial victory to Medtronic and Minnesota.398 Showing somewhat more deference to Minnesota than Minnesota showed to California, the California Supreme Court held that principles of judicial restraint and interstate comity counseled against the issuance of an injunction restraining Medtronic from maintaining parallel litigation in Minnesota.399 Acknowledging that such injunctions are extraordinary measures to be employed only in the most compelling circumstances, the California Supreme Court reversed the lower California courts and permitted the parties to pursue their separate actions to judgment in both states, despite the potential for conflicting results.400

The Minnesota appellate court, by contrast, found nothing improper in the trial court’s decision to enjoin Stultz and Advanced Bionics from litigating in California.401 Given the tendency of courts in states like California to apply their anti-noncompete policies broadly even to prospective employees who have not yet become California residents, we may see courts elsewhere starting to follow the Minnesota example of issuing anti-suit injunctions in a similar way to protect local employers from having their noncompetes invalidated in a hostile forum. One might reasonably question whether this is a positive development.

Eventually, if courts continue to ignore due process limits on choice of law, it may be necessary for Congress to intervene and enact a national choice-of-law statute governing enforcement of interstate noncompetes. Open debate in our national legislature might result in a broader consensus among the states as to when noncompetes should or should not be enforced across state lines. With the benefit of a clearer understanding of nationwide views on noncompetes, Congress could provide a national choice-of-law rule dictating the circumstances under which a contractual choice-of-law clause would control, and in the absence of a contractual term, which state’s law should govern. But until a national rule is adopted, it is imperative that courts be reminded of the limits imposed by due process and their obligation to avoid fundamentally unfair choice-of-law decisions when adjudicating disputes involving multi-state noncompetes.

397 Id.
398 Id. at 238 (“We hold that the trial court improperly issued the TRO enjoining Medtronic from proceeding in the Minnesota action. We also conclude, however, that the Minnesota action does not divest California of jurisdiction . . . .”).
399 See id. at 237 (“The comity principle requires that we exercise our power to enjoin parties in a foreign court sparingly . . . .”).
400 See id. at 237–38.
CONCLUSION

Principles of due process remain an important constraint on courts’ power to choose law in multistate disputes. The failure of courts in some states to recognize and observe applicable due process limits on choice of law, when combined with the “iron clad” rule of Full Faith and Credit for final judgments, has led employers and employees to engage in unfortunate races to the courthouse to either evade or enforce noncompetes. As a result, noncompetes are struck down merely because the employee was the first to file suit in a favorable forum, not because a careful, constitutionally proper choice-of-law analysis led to that result.

A renewed emphasis on due process considerations is essential if these cases are going to be decided by fair adjudications on the merits rather than which party obtained the earliest time-stamp on its complaint. By applying a due process “filter” as the first step in a two-step approach to choice of law in noncompete cases, courts can ensure that constitutional limits are respected and that parties are not incentivized to file suit prematurely in the hopes of securing a more favorable forum.

Finally, courts should ensure that the parties’ expectations at the time of contracting are honored when their noncompete is later litigated. This will have the additional advantages of promoting full and honest disclosure during contract negotiations and inspiring greater public confidence in the fairness and impartiality of the interstate judicial system.

Stephen M. Maurer*

I. INTRODUCTION

The nice thing about the open source (“OS”) phenomenon is that it changes faster than academics can study it. Ten years ago, most OS collaborations were organized around noncommercial, “fun” motives like altruism, hobbyist interest, and the like.1 By contrast, today’s OS projects are mostly commercial.2 Even if our theories and policy prescriptions were right ten years ago, the ground has shifted. This Article asks how judges and policymakers should manage the new commercial OS. In the process it uncovers a paradox. On the one hand, OS lets companies share costs. This potentially gives them the power to create far more software than ever before. But sharing also has a dark side. Since all OS companies offer consumers exactly the same shared codebase, no company can offer better shared software than its rivals. The result is a de facto cartel that suppresses competition and incentives to invest. Strangely, then, OS is self-limiting: Companies do share, but write far less shared code than they ought to.

This Article untangles the paradox of OS sharing and asks what judges and policymakers can do to help OS reach its full potential. Part II sets the stage by describing the rise of commercial OS over the past ten years. It also profiles a leading commercial OS collaboration (The Eclipse Foundation) and describes the various design issues that face such organizations. Part III examines how companies make investment decisions in OS, closed source (“CS”), and mixed OS/CS markets. Part IV uses these ideas to analyze when OS collaborations should and should not be permitted under the Sherman Act. It argues that OS collaborations can usually write licenses that satisfy the rule of reason. It also

* © 2012 Stephen M. Maurer. University of California Berkeley School of Law and Goldman School of Public Policy. 2607 Hearst Ave. Berkeley, CA. 94720-9657, smaurer@berkeley.edu. I wish to thank Bronwyn Hall, Deven Desai, Sebastian von Engelhardt, Andreas Freytag, Michael Katz, Jenny Kwan, Prasad Krishnamurthy, Mario Madden, Steve Mutkowski, Suzanne Scotchmer, Scott Selby, Jim Spohrer, Jeffrey Thompson, and Andrew Torrance for helpful conversations. This work was supported by generous grants from the Sloan Foundation, the National Science Foundation (Grant No. SES 20082737), and Nokia Corporation. Any oversights or errors are solely due to the author.

1 For a comprehensive review of these early theories, see Stephen M. Maurer & Suzanne Scotchmer, Open Source Software: The New Intellectual Property Paradigm, in 1 HANDBOOKS IN INFORMATION SYSTEMS: ECONOMICS AND INFORMATION SYSTEMS 285 (Terrence Hendershott ed., 2006).

See infra notes 7–9.
proposes two safe harbors in which OS collaborations should be presumed to be procompetitive. Part V examines the antitrust status of so-called “viral” licenses. It argues that very broad licenses (notably GPL)³ are unnecessarily restrictive and violate the Sherman Act. Part VI reviews how governments can use taxes, grants, and procurement policy to help OS sharing reach its full potential. Finally, Part VII presents a brief conclusion.

II. THE RISE AND RISE OF COMMERCIAL OPEN SOURCE

Academic understanding of OS has usually trailed the subject itself. When OS first appeared in the 1990s, scholars did little more than repeat the volunteers’ own narratives. In this telling, OS was a “movement” driven by psychological (“altruism”), political (“ideology”), or postmodern incentives (“the gift economy”) that defied conventional economic analysis.⁴ The more scholars studied OS, however, the less strange it seemed. Indeed, many OS incentives—for example a desire for education or to demonstrate one’s ability to potential employers—were transparently financial. By 2002 or so, economists had come to understand OS almost entirely in terms of familiar motives like education and signaling.⁵ At the same time, commercial incentives played a relatively small role and were seldom discussed.⁶

Worldwide OS production rose slowly during the phenomenon’s first decade. [Figure 1] Consider, for example, Riehle and Deshpande’s leading survey of open source activity on the Internet. As late as 2002, their sample contained just 500 OS projects.⁷ This changed dramatically, however, after software companies began

---

³ The “General Public License” or “GPL” is one of the oldest and best-known open source licenses. It was originally written by Richard Stallman for the GNU software project. GNU General Public License, WIKIPEDIA, http://en.wikipedia.org/wiki/GPL (last visited Feb. 6, 2012). For further details, see infra text accompanying notes 174–90.

⁴ For the best-known and most complete version of these arguments, see ERIC S. RAYMOND, THE CATHEDRAL AND THE BAZAAR: MUSINGS ON LINUX AND OPEN SOURCE BY AN ACCIDENTAL REVOLUTIONARY (rev. ed. 2001).

⁵ See, e.g., Josh Lerner & Jean Tirole, Some Simple Economics of Open Source, 50 J. INDUS. ECON. 197, 212–20 (2002); see also Maurer & Scotchmer, supra note 1, at 287–300 (OS communities include signaling and education as motives for creating OS products).

⁶ Lerner, supra note 5; Maurer & Scotchmer, supra note 1, at 287–300. Both sources treat commercial motives as a distinctly subsidiary phenomenon.

paying their software engineers to join OS collaborations as volunteers.\(^8\) In December 2000, IBM announced that it would invest a spectacular $1 billion in Linux.\(^9\) Four years later the total number of OS collaborations had grown more than fivefold. By 2006 it had doubled again to more than 4,500 projects.\(^{10}\)

Figure 1: Number of OS Projects, 1993–2007\(^{11}\)

---


\(^{10}\) Deshpande & Riehle, supra note 7, at 204.

\(^{11}\) Id.
Although exact figures are hard to come by, observers universally agree that most of the new projects are funded by for-profit companies that expect a clear dollars-and-cents return on their investment.\footnote{See Maurer & Scotchmer, supra note 1, at 286.} This implies that OS must be dramatically more commercial than it used to be. And indeed, industry observers estimate that commercial firms use three-quarters of today’s OS software.\footnote{Steve Lohr, Software War Pits Oracle vs. Google, N.Y. TIMES, Aug. 30, 2010, at B1 (remarking that days when OS was driven by sharing rather than profits “are long gone”).} Within this category, the great majority of projects follow business plans in which OS software is bundled with some proprietary product and sold to customers.\footnote{Barnett, supra note 8, at 1912 n.168.} This complementary good can either be a physical chattel like cell phones, a service like tech support, or closed source (“CS”) software.\footnote{There is also a second class of business models in which companies make OS software freely available to small users but charge fees to large companies. See, e.g., Dirk Riehle, The Commercial Open Source Business Model, 15 AM CONF. ON INFO. SYS. PAPER 104, 3 (2009), available at http://dirkriehle.com/wp-content/uploads/2009/04/commercial_v9_revision.pdf. Here, the main goal is less about code sharing than eliciting product ideas and bug fixes from lead users. Anecdotally, this model seems to be comparatively rare and we will ignore it in what follows.}

In light of these trends, it is hardly surprising to see leading scholars argue that the old, ideology-driven image of OS needs to be rethought.\footnote{For example, Professor Barnett remarks that the presence of paid employees is so pronounced that “[s]tandard characterizations of OSS development as the spontaneous coordination of . . . ideologically motivated volunteers . . . do not accurately describe at least the most successful applications in the current market.” Barnett, supra note 8 at 1895. Similarly, Professors Lerner and Schankerman observe that, “It is the conventional wisdom among many observers, analysts, and advocates of open source software that its development of [sic] is driven by a widely dispersed pool of voluntary contributors. But the fact is that corporations increasingly invest large amounts of money to finance open source development, both in terms of direct finance to other companies and through paying their employees to engage in such activity.” Josh Lerner & Mark Schankerman, The Comingled Code: Open Source and Economic Development 90 (2010).} Yet in many ways, the new commercial OS collaborations are even stranger than the traditional “fun” ones. In the “Old Economy,” companies built elaborate hierarchies to monitor and control their workers.\footnote{NAT’L BUREAU OF ECON. RESEARCH, 2 INNOVATION POLICY AND THE ECONOMY 8 (Adam B. Jaffe et al. eds., 2002).} In the “New Economy,” they encourage
employees to join self-governing OS collectives on company time.\textsuperscript{18} Furthermore, companies do little or nothing to monitor these activities.\textsuperscript{19}

The Eclipse Foundation. To get a better feel for commercial OS, consider the Eclipse Project. In April 1999, IBM began development of a new “Integrated Development Environment” for professional software developers.\textsuperscript{20} Two-and-a-half years later, IBM released Eclipse 1.0 as a conventional CS software product.\textsuperscript{21} Just one month later, however, it transferred the codebase—then valued at $40 million—to an OS “Eclipse Project” for further development.\textsuperscript{22} Legally, the key step was adopting a “Common Public License”\textsuperscript{23} so that users could freely use, modify and distribute the code to others.\textsuperscript{24} However, non-IBM participation remained disappointing.\textsuperscript{25} IBM reacted by surrendering still more control. This was done by moving the Eclipse project to an independent nonprofit corporation in February 2004.\textsuperscript{26} At this point, large numbers of outside developers began contributing to the code base.\textsuperscript{27} Today, Eclipse has roughly 150 member companies, and roughly one-third of these have contributed at least some software to the project since 2001.\textsuperscript{28}

OS sharing has yielded significant value for IBM. On the one hand, non-IBM members have supplied more than half of all contributions to the project codebase

\textsuperscript{18} Id. Companies’ readiness to support paid OS volunteers can be remarkably explicit. For example, the Eclipse Foundation’s rules require “Strategic Developer”-level corporate sponsors to assign at least eight full-time software developers to the project. “Strategic Consumer”-level members who contribute full-time developers pay lower dues. \textit{Types of Membership}, \textsc{ECLIPSE Found.}, \url{http://www.eclipse.org/membership/become_a_member/membershipTypes.php} (last visited Jan. 18, 2012).

\textsuperscript{19} \textsc{Lerner & Schankerman}, \textit{supra} note 16, at 41.

\textsuperscript{20} \textit{Eclipse Project Briefing Materials}, \textsc{ECLIPSE Found.} (2003), \url{http://www.eclipse.org/eclipse/presentation/eclipse-slides.pdf}.


\textsuperscript{22} \textit{Eclipse Project Briefing Materials}, \textit{supra} note 20.

\textsuperscript{23} \textit{Common Public License - v 1.0}, \textsc{ECLIPSE Found.} (Apr. 16, 2009), \url{http://www.eclipse.org/legal/cpl-v10.html}.

\textsuperscript{24} Users can even distribute the modified code under their own CS licenses as long as they make Eclipse’s underlying source code freely available. \textit{Eclipse Public License - v.1.0}, \textsc{ECLIPSE Found.}, \url{http://www.eclipse.org/legal/epl-v10.html} (last visited Jan. 17, 2012).

\textsuperscript{25} See Spaeth et al., \textit{supra} note 21, at 419.

\textsuperscript{26} Id. The new body is called the Eclipse Foundation. \textit{See also Bylaws of Eclipse Foundation, Inc.}, \textsc{ECLIPSE Found.} (Aug. 15, 2011), \url{http://www.eclipse.org/org/documents/Eclipse%20BYLAWS%202011_08_15%20Final.pdf}.

\textsuperscript{27} See Spaeth et al., \textit{supra} note 21, at 419.

\textsuperscript{28} A complete list of Eclipse’s member companies can be found at \textit{Explore the Eclipse Membership}, \textsc{ECLIPSE Found.}, \url{http://www.eclipse.org/membership/exploreMembership.php} (last visited Jan. 17, 2012). The list includes various entities (e.g., publishing houses) that seldom write code. For a list of companies that have actually made code deposits, see \textit{Eclipse Dashboard}, \textsc{ECLIPSE Found.}, \url{http://dash.eclipse.org/dash/commits/web-app/summary.cgi} (last visited Jan. 17, 2012). More than two dozen nonmember companies have also made code deposits over the life of the project.
since 2007. On the other hand, non-IBM companies have incorporated Eclipse into a host of new products. This has made Eclipse the leading tool for developing software for mobile devices like cell phones, GPS units, and MP3 players. Eclipse is also widely used to develop enterprise software for data-intensive problems like train scheduling and banking.

Why do companies participate? The “business advantages,” according to Eclipse, depend on cost sharing. At first blush, this rationale may sound homely to readers accustomed to claims that OS produces software more efficiently than CS methods, is better at finding and fixing bugs, or generates streams of clever ideas from users. But consider the math: OS advocates would be overjoyed if they could cut costs or increase quality by, say, 20 percent. From a business standpoint, this means that companies could now pay eight workers instead of ten. At first glance, cost sharing, in which ten workers still do the work of ten, seems much less wonderful. But consider what happens when just two companies share. Now each partner pays just five workers. And these savings keep growing for each

29 E-mail from Mike Milinkovich, Executive Director, Eclipse Foundation, Inc., to author (Jan. 4, 2012) (on file with author).
31 Taft, supra note 30.
33 More precisely, Eclipse says that it allows “individuals and companies to collaborate on projects that would be difficult to achieve on their own.” Eclipse Public License (EPL) Frequently Asked Questions, ECLIPSE FOUND. http://www.eclipse.org/legal/eplfaq.php (last visited Jan. 17, 2012). Eclipse adds that its model has the further “technical advantage of turning users into potential co-developers.” Id.
34 RAYMOND, supra note 4, at 129–32.
35 Id. at 33–36.
firm that joins the collaboration thereafter. On any reasonable assumption, then, we expect sharing to dominate the business case for open source. What kinds of companies join Eclipse? Table 1 profiles the top ten companies that have donated software to the Project from 2001 to 2010. These can be conceptually divided into four groups:

*IBM.* Despite sharing, it remains true that IBM has contributed far more effort than any other company. This has included supplying one-fourth (24.3%) of all programmers and two-fifths (42.6%) of all software deposits between 2001 and 2010. Even so, IBM’s glass is more than half full: After all, nearly three-fifths of all deposits were paid for by someone else.

*Other Dominant Companies.* The next four companies in Table 1—Oracle, Intalio, Cloudsmith and Actuate—contributed an additional 17 percent of all software deposits. Together with IBM, this shows that about 60 percent of the total Eclipse effort between 2001 and 2010 has come from five large players. To this point, the picture is not too different from a conventional joint venture.

*Small Companies.* Eclipse’s records show that at least eighty-six additional companies have contributed “commits” since 2001. While the data are uncertain, these donations probably account for roughly 30 percent of all commits. Thirty-four of these companies contributed at least 10,000 commits each.

---

38 Of course, these estimates assume that OS sharing is perfect. This is plainly untrue. Most notably, game theory suggests that OS collaborations should suffer from so-called “game-of-chicken” effects in which each participant delays creating software in hopes that somebody else will do the work for them. In equilibrium this reduces, but does not eliminate, OS sharing. See, e.g., Justin Pappas Johnson, *Open Source Software: Private Provision of a Public Good,* 24 J. ECON. & MGMT. STRATEGY 637, 641–44 (2002). 39 There are numerous ways to measure software donations. This article relies on the number of “commits,” i.e., separate instances in which a programmer deposits new or revised code in the project repository. 40 Eclipse data from 2001 through 2010 provided by Mike Milinkovic and Wayne Beaton (Eclipse Foundation) and analyzed by Severin Weingarten (Friedrich Schiller University, Jena) [hereinafter Weingarten data]. Weingarten describes how this data was compiled in Severn Weingarten, *On the Economics of Commercial Open Source Software: the Case of the Eclipse Project* (Mar. 31, 2012) (unpublished Master’s thesis, Fredrich-Schiller University Jena) (on file with Utah Law Review). 41 Weingarten Data, *supra* note 40. 42 The Eclipse data record commits by hundreds of individual programmers with no known employer. This figure is misleading, however, because many “[p]eople categorized under ‘individual’ may be associated with a company, just not an Eclipse member company that has signed a Member Committer Agreement.” *Dash Project/Commits Explorer,* ECLIPSE FOUND., (Dec. 20, 2010), http://wiki.eclipse.org/Dash_Project/Commits_Explorer.
Traditional OS Participants. The remaining Eclipse deposits come from sources that resemble traditional, noncommercial OS communities. These include nonprofits, academic organizations, and hundreds of individuals. While these participants account for relatively few deposits—probably fewer than 10 percent—this may understate their value as lead users facing unusual computing problems. Similar user communities are known to be a valuable source of bug patches and new product ideas.

Sharing is not the whole story, of course. Companies must also find a way to recover their investments. Table 1 profiles the top ten corporate software donors since 2001. Strikingly, all ten companies earn their living by making and selling CS products like hardware, software, or IT consulting services. Directly or indirectly, these sales pay for Eclipse.

Making Sharing Work: Collaboration Architecture. The argument that Eclipse and other OS collaborations are driven by sharing deserves a closer look. After all, CS firms can also share costs by signing joint venture agreements. Why should OS work any better? One partial answer is that ordinary joint ventures are deeply flawed. The main reason is that members usually find it difficult to monitor what their partners—or even their own employees—are doing. This leads to recurring principal-agent problems in which participants try to (a) hijack joint venture research in directions that favor themselves, or (b) shirk work altogether. On the other hand, OS methods have a well-deserved reputation for transparency. Could this cure the principal-agent problems that afflict CS sharing? IBM and its Eclipse partners are clearly betting that it will.

To estimate the size of this effect, consider that just 10 percent of the programmers listed as “individuals” contributed three-quarters (74%) of all “individual” deposits between 2001 and early 2010. Eclipse Dashboard, ECLIPSE FOUND., http://dash.eclipse.org/dash/commits/web-app/summary.cgi (last visited May 3, 2012). Furthermore, each of these contributors donated at least 10,000 commits apiece. Such effort strongly suggests a corporate sponsor. We therefore conclude, somewhat speculatively, that unpaid volunteers supplied fewer than 10 percent of all Eclipse commits between 2001 and 2010.

43 Weingarten data, supra note 40.
44 At least five academic and nonprofit institutes donated commits from 2001 to 2010. The largest were CEA LIST (1.0%) and INRIA (0.2%). Weingarten data, supra note 40.
45 See VON HIPPEL, supra note 36, at 93–95.
46 This pattern is similarly evident in the Linux collaboration as well as various OS projects (e.g. Android) that write software for use in mobile telephones. Readers can find a detailed statistical description in Barnett, supra note 8, at 1902–10, 1924.
As usual, the devil is in the details. The Eclipse Foundation tries to address transparency issues at two levels. The first set of solutions relates to high-level decision-making and would not be out of place in a traditional joint venture. Overall governance is committed to the Eclipse Foundation’s Board of Directors which includes (a) representatives of fifteen large corporate sponsors, and (b) six elected members representing the committer and add-on communities.\textsuperscript{48} The Board sets the Foundation’s operating budget, decides which new software projects to develop, and can also amend the Foundation’s OS license.\textsuperscript{49} Because Board membership is diverse, no single interest group—including IBM and other large players—has nearly enough votes to hijack the collaboration.\textsuperscript{50}

The Eclipse Foundation’s second set of strategies for ensuring transparency takes a page from traditional OS methods. Beneath the Board level, companies play little or no formal role in prioritizing and performing work.\textsuperscript{51} Instead, volunteers are encouraged to discard their corporate affiliations and interact as individuals. This is reinforced by governance rules that vest leadership in individual programmers (“committers”) whose status has nothing to do with company affiliation and follows programmers when they change employers. Committers are said to be chosen on the basis of “meritocracy,” “contributions,” and “peer acclaim.”\textsuperscript{52}

Superficially, consigning production to a crowd of unsupervised programmers sounds like a formula for disaster. On the other hand, we have seen that traditional outside monitoring probably would not work in any case. This makes it sensible to rely on collaboration insiders who know the project first-hand. This strategy can only succeed, however, if insiders’ incentives are aligned with their employers’ goals. Commercial OS does this in several ways:

\textit{Efficiency Wage.} Probably the most obvious strategy is for employers to pay insiders above-market wages and then fire them if the collaboration

\textsuperscript{48} The current Board includes representatives from SAP, Talend, Actuate, OBE, IBM, Innoo, Oracle, CA Technologies, BREDEX, itemis, Sonatype, and six individuals elected by the committer and add-on provider communities. \textit{Eclipse Foundation Board of Directors}, \textsc{ECLIPSE FOUND.}, http://www.eclipse.org/ org/foundation/directors.php (last visited Jan. 17, 2012). Professor Barnett has emphasized that many other OS collaborations including Linux, Ubuntu, GNOME, Firefox, and Apache similarly rely on independent boards to limit opportunism by individual sponsors. Barnett, \textit{supra} note 8, at 1899–1900.

\textsuperscript{49} The Board also hires and fires the Foundation’s director, sets membership fees, terminates and reinstates members, and sets IP and antitrust policy. \textit{See Membership Rights}, \textsc{ECLIPSE FOUND.}, http://www.eclipse.org/membership/become_a_member/memberRights.php (last visited Jan. 17, 2012).

\textsuperscript{50} As in a private corporation, fundamental changes to the bylaws and other ground rules must be approved by three additional classes of voting members along with programmers who have achieved leadership (“Committer”) status within the project. \textit{Id.}


\textsuperscript{52} \textit{Id.} § 4.1.
fails to deliver value. These “efficiency wages” create a strong incentive to please employers.\(^{53}\) They also work indirectly by encouraging lower-level insiders to work hard so that they, too, can receive efficiency wages in the future.

**Collaboration Assets.** Insiders can usually extract rents from the collaboration over and above the value of their labor.\(^{54}\) These assets include (a) the added productivity that collaboration volunteers gain from a long history of working together, (b) the ability to extract programming effort from individual volunteers pursuing traditional OS incentives like education or signaling,\(^{55}\) and (c) money and in-kind contributions from corporate sponsors to the collaboration as a whole.\(^{56}\) By definition, these assets and rents are only as durable as the collaboration itself. This gives insiders a powerful incentive to keep sponsors happy.

**Self-Policing.** Because OS collaborations are transparent to insiders, no single insider can shortchange a sponsor without the others noticing. But why should the other insiders permit this? After all, their income also depends on the collaboration’s reputation for delivering value. This gives each insider a direct stake in seeing that all sponsors receive value.

Are any of these arguments correct? The jury is still out.\(^{57}\) Nevertheless, companies have clearly found them sufficiently persuasive to justify ambitious

---


\(^{54}\) Lerner \\& Tirole, \textit{supra} note 5, at 217.

\(^{55}\) Individual programmers often donate unpaid labor to commercial OS collaborations as a way of advertising their skills to prospective employers. See, e.g., Lerner \\& Tirole, \textit{supra} note 5, at 217. Here, the prospect of a future job acts as a prize to induce current effort.

\(^{56}\) Typical examples include paying for workshops and conferences, purchasing hardware, and relicensing existing CS code under the GPL. Joel West \\& Siobhan O’Mahoney, \textit{Contrasting Community Building in Sponsored and Community Founded Open Source Projects}, 38 \textit{Haw. Int’l Conf. on Sys. Sci.} 8 (2005), available at http://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1001&context=org_mgmt_pub.

\(^{57}\) The steep disparity in effort between the Eclipse Foundation’s first four corporate contributors and the remaining 178 companies suggests that OS sharing is highly imperfect. This may be because companies intentionally delay effort in hopes that someone else will do the work. See, e.g., Johnson, \textit{supra} note 38, at 641–44 (presenting model in which OS developers balance the expected costs and benefits of writing new programs against the chance to free-ride if someone else does). Alternatively, small companies may sometimes withhold effort knowing that their contributions will be miniscule in any case.
experiments like Eclipse. Until we learn differently, then, it is only prudent to set policy on the assumption that commercial OS sharing is here to stay. This makes it important to understand how OS affects the economy along with the challenges it poses for judges and policymakers.

III. THE OS/CS TRADEOFF: COMPETITION, SHARING, AND CARTELIZATION

The rise of commercial OS poses a complex investment problem for companies. On the one hand, OS sharing dramatically cuts each company’s individual software development costs. On the other hand, shared code strengthens its competitors. This can reduce profits to the point where OS no longer makes sense. Finally, each OS company offers consumers the same shared software as every other OS company. This leads to a de facto cartel that shelters companies from competition and limits their incentives to invest in software.

This Part explores OS economics in detail. It begins by asking how companies in a traditional all-CS industry decide how much to invest in software. The Article then explores how sharing changes this analysis for an all-OS industry before finally turning to mixed markets in which OS and CS companies compete with one another. It demonstrates that strong CS competition can force OS collaborations to produce far more software than a pure-OS industry would. However, this Part also shows that this condition seldom occurs in practice. Instead, market forces tend to produce information technology (“IT”) ecosystems in which there are not enough CS firms to overcome the cartel effect. This systematically reduces the effort that OS firms would otherwise invest in shared code production. The resulting gap between OS software’s potential and its actual performance poses a central challenge for judges and policymakers.

A. How CS Firms Invest

This Part begins by asking how companies in a traditional, all-CS industry decide how much to invest in software. Consider, for the sake of definiteness, a simple scenario in which companies create software in Year One, and then manufacture and sell a product that contains the software in Year Two. In both cases, we should expect (a) large companies to shoulder more than their “fair share” of effort, and (b) total collaboration effort to be smaller than it otherwise would be.

58 The Eclipse Foundation goes out of its way to warn prospective members that it “provides the same opportunity to all” and that “there are no rules to exclude any potential contributors which include, of course, direct competitors in the marketplace.” Eclipse Development Process, supra note 51, § 2.1.

Two product can be a chattel (for example, a cell phone), a service (for example, technical support), or CS software. For simplicity, this Article assumes that the quality of the Year Two product depends entirely on how much software it contains. Readers should note that this scenario includes two business decisions: (a) How much to invest in shared software development (“quality”) in Year One, and (b) How many goods and services to manufacture and sell in Year Two. Each of these decisions provides a separate and distinct channel for companies to compete.

Note that a CS company will only invest if its expected profit from sales in Year Two is large enough to cover both the costs of making the product in Year Two and its software investment in Year One. Consider its Year Two profits first. These can be calculated using methods that are routinely taught in college economics courses. The result is more or less what you would expect. As competition increases, each CS firm produces and sells more goods. Because prices fall, however, its total profits decline. We will refer to this dynamic as “Year Two competition” in what follows.

The Year Two solution, in turn, sets up the Year One analysis. While Year Two profits clearly depend on the quantity of goods sold, they also depend on the quality of those goods, that is, how much software is written in Year One. Detailed calculation shows that CS companies’ Year Two profits (a) increase with the amount of software they write in Year One, and (b) decline when their rivals create competing software. We will refer to these effects as “Year One competition.”


Formally, the only limitation is that the product must be “excludable,” i.e., that sellers can control who uses it. Physical goods are clearly excludable since ordinary property law lets sellers control access. Similarly, service-providers can stop supplying users any time they choose. The case is subtler for software that consumers can copy at little or no cost. Despite this, companies can usually achieve reasonable excludability by using technical protections, invoking intellectual property law, and/or withholding source code and tech support.

Briefly, each company calculates how many goods “x” to manufacture for every possible industry-wide output “X.” The problem is then solved by finding the (at most) handful of cases in which the calculated x’s for every company add up to a self-consistent X. See, e.g., WALTER NICHOLSON, MICROECONOMIC THEORY: BASIC PRINCIPLES AND EXTENSIONS 639–44 (6th ed. 1995).

von Engelhardt & Maurer, supra note 59, at 14–24.

Id.
Figure 2 summarizes these arguments. Here, the horizontal axis denotes how similar these companies’ Year Two products are to one another with the degree of similarity increasing from left (completely different) to right (completely interchangeable). Readers should note, however, that product similarity also increases competition. For this reason, we will usually think of the horizontal axis as an index of Year Two competition. Thus we can interpret the dashed line as showing how much software a typical CS company creates in Year One as a function of expected competition in Year Two. Production is highest on the left-hand side where companies expect to face little, if any, Year Two competition. In this case, the Year Two monopoly acts like an additional intellectual property incentive, giving the company maximal freedom to raise prices if and when consumers demand more of its software-enhanced products. Conversely, CS software production falls as competition increases. This is because the company knows that its ability to charge higher prices will be limited by the existence of similar, equally improved competing products.

But this is not the whole story. Instead, CS software production levels off and even starts to rise again in right-hand portion of Figure 1. How is this possible? The answer is quality competition. When products are nearly interchangeable, even tiny quality differences can persuade consumers to switch from one company to another. This leads to a kind of arms race in which each company writes large amounts of software to keep its rivals from stealing customers. Like real arms races, these investments cancel each other out, leaving each company no better off—and considerably poorer—than it was before. This process is painfully

---

64 Adapted from id. at 16.

65 Competition also depends on the number of firms in the market. This effect is qualitatively similar to product similarity and we will normally ignore it in what follows.
familiar in Silicon Valley where CEOs frequently complain about razor-thin margins for “commoditized” (that is, generic) software.66

It is easy to understand why Silicon Valley CEOs should see “commoditization” as a disaster. Their business models center on being the first—and for a time, the only—company to deliver the next breakthrough product (“Killer App”) to consumers.67 These temporary monopolies are far more lucrative than creating new versions of existing products.68 But what should the rest of us think? Outside Silicon Valley, most products are already commoditized—that is, they are produced by multiple competing companies. But in that case, our usual antitrust intuition suggests that commoditization is good for society. And indeed, generic economic models of CS production suggest that the parameters under which commoditization would hurt industry more than it helps consumers are extremely narrow.69 While it is theoretically possible for commoditization to hurt society as a whole, such cases are probably rare and may not exist at all.70

B. How OS Firms Invest

Now consider how a company in an all-OS industry invests. Start by noticing that once the software exists, an OS company’s Year Two choices are the same as a CS company’s. This means that any differences must relate to the OS company’s Year One strategy. Here, sharing introduces two important changes. On the one hand, it almost always reduces per-firm costs to below the CS case. On the other, OS sharing also dilutes the payoff from investing. This is because more software not only enhances the investing company’s own products, but also makes its competitors’ products more attractive.

To see how these effects interact, consider the solid line in Figure 1. OS investments are a no-brainer on the left hand side, where competition is weak. This

---

66 See, e.g., Ian Murdock, Open Source and the Commoditization of Software, in OPEN SOURCES 2.0, at 91 (Chris Dibona et al. eds., 2006), available at http://commons.oreilly.com/wiki/index.php/Open_Sources_2.0/Open_Source:_Competition_and_Evolution/Open_Source_and_the_Commoditization_of_Software. Commoditization appears to be a natural outcome for models based on Cournot competition. Interestingly, models based on Bertrand competition seem to lack this feature. (Professor Prasad Krishnamurthy, personal communication). This sort of disagreement between models is not unusual. The fact that Silicon Valley executives routinely complain about “commoditization” suggests that Cournot models are on the right track.

67 See Murdock, supra note 66, at 93.

68 They may even be socially justified if the lure of large rewards induces companies to develop new innovations earlier or more reliably. See, e.g., SUZANNE SCOTCHMER, INNOVATION AND INCENTIVES 112 (2004).

69 von Engelhardt & Maurer, supra note 59, at 22–24.

70 The reason is that companies in concentrated industries can usually take steps to mitigate commoditization. Here, the normal strategy is for each company to optimize its product for a different submarket. This suggests that the welfare losses associated with extreme commoditization will only occur in those rare industries where technical or market considerations force companies to make nearly identical products.
is because companies can split costs without threatening each other’s revenues. What could be better? On the other hand, revenues decline as products become more interchangeable and quantity competition increases. This explains why OS companies, like their CS counterparts, produce less and less software as competition increases along the right-hand side of Figure 1.

Even so, there is a surprise. Unlike the CS case, OS output never encounters a commoditization crisis in which an arms race between firms revives incentives to produce software. Instead, OS software production never stops falling. How can this be? The reason is sharing. By definition, no company in an all-OS industry can offer consumers better shared software than any other company. This suppresses quality competition much as a formal cartel would. (In fact, the incentives to invest are actually worse than a cartel.) While OS companies invest in software, then, they stop writing sooner than they would if quality competition existed. This leads to the usual monopoly result—an undersupplied product.

---

71 Actually, one can imagine something better. As Dr. Schmidke points out, many goods are complements so that improvements to one product increase demand for both. In this situation, firms that invest in R&D confer benefits not only on themselves but also others. Since only the former are compensated, this leads to underinvestment. OS sharing gets around this externality by letting companies subsidize each other’s products. Richard Schmidtke, Private Provision of a Complementary Public Good 8–9 (CESifo, Working Paper No. 1756, 2006).

72 von Engelhardt & Maurer, supra note 59, at 8.

73 This statement assumes that companies cannot write separate, proprietary modules that “turbocharge” the shared codebase’s performance. This is reasonable. First, OS licenses typically contain “viral” terms that limit how closely proprietary modules can interact with the underlying codebase. Second, letting companies divert improvements into proprietary modules would immediately cripple OS sharing. Indeed, sharing stops entirely in our model. Knowing this, we would expect consciously parallel conduct to avoid the practice. Finally, the idea of using separate proprietary modules to turbocharge performance is often impractical on marketing and technical grounds. We return to these issues infra Part V.

74 Ironically, OS collaborations actually produce less software than an explicit cartel would. The reason is that a well-designed cartel maximizes industry-wide profits. By contrast, participants in an OS collaboration only care about their own profits and ignore their work’s impact on the profits of other members. This reduces their incentive to produce code in the first place. Maurer & Engelhardt, supra note 59, at 18, 39–40.

Readers should note that Dr. Llanes and his co-authors do not find cartel effects in their models of OS/CS investment. Llanes & de Elejalde, supra note 59, at 34–35; Ramon Casadesus-Masanell & Gaston Llanes, Mixed Source 5–6 (NET Institute, Working Paper No. 09-06, 2009). This seems to be an artifact of the authors’ assumption that companies make their Year One and Year Two decisions in one single instant. Importing this assumption into the Maurer and von Engelhardt model likewise suppresses cartel effects. While mathematically convenient, the Llanes et al. decision fails to capture the very different lead times for writing software compared to making physical goods.

75 More specifically, companies stop investing when their expected profit is largest. Here, the basic intuition is that (a) consumer willingness to pay (and hence revenue) is approximately linear in program size, but (b) development costs increase steeply for large
We have come to a paradox. On the one hand, OS companies’ ability to share costs means that they can produce far more software than any CS competitor. On the other, this same sharing creates a de facto quality cartel that suppresses OS companies’ incentives to invest in the first place. We therefore expect that OS collaborations facing stiff Year Two competition will invest less effort in writing software than a CS company would. This will still be good for society in the usual case where the shared, collaboration-wide OS effort produces more software than any CS company can fund on its own. At the same time, it remains true that OS companies would have invested still more absent the cartel effect. For OS-enthusiasts and policymakers, the glass is only half-full.

C. Can Competition Fix the Problem?

We have seen that OS promotes sharing but suppresses quality competition among collaboration members. On the other hand, antitrust law reflects the belief that competition cures most ills. Can CS companies deliver the missing quality competition? Simple calculations are encouraging. For example, generic models suggest that IT industries deliver maximum benefits when 15–20 percent of companies practice OS and the rest are CS. [Figure 3] Reassuringly, this

software projects. This suggests that OS companies will normally stop writing software once development costs start to increase faster than expected revenues. CS companies do not have this option. Instead, quality competition forces them to continue investing in new software despite “commoditization.”

76 In retrospect, this should not surprise us. Scholars have worried since the 1980s that conventional CS sharing (i.e. joint R&D ventures) could suppress competition in innovation. This literature is often strikingly similar to the current analysis. For example, Professor Katz argued in 1986 that R&D joint ventures eliminated duplication but also reduced members’ incentives to invest. Michael L. Katz, An Analysis of Cooperative Research and Development, 17 RAND J. ECON. 527, 527–543 (1986). The latter effect could, however, be mitigated where (a) joint venture members’ downstream products faced minimal competition, or (b) the joint venture was just one of several R&D programs operating in the industry. Id. Katz later served as Deputy Assistant Attorney General for Economic Analysis from 2001 to 2003. Some courts have similarly expressed concern that joint ventures could operate to suppress R&D incentives. See, e.g., Addamax Corp. v. Open Software Found., Inc., 888 F. Supp. 274, 282–86 (1995) (joint R&D venture could potentially cartelize software production by participants), aff’d, 152 F.3d 48, 52 (1st Cir. 1998).


78 von Engelhardt & Maurer, note 59. Careful readers will note that OS output in Figure 3 actually peaks when OS companies constitute about 10 percent of the market. Social benefits would also peak at this point if they depended solely on the amount of OS software produced. In fact, however, CS software also generates benefits by serving consumers who prefer to pay low prices for low quality software. This explains why social benefits for the combined (OS + CS) market only peak when 15–20 percent of all companies pursue OS business plans. Id. at 27.
percentage is more or less identical to the rule of thumb that the Department of Justice usually invokes when measuring cartelization threats, for example, in mergers.79

Figure 3: Software Production vs. Percent of Industry Members Adopting OS Methods.80

Finally, Figure 3 assumes a scenario in which a single OS collaboration competes with multiple CS companies. However, one can imagine still more favorable scenarios in which several OS collaborations compete with one another. Such situations promise the same level of competition as Figure 3 with radically increased sharing. For example, the Eclipse collaboration currently competes with both proprietary (JDeveloper) and open developer tools (NetBeans, IntelliJ).81 Similar examples in which open products compete against each other include Apache Harmony versus OpenJDK (Java runtimes),82 Pentaho versus JasperSoft (business reporting software),83 and Jetty versus Tomcat (web servers).84 The

80 von Engelhardt & Maurer, supra note 59, at 21.
picture for Content Management Systems—in which OS collaborations Plone, Drupal, PostNuke, and Joomla all compete with one another—is even more encouraging. The main drawback is that some or all of these projects could fall by the wayside over time.

D. Market Imperfection No. 1: Lock-In

We have seen that the benefits of sharing are largest in markets where CS companies outnumber their OS competitors by roughly four to one. Do market forces guarantee this ratio? The answer is almost certainly “no.” One reason is that infant industries will often be born in an All-OS or All-CS state. Such industries can become “locked in” so that mixed OS/CS markets never emerge. To see why, consider an infant industry that consists entirely of CS companies. Clearly, OS companies will only enter the market if they expect to earn a profit. However, detailed calculation shows that this may not be possible in markets where Year Two competition is strong. For this reason, we expect many All-CS markets to exclude OS indefinitely. In a few cases, incumbents may deliberately select CS so that would-be entrants cannot take advantage of low-cost OS codebases to enter the market. Here, we expect companies to choose CS despite the fact that OS sharing would yield higher profits in the short run.

Lock-in can also occur in All-OS markets where Year Two competition is too weak for would-be CS entrants to earn a profit. Here too, we expect market forces to block the emergence of more balanced markets in which OS and CS companies compete with one another.

E. Market Imperfection No. 2: Too Many OS Firms

The models that predict lock-in require fairly special parameters. This implies that mixed OS/CS markets should eventually emerge in many, if not most, cases. Even when this happens, however, the market will encourage too many

---

85 I am indebted to the Eclipse Foundation’s Mike Milinkovich for these examples.
86 See supra text accompanying notes 76–78.
87 This result is more or less inevitable where the young industry consists of just one or two companies. Because OS sharing needs at least two companies, we naively expect more All-CS industries than All-OS ones.
88 von Engelhardt & Maurer, supra note 59, at 28, 32–36.
89 Id. at 15, 32–36. The reason for this result is apparent from Figure 2, which shows that OS collaboration investments go to zero as quantity competition increases.
90 Id. at 14–16, 32–37.
91 Id. at 19–22. The logic is reminiscent of predatory pricing, in which firms deliberately incur short-term losses in order to charge higher prices once their rivals have been driven from the market.
92 Id. at 33–36.
firms to practice OS. The reason, once again, is the cartel effect, which suggests that any given group of OS companies will normally be more profitable than the same number of CS companies. This implies that most companies will adopt OS business models. Figure 4 (dotted line) shows what happens in the case where entry continues until the profits of both OS and CS companies fall to zero. For most values of competition, the number of OS companies hovers at roughly 60–70 percent. However, these companies tend to be small so that OS’s total market share remains low. Conversely, the CS sector is dominated by a handful of large companies that serve most of the market. This pattern is frequently seen in real software markets.

This systematic over-supply of small OS firms has profound implications. Ideally, policymakers would like markets to occupy the left-hand side of Figure 3 where software output is greatest. Because of the cartel effect, however, we expect real OS collaborations to fall on the right-hand side and produce less software than

---

93 See supra text accompanying notes 69–74.
94 OS companies’ small average size follows from the cartel effect. On the one hand, high profits increase the number of companies that practice OS. On the other hand, the cartel effect suppresses output. This means that average per-firm output must be small.
96 von Engelhardt & Maurer, supra note 59, at 31.
their CS competitors. This anemic result is still better than having each OS company write wastefully duplicative software. Nevertheless, the situation is far from ideal. Parts IV through VI will evaluate judges’ and policymakers’ options for closing this gap.

F. Beyond Simple Models.

So far, we have analyzed OS output and the cartel effect using deliberately generic assumptions. Nevertheless, details matter. We close this section by asking how more complicated—but still plausible—assumptions could change our understanding:

1. Inefficient Sharing

We have assumed that OS sharing is efficient. Though plausible, this remains to be seen. Less efficient sharing would reduce the benefits of OS while leaving the cartel effect unchanged. This suggests that judges and policymakers should scrutinize OS more closely where sharing is imperfect.

2. Idiosyncratic Consumers

We have assumed that products compete on both quantity and quality. However, software consumers sometimes display loyalty to particular products despite large differences to price and/or quality. Here, quality competition will likely be weak in any case. In these circumstances, policymakers and judges

---

97 Experience with real world markets confirms this intuition. Despite considerable argument, there is no clear evidence that OS code performs substantially better than equivalent CS products. The relevant literature is discussed at length in, for example, Maurer & Scotchmer, supra note 1, at 285–322.

98 Economists usually refer to the wasteful duplication associated with conventional CS production as “business stealing” or “me-too products.” See, e.g., Joachim Henkel & Eric von Hippel, Welfare Implications of User Innovation, 30 J. TECH. TRANSFER 73, 78–79, 83 (2005).

99 See supra text accompanying notes 56–94.

100 See supra text accompanying notes 55–56.

101 See supra text accompanying notes 56–94. Formally, the assumption leads to models that feature classically linear demand curves. Maurer and von Engeldhardt, supra note 59, at 10.

102 Brand loyalty is commonly represented by rank-ordering products along an imaginary line according to their attributes. These models are usually constructed so that consumers who prefer a given set of attributes will often select “nearby” products over higher-quality, but more distant alternatives. For an application of these ideas to commercial open source, see Llanes & de Elejalde, supra note 59.

103 Detailed calculations by Doctors Llanes and De Elejalde confirm this intuition for a model containing two technology sectors. Id. at 10–20.
could reasonably conclude that any additional damage from cartel effects is not worth worrying about.

3. Multiple Technologies

This Part has assumed for simplicity that product quality is derived from a single, indivisible technology called “software.” In fact, quality usually depends on multiple technologies. This suggests that companies may choose to develop some technologies using OS sharing and others using CS methods. If so, the benefits of OS sharing—but also the drawbacks associated with OS cartels—will be diluted so that the basic OS/CS choice becomes less important.104

4. Product Differentiation

We have assumed that OS sharing reduces the wasteful duplication inherent in having each company write its own software.105 However, duplication also increases the odds that at least one program will arrive sooner—or perform better—than the others. Moreover, duplicative programs may not be wasteful if each is tailored to a different user group. Judges and policymakers should be careful to look for these effects before they accept the benefits of OS sharing at face value.

5. Limited Entry

We have made the conventional assumption that new companies can and will enter the market until incumbents’ profits fall to zero.106 If entry is slow, however, it may make more sense to assume that the number of companies is fixed and ask how many will choose OS business plans. Fortunately, this new assumption leads to qualitatively similar results. We still expect the cartel effect to make OS

---

104 This is indeed what happens in Doctors Llanes and De Elejalde’s analysis when they split their technology sector into two pieces, only one of which is suitable for OS sharing. Id. The case may be different where technologies splinter into fragments that are too small for efficient R&D sharing so that each company takes responsibility for whichever project(s) offer the most value to its business. As Professor Henkel has emphasized, this boosts total investment and makes OS sharing more attractive. Joachim Henkel, The Jukebox Mode of Innovation — a Model of Commercial Open Source Development 10–11 (Danish Research Unit for Industrial Dynamics, DRUID Working Paper No. 06-25, 2006), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=578142. Judges and policymakers should be especially tolerant of cartel effects in these circumstances.
105 See supra text accompanying notes 55–74.
106 See supra text accompanying notes 55–95.
companies systematically more profitable and therefore more numerous than CS companies.\textsuperscript{107}

Based on this discussion, it is reasonable to think that more complex fact patterns will sometimes dilute or qualify our arguments. The basic themes, however, are robust. These include the benefits of OS sharing, the OS cartel effect, and the resulting oversupply of OS companies. The next three sections explore how judges and policymakers should exploit these insights.

\section*{IV. THE SHERMAN ACT (A): RULE OF REASON}

Section 1 of the Sherman Act reaches “\[e\]very contract, combination . . . or conspiracy, in restraint of trade . . . .”\textsuperscript{108} Under the rule of reason, judges and policymakers must intervene whenever an agreement’s “anticompetitive effects on trade outweigh its procompetitive effects.”\textsuperscript{109} We have seen that OS cost sharing eliminates wasteful duplication and lets companies write far more software than they could before.\textsuperscript{110} However, we have also seen that cartel effects will often suppress the second benefit.\textsuperscript{111} Does this diminished OS still offer enough benefits to satisfy the rule of reason?

This Part reviews the antitrust rules for CS cost sharing (that is, joint ventures) and uses this baseline to develop a rule of reason analysis for OS collaborations. We will see that it is almost always possible to design OS collaborations that satisfy the rule of reason for operating systems. However, the rule of reason case for applications programs is much closer and should normally be viewed with suspicion. I also suggest two “safe harbor” exceptions in which OS collaborations should be presumed valid. The Part concludes with a short discussion of monopolization issues under Section 2 of the Sherman Act.\textsuperscript{112}

\textsuperscript{107} Examples of such models can be found in Llanes & DeElejalde, \textit{supra} note 59, at 18, and Casadesus-Masanell & Llanes, \textit{supra} note 74, at 19–25.

\textsuperscript{108} 15 U.S.C. § 1 (2006). Liability does not depend on whether the challenged contract is memorialized in a single, overarching agreement or—as with OS licenses—a series of bilateral agreements. Interstate Circuit, Inc. v. United States, 306 U.S. 208, 208, 226 (1939) (inferring conspiracy from multiple bilateral agreements where parties agreed “knowing that concerted action was contemplated and invited”).

\textsuperscript{109} Nat’l Collegiate Athletic Ass’n v. Bd. of Regents of Univ. of Oklahoma, 468 U.S. 85, 104, 113 (1984) (agreements that raise prices or restrict output bear a “heavy burden” of showing that “apparent deviation” from normal free market principles is “competitively justifie[d]”).

\textsuperscript{110} See \textit{supra} text accompanying notes 55–95.

\textsuperscript{111} See \textit{supra} text accompanying notes 71–75.

A. Existing Joint Venture Law

There is still surprisingly little case law and commentary analyzing OS collaborations under the Sherman Act. Fortunately, the antitrust treatment of CS sharing (that is, joint research and development (“R&D”) ventures) has been extensively discussed, though seldom litigated. For most of the twentieth century, antitrust authorities evaluated joint R&D ventures like any other cartel. Debates over US competitiveness in the 1980s, however, led the US government, Congress, and scholars to a new appreciation that cost sharing could encourage companies to set more ambitious R&D goals. In the words of Professor Edward Correia:

“[A] single firm is frequently unwilling to make the investment necessary to enter a market or solve a technological problem on its own, but it will invest a smaller amount in a collective effort with a better chance of success. If courts incorrectly assume that individual collaborators will engage in the same effort on their own the result is to discourage collaborative investment on the mistaken theory that there will be even more investment if firms act independently.”

This new appreciation for sharing was nevertheless tempered by fears that cartelization might limit companies’ incentive to invest in R&D. Even if sharing did suppress R&D, however, joint ventures would still satisfy the rule of reason unless these “disincentives” were “sufficiently strong” so that “the decreased competitive pressure to innovate will outweigh whatever economies of scale or other efficiencies are likely to be generated by the collaboration.” Tests of this criterion included asking (a) whether the agreement “eliminates innovations that would have been cost-justified in a competitive innovation market,” (b) whether

---

113 But see Wallace v. IBM, 467 F.3d 1104, 1107 (7th Cir. 2006) (holding that GPL provision setting license fees equal to zero was not predatory pricing because GPL “keeps price low forever” and cannot “lead to monopoly prices in the future”).


115 See id.


117 Id. at 759.

118 Id.

119 13 HERBERT HOVENKAMP, ANTITRUST LAW: AN ANALYSIS OF ANTITRUST PRINCIPLES AND THEIR APPLICATIONS § 2115, at 114 (2d ed. 2005); see also Northrop Corp. v. McDonnell-Douglas Corp., 705 F.2d 1030, 1052–53 (9th Cir. 1983) (holding that
the parties would “have developed [the product] more quickly alone,” 120 and (c) “what the parents would have done but for the joint venture.” 121 More concretely, Congress and the Department of Justice also created a new safe harbor for joint R&D ventures. 122 This “five effort rule” created a presumption that joint R&D ventures are justified so long as they face competition from at least four other comparable programs. Strikingly, this new rule did not create an especially favorable test for CS sharing. Instead, it merely repeated the Department of Justice’s usual rule of thumb for evaluating market concentration in merger cases. 123

Formally, none of this marked a significant departure from previous learning. Nevertheless, the change was real. Prior to the 1980s, antitrust authorities had viewed CS sharing with suspicion. 124 This underlying attitude was now reversed by a new intuition that “[t]he optimal amount of innovation can ordinarily be achieved through competition or cooperative innovation ventures.” 125

B. A Rule of Reason Approach to OS Collaborations

We have seen that OS’s chief benefit is cost sharing and that its main vice is cartelization. 126 Furthermore, these effects cannot be disentangled. Firms might not invest without the promise of shared code, but this same promise also ensures cartelization. Judges applying the rule of reason must therefore decide whether the (admittedly truncated) benefits of OS sharing justify the cartel effect. Part III has argued that this balance should normally favor OS. This suggests a kind of informal intuition—as in the joint venture case—that most OS collaborations can be designed in ways that satisfy the rule of reason. On the other hand, this judgment may not hold for more complicated fact patterns. Judges should be particularly alert to instances where the value of OS sharing is doubtful, for example where (a) OS sharing is inefficient, (b) otherwise duplicative CS

a joint venture to invent military aircraft that neither partner could develop separately was not a per se violation).

120 See HOVENKAMP, supra note 119, at 118.
121 Correia, supra note 116, at 760.
123 Professor Correia has pointed out that an industry comprised of five identically sized competitors has a Herfindahl index of 2000, whereas the Justice Department scrutiny is triggered above 1800. Correia, supra note 116, at 759 n.84.
124 Id. at 756–58.
125 HOVENKAMP, supra note 119. This expansive proposition is said to be founded on economics research suggesting that even “a monopolist will not underinvest in research” in most real world situations. Correia, supra note 116, at 759 & n.82.
126 See supra Part III.
programs have been optimized to serve different user groups, or (c) competing R&D programs provide useful redundancy against failure.

The potential damage from cartelization will also vary from case to case. OS collaborations are especially likely to satisfy the rule of reason where cartelization effects either do not matter or matter very little. At least three situations fit this description:

1. Limited Absorptive Capacity

All projects eventually encounter diminishing returns beyond which further investment offers little improvement. The question is whether cartelization chokes off investment before or after this point is reached. In general, courts should be more willing to tolerate cartelization where the evidence suggests the project cannot usefully absorb additional resources.

2. Low-Priority Projects

We have assumed that companies fund every project that offers a positive return on investment. In Silicon Valley, however, capital is almost always scarce so that only the most promising projects are funded. Cartelization only matters if it affects these favored projects; the rest will be discarded in any case.

3. Pathological Competition

We noted in Part III that it is possible to construct scenarios in which quality competition drives software production so far into diminishing returns that manufacturers lose more value than consumers gain. That said, such situations are almost certainly rare and courts should greet such claims with skepticism.

Deciding when pathological competition exists is clearly fact dependent. Nevertheless, the most important variable—the type of software under development—can be anticipated. For convenience, we consider three types of products: (a) operating systems, (b) breakthrough programs that establish entirely new categories of software (“Killer Apps”), and (c) new applications that perform mostly familiar functions (“Commoditized Apps”).

a) Operating Systems

Operating systems provide the clearest candidate for a software product in which the downsides of cartelization can be safely ignored. Courts could sensibly conclude that existing operating systems are already deep into diminishing returns. First, their effect is mediated through application programs. This means that further improvements would exert, at best, only an indirect influence on the look, feel, and functionalities consumers actually receive. Second, operating systems often grow by absorbing existing functions from apps. The benefits of accelerating this process are probably small. Finally, the division of labor between apps and
operating systems is technically flexible. For this reason, failures to develop operating system features can often be remedied at the apps level.

Engineers say that “operating systems should be open and applications programs should be closed.”\(^{127}\) Apart from the empirical observation that most OS collaborations develop operating systems, however, the policy reasons for this intuition are unclear. Our analysis validates the first (“operating systems”) part of this statement. We now turn to the second part, that is, whether “application” programs should normally be closed.

\(\textit{b) Killer Apps}\)

Part III’s arguments were predicated on a tacit assumption that program quality depends almost entirely on the amount of code that is written. However, not all software derives its value from effort that “is sold by the pound.”\(^{128}\) Instead, application programs often derive significant earning power from clever new features.\(^{129}\) Here, the most extreme case consists of so-called “Killer App” products that discover previously overlooked needs or invent new functionalities.\(^{130}\) Any attempt to organize Killer Apps as an OS collaboration would therefore gain little from cost sharing while still suppressing incentives to innovate. This would almost certainly violate the rule of reason.

Fortunately, the threat is remote. Indeed, the empirical evidence suggests that OS collaborations avoid idea-driven products.\(^{131}\) Part of the reason is practical—because innovative ideas are (by definition) unanticipated, it is almost impossible to write an OS agreement that defines which ideas must be shared. This destabilizes OS collaborations by encouraging members to develop their best ideas as CS products. There is also a deeper problem. Even if an OS cartel were possible, most Silicon Valley CEOs, venture capitalists, and shareholders probably prefer high-risk, large-payoff races to develop the next Killer App.\(^{132}\) The prospect of small-but-predictable rewards does not appeal to them.\(^{133}\)

\(^{127}\) I am indebted to the late Professor Richard Newton for this observation.

\(^{128}\) I am grateful to the Eclipse collaboration’s Mike Milinkovich for coining this phrase.


\(^{130}\) Id.


\(^{133}\) See id. Hardware and service providers are even less likely to want this trade. “Killer Apps” have historically been instrumental in persuading consumers to upgrade their computer systems. See Scannell, supra note 129.
c) Commodified Apps

Over time, even the most unique products are imitated and become generic. At this point, the payoffs from making still more clever improvements become smaller and smaller. Companies may find it tempting to cartelize such software and refocus their cleverness on Killer Apps.

For now, it is hard to think of a convincing example in which OS has successfully cartelized an applications market. However, the potential is there. Consider, for example, the market for business desktops in which Microsoft’s dominant CS desktop products currently face four OS competitors. On the one hand, this five-way competition surely minimizes cartelization risk for the immediate future. On the other hand, it is easy to see how many of today’s OS collaborations could merge or drop by the wayside. If this happens, the cartel effect would grow steadily stronger.

1) Proposed Safe Harbors

Our argument that society receives maximal value when roughly 15–20 percent of all software companies practice OS sharing closely resembles the US Department of Justice’s five effort rule for joint ventures. This is no coincidence. Part III has argued that the cartel effect provides the main brake on OS output. It only makes sense that the effect should be smallest in industries that satisfy the Justice Department’s normal rule-of-thumb for judging market concentration.

The five effort rule will, however, need to be revised before it can be applied to OS collaborations. The reason is that our arguments do not depend on effort per se, but only on the number of OS companies which have joined the collaboration. For this reason, an OS safe harbor should be based on membership rather than effort. This revised safe harbor would immunize any OS collaboration that included less than 20 percent of all current industry members.

It also makes sense to create a second safe harbor for OS collaborations that develop operating systems. Such collaborations usually derive large benefits from sharing while presenting only minimal cartelization risks. Here, the main practical difficulty is the blurred line between operating systems and application programs. This would encourage OS collaborations to relabel application projects as “operating systems” in order to evade judicial scrutiny. Fortunately, this danger can be managed by limiting the safe harbor to programs that (a) are typically found

---

134 The competitors are Sun’s Star Office, Oracle Open Office, Gnome, and the K Desktop Environment.

135 See supra Figure 4.

136 Similarly, we have argued in Part III that the amount of effort that OS companies invest can vary dramatically depending on how much competition they face.

137 See supra Part III.
in today’s operating systems or (b) have become so generic that their inclusion in an operating system is no longer surprising.  

2) Section 2 Claims?

So far we have concentrated on the idea that OS might violate Section 1 of the Sherman Act. This is reasonable since a Section 2 claim would require proof that violators already possessed a substantial share of the market. OS companies hardly ever meet this test.

Section 2 issues could still arise on the CS side. Here, the most likely issue is lock-in. We have seen that incumbents can sometimes deliberately adopt CS to block entry by new competitors. Still, this behavior is unlikely to give rise to a Sherman Act violation. The reason lies in the well-settled case law holding that Section 2 does not create a duty to aid competitors. In effect, plaintiffs would have to argue that the OS collaboration was an “essential facility,” with the added twist that incumbents could not refuse to build such a facility in the first place. Given the disfavored status of essential facility claims under modern antitrust law, this theory would almost certainly fail.

V. THE SHERMAN ACT (B): VIRAL LICENSES

As discussed above—the cartel effect notwithstanding—OS sharing is often justified under the rule of reason. At this point, judges must accept some suppression of output. Further attempts to mitigate the cartel effect, if they come at all, will require government intervention.

On the other hand, there is still much for antitrust courts to do. In particular, judges will need to make sure that whatever cartelization does occur is truly unavoidable. Here, the most urgent issues involve so-called “viral” licenses that

---

138 The second inquiry would not be free from controversy. Courts have been reluctant to second-guess defendants’ claims that technical efficiency justifies bundling traditionally separate applications into a single program. See United States v. Microsoft Corp., 253 F.3d 34, 84 (D.C. Cir. 2001).
139 See supra text accompanying note 28.
141 See supra Part III.D–E.
142 See supra Part III.D.
144 For a recent review, see Robert Pitofsky et al., The Essential Facilities Doctrine Under U.S. Antitrust Law, 70 ANTITRUST L.J. 443, 443–45 (2002).
145 See infra Part VI.
146 “Viral” or “copyleft” licenses contain terms that require “anyone who redistributes the software, with or without changes, [to] pass along the freedom to further copy and
require consumers who use OS software to license any improvements or extensions on similarly open terms.

This Part presents a framework for deciding when viral license terms do and do not violate the antitrust laws. We begin by reviewing judges’ authority to strike down unnecessarily broad licenses under the Sherman Act. We then explore how viral licenses restrain (and in some cases also facilitate) competition. Next, we discuss various popular licenses that OS collaborations have used in the past. In the process, we find evidence that GPL and other so-called strong viral licenses are unnecessarily broad and that even some weaker licenses (for example, Mozilla) should often be viewed with suspicion. We close by identifying circumstances in which companies are most likely to adopt viral licenses as a deliberate cartelization strategy.

A. Antitrust Law and Restrictive Licenses

Many and perhaps most OS collaborations can be implemented in ways that satisfy the rule of reason. However, this does not immunize every “naked restraint” (more precisely: “unnecessary output limiting appendage”) that OS members decide to include in their licenses. Indeed, such terms are per se illegal. Do viral terms fit this definition? There is not much doubt that their announced purpose—encouraging companies that would normally create CS products to join OS instead—reinforces the cartel effect and therefore is “output limiting.” The harder question is whether such restraints are “necessary.” We will argue below that weak viral terms may sometimes be needed to stabilize OS collaborations in the first place. This is probably enough to escape per se illegality under current law. On the other hand, judges must still conduct a rule of reason change it. “What is Copyleft?, GNU OPERATING SYSTEM, http://www.gnu.org/copyleft/ (last visited Mar. 11, 2012). Although originally pejorative, the “viral” label has largely lost this connotation through indiscriminate usage. This Article uses the term purely for its descriptive value. Readers interested in the ongoing semantic debate should consult WIKIPEDIA. See Talk: Viral License, WIKIPEDIA, http://www.en.wikipedia.org/wiki/Talk:Viral_license (last modified Dec. 7, 2011).

148 See supra Part IV.


151 Doctrinally, the existence of a potential justification transforms the viral terms from a “naked restraint” to an “ancillary agreement.” The threshold for this transformation is still unclear. As Professor Piraino has emphasized, courts are divided over whether restraints must be “plausibly related to the venture’s procompetitive effects, reasonably related to those effects, or essential to achieving the effects.” Thomas A. Piraino, Jr., A Proposed Antitrust Approach to Collaborations Among Competitors, 86 IOWA L. REV. 1137, 1189 (2001). Piraino argues that defendants should not be asked to prove “that a restraint was so essential that, but for the restraint, the venture could not operate at all” but only that “in the opinion of the joint venture partners, [the restraint was] reasonably
inquiry to see whether the benefits of OS sharing could be achieved by adopting some less restrictive alternative. I argue below that such alternatives often exist, particularly for strong, GPL-type licenses.

Finally, suppose that a court finds that a particular viral license is overbroad. What relief can it grant? Unlike most Sherman Act cases, simple injunctive relief ordering parties “to cancel, shorten, or modify outstanding agreements” among two or more defendants may be impractical. This is because the number of OS users will often be so large so that courts cannot identify, much less assert jurisdiction over every licensee. This is unlikely to be problematic for the vast majority of OS licenses that give named “stewards” the power to issue amended terms. Here, it should normally be sufficient for the court to assert jurisdiction over the steward and order him or her to issue a less restrictive alternative license. For licenses without stewards, courts may have to content themselves with a declaration that the existing viral license violates the antitrust laws. This should allow users to assert a copyright misuse defense against anyone who tries to enforce the original license.

B. Viral Economics

The Free Software Foundation (“FSF”) invented GPL to encourage (or less politely, to compel) programmers to switch from CS to OS licenses. In its

necessary for their success.” Id. at 1190. This lower, subjective standard would make it much harder to strike down relatively weak viral licenses like Mozilla.

152 HOVENKAMP, supra note 149, ¶ 1913, at 373–75 (citing Sullivan v. Nat’l Football League, 34 F.3d 1091, 1112 (1st Cir. 1994); Wilk v. Am. Med. Ass’n, 895 F.2d 352, 378 (7th Cir. 1990)).


154 Most moderate and strong viral clauses include steward clauses. Examples include the Free Software Foundation’s GPLv2, see infra note 177, ¶ 9; GPLv3, see infra note 185, ¶ 14; and LGPL, see infra note 165, ¶ 6; Netscape’s Mozilla License, see infra note 198, ¶ 6; Apple Computer’s Public Source License, see infra note 204, ¶ 7; Sun’s Sun Industry Standards License, see infra note 215, ¶ 6; the IBM Public License, see infra note 207, ¶ 7; the IBM Common Public License, see infra note 208, ¶ 7; and the Eclipse Public License, see infra note 209, ¶ 7. Among the licenses reviewed for this article, only the very minimal Apache License, see infra note 221, and licenses modeled on the Berkeley Software Distribution license, see infra note 224, fail to name a steward.


original form, FSF’s argument was based on the assumption that GPL would be used to license “unique” OS modules that no CS program could match.\footnote{157} Programmers who wanted to use the modules would therefore have to adopt GPL themselves. This central assumption that a module could be “unique” was probably reasonable in a world dominated by small individual programmers who lacked the resources to duplicate large code bases. However, it was always more doubtful for commercial developers\footnote{158} who, as FSF never tired of pointing out, had “the advantage of money.”\footnote{159} For companies, the question was less whether a supposedly “unique” program could be duplicated—clearly it could\footnote{160}—but whether the profits from doing so were greater than those that could be earned by adopting GPL.

The rise of commercial OS, then, has rendered FSF’s original “uniqueness” argument obsolete. Even so, the basic instinct is sound. To see why, consider how a company would go about making OS/CS decisions for a group of proposed software modules. Following Part III, a CEO would start by estimating the relative profitability of developing each module using OS and CS methods. This would allow her to sort the modules into four categories:

(I) OS is much more profitable than CS;
(II) OS is slightly more profitable than CS;
(III) OS is slightly less profitable than CS; and
(IV) OS is much less profitable than CS.

At this point our CEO’s investment decision would depend on how the OS license was structured. A traditional, nonviral license would let the company make separate module-by-module choices. In this case, we would expect our CEO to develop the Category I and II modules as OS and the Category III and IV modules as CS. Viral licenses change this result by requiring the company to make take-it-

\footnote{157} FSF based its argument on experience with a program called “GNU Readline” which reportedly had “a significant unique capability.” FSF believed that “Releasing [Readline] under the GPL and limiting its use to free programs [gave] our community a real boost. At least one application program is free software today specifically because it was necessary for using Readline.” Richard Stallman, \textit{Why You Shouldn’t Use the Library GPL for Your Next Library}, GNU OPERATING SYSTEM (FEB. 1999), http://www.gnu.org/software/chinese/www/why-not-lgpl.2002-09-26.html.

\footnote{158} FSF seems to have assumed that commercial companies would continue to practice CS long after most individual programmers adopted OS. This explains its curiously muted boast that “Nowadays, as companies begin to consider making software free, \textit{even some} commercial projects can be influenced in this way.” \textit{Id.} (emphasis added).

\footnote{159} \textit{Id.}

\footnote{160} Because duplication takes time, it remains possible that some OS modules could be temporarily unique. However, this probably does not happen very often. The reason is that temporary uniqueness implies a head start, that is, that CS companies only learn about the OS module after it has been written. Today’s companies monitor OS developments so closely that such surprises seem unlikely.
or-leave-it decisions for bundles containing multiple modules. Here, shrewd license design can increase the number of modules that the company develops in OS mode by, for example, bundling some Category III and IV modules with a much larger number of Category I and II modules. The trick, of course, is to make sure that the former outnumber the latter. Viral licenses that include too many Category III and IV modules could easily backfire by persuading our CEO that CS development was more lucrative after all.

In the real world, designing licenses with just the right infectiveness is a rough-and-ready business. GPL solves this problem by using objective proxies based on the extent to which new modules copy or dynamically link to preexisting OS code. Specifying multiple or unusual links (a) limits the number of Category III and IV modules that companies are asked to embrace, and (b) provides at least some evidence that the module’s function is similar to the underlying OS software and therefore more likely to fall within Category III than Category IV. On the other hand, FSF acknowledges that these proxies can sometimes be wrong. It therefore encourages OS programmers to use the so-called Lesser General Public License ("LGPL") in cases where the normal GPL license would be ineffective or counterproductive. Unlike the GPL, the LGPL permits dynamic linking which makes it markedly less infective.

Suppose, then, that viral terms succeed in persuading companies to adopt OS more often. What are the downsides? First and most obviously, viral licenses increase the number of companies that choose OS development for any given module. This aggravates the already large imbalance between OS and CS companies and reinforces the cartel effect. Second, software development is not an

---

161 In principle, OS architectures can also be manipulated to change the boundaries between modules. This is because programmers have considerable technical freedom to place software in one “container” instead of another. Greg R. Vetter, “Infectious” Open Source Software: Spreading Incentives or Promoting Resistance?, 36 RUTGERS L.J. 53, 101–13 (2004). This means that it should be possible to gerrymander software projects so that the functions that companies prefer to develop using CS methods are concentrated in as few modules as possible. Such a strategy would pose significant antitrust problems given courts’ admitted reluctance to second-guess software designers’ architecture choices. See, e.g., United States v. Microsoft Corp., 253 F.3d 34, 84 (D.C. Cir. 2001) (expressing reluctance to second-guess engineering decision to combine previously separate software programs).

162 Category III modules may, however, be developed less robustly than they would be under a CS business plan.

163 See Lerner & Tirole, supra note 5, at 202–04.

164 Applications programs are said to use “static links” when code from preexisting software is copied and permanently incorporated during installation. By contrast, “dynamic links” install code temporarily while the application is running and later erase it again. Vetter, supra note 161, at 82–85, 104–06.


166 See Vetter, supra note 161, at 104–05, 199.

167 See supra Part V.A.
all-or-nothing process. Companies that would have preferred to use CS methods will normally spend less if viral terms force them to use OS methods instead. This suggests a Faustian bargain: broad viral terms produce more OS software, but this gain is more than offset by lost CS production.168

C. Justifications

If viral terms are the problem, why not get rid of them? Unfortunately, things aren’t so simple. The reason is that some viral terms may still be needed to keep collaborations from unraveling. To see why, consider a company that would like to improve or extend an existing OS module. Plainly, its most profitable business strategy is to copy the underlying OS module at zero cost and then sell CS improvements for whatever the market will bear. This, of course, is a formula for disaster. If every company does this, OS will disappear entirely.169

Viral provisions block this outcome by forcing companies that use OS code to donate their improvements back to the community.170 The question is, how narrow can this restriction be and still ensure stability?171 Logically, viral terms should be at least as broad as the underlying OS module. On the other hand, this might not be enough. The problem, as Professor Vetter has noted, is that programmers have considerable technical freedom to move software from one module to another.172 This means that viral clauses may have to be broader than the original OS module to enforce stability. Probably the simplest solution is to extend the viral term to any module that includes code copied from an underlying OS program. This approach is found in the widely used Mozilla license and its imitators.173 This may not be sufficient, however, in cases where companies can write a second module that links to the first module and causes it to behave as if it had been rewritten. In such cases, the viral provision must cover both the first module and at least some of the modules that link to it.174

168 See supra Part V.A; see also Vetter, supra note 161, at 152–53.
170 Vetter, supra note 161, at 155–56 n.272.
171 Professor Katz has noted a similar problem in the context of joint R&D ventures. He warns that “[a]ntitrust authorities should be wary of agreements that attempt to limit the R&D that a member firm may conduct without sharing. Firms should have to prove that the restraint is essential to the proper functioning of the cooperative agreement.” Katz, supra note 76, at 542.
174 The license does not have to proscribe all or even most links. It only has to narrow the programmer’s toolbox of possible links to the point where evading OS status is no longer worth the trouble. This is similar to the usual argument that the market power conferred by patents is measured by the “inventing around” costs. Scotchmer, supra note 68, at 105, 107–12.
That said, the argument for viral provisions contains an important loophole. We have assumed that companies can sell their CS improvements on the open market. This may not be true for a variety of reasons. First, the modifications may be so specialized or have such limited value that finding potential buyers is not worth the transaction cost. Second, consumers may find it prohibitively expensive to judge software quality for themselves. In this case, they may prefer OS software, not because it is better, but because the collaboration has endorsed it. Finally, consumers may worry that CS software could stop working if and when the “official” software changes. In any of these cases, viral terms could be completely unnecessary.

This is as far as theory can take us. At this point, proving the existence of less restrictive alternatives becomes an empirical question. As Professor Hovenkamp has remarked:

[P]laintiffs cannot be permitted to offer possible less restrictive alternatives whose efficacy is mainly a matter of speculation. . . . Proffered less restrictive alternatives should either be based on actual experience in analogous situations elsewhere or else be fairly obvious. Tending to defeat such an offering would be the defendant’s evidence that the proffered alternative has been tried but failed, that it is equally or more restrictive, or otherwise unlawful.

Fortunately, commercial OS collaborations have used many different viral licenses over the past decade. We now turn to this evidence.

1. Case 1: Strong Viral Terms

The broadest and most stringent viral terms are invariably found in FSF’s GPLv2. This is the best-known and strongest viral license. Formally, it extends to any software “based on” protected code. While the scope of this phrase has been much debated, it seems to include (a) files that include any verbatim portion of the preexisting code, (b) entirely new modifications to that code, (c) any interface files, compiler and

---

175 See supra text accompanying note 169.
176 Hovenkamp, supra note 149, at 375–76.
177 GNU General Public License Version 2, GNU OPERATING SYSTEM, para. 0 (June 1991), http://www.gnu.org/licenses/old-licenses/gpl-2.0.html (last visited Jan. 15, 2012) [hereinafter GPLv2].
179 GPLv2, supra note 177, at para. 0.
180 Id.
installation scripts needed to run that code, and (d) all “derivative work,” a term that arguably incorporates the full reach of “derivative works” under US copyright law. The concept of verbatim copies is also said to include both static links that copy software prior to use and dynamic links which make transient copies while the program is running. FSF argues that the license’s scope should normally be evaluated based on the number and type of interactions between the underlying OS software and any new software. To the extent that viral terms apply, programmers must adopt GPLv2 for their own work and cannot charge royalties.

GPLv3. This license was designed to simplify GPLv2’s sometimes obscure “based on” language. This is done by extending the viral term to any software that “cop[ies] from” or “adapt[s] all or any part” of the underlying OS software. Programmers who write such software must license it at zero royalties under GPLv3 and also supply enough “corresponding source” software for users to run the code.

Oracle’s Berkeley Database License. The scope of this license includes both preexisting OS code and “modifications.” Programmers whose work falls within this definition must make their source code

---

181 Id. at para. 3. The license contains a “special exception” for “anything that is normally distributed . . . with the major components . . . of the operating system on which the executable runs . . . .” Id.
182 Id. at para. 0.
183 GPLv2 does create an exception for code “written entirely by you.” Id. at para. 2. However, this is implemented in a fairly minimal way by excluding software only when it is distributed as a “separate work.” Id. GPLv2 also creates exceptions for software that is part of a “mere aggregation” on a storage medium. Id.
184 FSF has suggested that courts should decide which links qualify based on both “the mechanism of communication (exec, pipes, rpc [sic], function calls within a shared address space, etc.) and the semantics of the communication (what kinds of information are interchanged).” Ron Phillips, Deadly Combinations: A Framework for Analyzing the GPL’s Viral Effect, 25 J. MARSHALL J. COMPUTER & INFO. L. 487, 493–94 (2008).
185 GPLv2, supra note 177, at paras. 10, 14.
187 Id. at para. 10.
188 Id. at para. 6. Significantly, the term excludes “system libraries” and “general purpose” tools. Id. at para. 1. This presumably reflects a realization that companies faced with such demands would reject GPL entirely.
190 Id.
available to users together with “any accompanying software that uses” it.\textsuperscript{191}

There are several reasons to think that GPL and its progeny are much broader than stability requires. First, FSF designed GPL to have the maximum possible impact. Nothing in the written record suggests that FSF thought that viral clauses could be too strong, let alone that the ratio of OS to CS companies could ever be too high.\textsuperscript{192} Second, FSF knew that narrower licenses were possible and stable. The existence of an LGPL that lets CS programs dynamically link to OS libraries proves this.\textsuperscript{193} Finally, Linus Torvalds has repeatedly relaxed GPL so that programmers can write CS programs that make “normal system calls” on LINUX\textsuperscript{194} and even insert CS “loadable kernels” into LINUX itself.\textsuperscript{195} Despite this, the LINUX collaboration shows no signs of instability.

2. Case 2: Weak Viral Terms

LINUX apart, “infectious terms have been GPL’s least imitated feature.”\textsuperscript{196} Probably the biggest reason is that broad viral terms make it harder for software companies to earn revenue from complementary CS programs. Relaxing this restriction strengthens OS collaborations by allowing more companies to participate.\textsuperscript{197} This may explain why commercial OS collaborations overwhelmingly choose so-called “weak copyleft licenses.” Examples include:

\textsuperscript{191} Id. at para. 3. The license is tempered by Oracle’s statement that it is prepared to negotiate royalty-bearing licenses that waive these obligations. See id. This suggests that— unlike FSF’s licenses—the Oracle license is mainly a negotiating ploy.

\textsuperscript{192} See, e.g., Stallman, supra note 157.

\textsuperscript{193} Some commentators have also argued, contrary to FSF, that GPLv.2 also permits dynamic links. See, e.g., Lothar Determann, Dangerous Liaisons—Software Combinations as Derivative Works? Distribution, Installation, and Execution of Linked Programs Under Copyright Law, Commercial Licenses, and the GPL, 21 BERKELEY TECH. L.J. 1421, at 1491, 1496 (2006).

\textsuperscript{194} Greg R. Vetter, Claiming Copyleft in Open Source Software: What if the Free Software Foundation’s General Public License (GPL) Had Been Patented?, 2008 Mich. St. L. Rev. 279, 311–12. The situation has been muddied by Torvalds’ suggestion that Linux does extend to software that is “so obviously Linux-specific that it simply wouldn’t make sense without the Linux kernel.” Mitchell L. Stoltz, The Penguin Paradox: How the Scope of Derivative Works in Copyright Affects the Effectiveness of the GNU GPL, 85 B.U. L. Rev. 1439, 1452–53 (2005). Conversely, Torvalds has argued that GPL does not reach preexisting software that has been modified to place calls on the Linux kernel. Vetter, supra note 161, at 154–55.

\textsuperscript{195} John Tsai, Note, For Better or Worse: Introducing the GNU General Public License Version 3, 23 BERKELEY TECH. L.J. 547, 559–60 (2008).

\textsuperscript{196} Vetter, supra note 161, at 151.

\textsuperscript{197} For example, Sun decided not to use GPL because it wanted collaboration members to be able to create “larger works for commercial purposes.” FAQ: Common Development and Distribution License (CDDL), OPEN SOLARIS, (Nov. 10, 2011, 3:38 PM),
Mozilla Licenses. The “Mozilla Public License”\footnote{Mozilla Public License Version 2.0, \url{https://www.mozilla.org/MPL/2.0/} (last visited Feb. 6, 2012).} covers all underlying OS code and “modifications.” The latter are defined to include (a) additions or deletions from existing files, and (b) any new module that includes copies of licensed code.\footnote{Id. at para. 1.9.} In practice, it is usually interpreted in ways that permit CS programs to include unlimited dynamic and static links to the underlying software and also construct “larger” CS works that bundle OS and CS software together.\footnote{See, e.g., GNU General Public License, \url{http://en.wikipedia.org/wiki/GNU_General_Public_License} (last visited Jan. 16, 2012).} Covered software must, however, be made available to users as source code on a royalty-free basis.\footnote{Mozilla Public License Version 2.0, supra note 198 at para. 2.1.} Sun’s “Common Development and Distribution License”\footnote{Common Development And Distribution License Version 1.0, \url{http://www.opensource.org/licenses/cddl1.php} (last visited Jan. 18, 2012).} and “Sun Public License”\footnote{Sun Public License Version 1.0, \url{http://java.sun.com/spl.html} (last visited Jan. 18, 2012).} contain similar provisions.

Apple Public Source License.\footnote{Apple Public Source License Version 2.0, \url{http://www.opensource.apple.com/license/apsl/}.} This license asserts Mozilla-type breadth but permits companies to sell CS modifications provided that the source code is supplied to users.\footnote{Id. at para. 2.2(c).} Here, the formal ability to license extensions is undercut by the requirement to distribute source code. This makes it easier for consumers to evade licensing restrictions on, for example, transferring the software to unauthorized users. Similar requirements can also be found in Sybase’s “Sybase Open Source License.”\footnote{The Sybase Open Watcom Public License 1.0, \url{http://www.opensource.org/licenses/sybase.php} (last visited Jan. 18, 2012).}

IBM/Eclipse Licenses. The “IBM Public License,”\footnote{IBM Public License Version 1.0, \url{http://www.opensource.org/licenses/ibmpl.php} (last visited Jan 18, 2012).} “IBM Common Public License,”\footnote{Common Public License (CPL) -- V1.0, \url{http://www.ibm.com/developerworks/library/os-cpl.html} (last visited Jan 18, 2012).} and Eclipse Foundation “Eclipse Public License”\footnote{Id. at para. 2.2(c).} all

\url{http://hub.opensolaris.org/bin/view/Main/licensing_faq#HCanIredistributeorselltheOpenSolarissource.}
assert Mozilla-type breadth.\(^{210}\) However, they are much less stringent since they let users sell the underlying OS code (with or without modifications) under CS licenses provided that they make their source code available.\(^{211}\) Furthermore, the Eclipse license does not reach “separate modules” unless they qualify as “derivative works” under US copyright law.\(^{212}\) While Eclipse concedes that the precise boundaries of this exception are unclear,\(^{213}\) companies are clearly permitted to write CS licenses for Eclipse plug-ins that contain 100 percent new code and “implement functionality not currently in Eclipse.”\(^{214}\)

**Sun Industry Standards Source License (SISSL).**\(^{215}\) Sun provided its Open Office software under this license until 2006.\(^{216}\) The unusual document asserts Mozilla-type breadth, but permits programmers considerable freedom to write CS software. Thus, they can sell modifications under CS licenses provided that (a) they continue to make the original source available under the SISSL license, and (b) their modifications meet Open Office standards. Conversely, companies whose products fail to satisfy Open Office standards must publish a list of all deviations and/or the modification’s source code.\(^{217}\) In either case, the obligation to disclose source code is much lower than for other Mozilla-type licenses.


\(\)\(^{210}\) See IBM Public License Version 1.0, supra note 207, at para. 1; Common Public License (CPL) -- V1.0, supra note 208, at para. 1; Eclipse Public License - v 1.0, supra note 209, at para. 1.

\(\)\(^{211}\) IBM Public License Version 1.0, supra note 207, at para. 3(b)(iv); Common Public License (CPL) -- V1.0, supra note 208, at para. 3(b)(iv); Eclipse Public License - v 1.0, supra note 209, at para. 3(b)(iv).

\(\)\(^{212}\) IBM Public License Version 1.0, supra note 207, at para. 1(b)(ii); Common Public License (CPL) -- V1.0, supra note 208, at para. 1(b)(ii); Eclipse Public License - v 1.0, supra note 209, at para. 1(b)(ii).


\(\)\(^{214}\) Id. at para. 27.


\(\)\(^{217}\) Sun Industry Standards Source License - Version 1.1, supra note 215, at para. 3.1.
The popularity of the foregoing licenses provides powerful evidence that even relatively narrow viral terms are enough to stabilize OS collaborations. First, none of the licenses ban CS licenses that link to preexisting code. Indeed, the IBM and Eclipse licenses grant companies an unrestricted right to sell separate CS “plug-in” modules provided that they contain entirely new software. Second, only the Mozilla license completely bars users from selling modified code under CS licenses. That said, all of the above-listed non-Mozilla licenses also restrict this right, usually by requiring users to make the source code for their modifications available to users. Because these terms make CS licensing more difficult, they are at least arguably related to stability.

3. Case 3: Minimally Viral Terms

Finally, many commercial projects contain only minimal viral terms. Examples include:

Apache License. This license formally reaches all software that is based on or derived from the underlying OS code. However, it only requires licensees to reproduce the preexisting software’s copyright, trademark, attribution, and legal notices. Users are otherwise free to sell modified (and even original) versions of the underlying software under CS licenses.

Berkeley Software Distribution (BSD). BSD formally covers all “redistribution with or without modification.” However, licensees’ obligations are limited to retaining certain disclaimers and notices. This gives users an unrestricted right to license both original and

---

218 IBM Public License Version 1.0, supra note 207, at para. 2; Common Public License (CPL) -- V1.0, supra note 208, at para. 2; Eclipse Public License - v 1.0, supra note 209, at para. 2.
219 Mozilla Public License Version 1.1, supra note 198, at para. 2.1.
220 Requiring users to disclose source code for their modifications almost certainly makes CS less attractive. Given a choice, practically all commercial CS companies withhold source code as a way of discouraging unauthorized use.
222 Apache License, Version 2.0, supra note 221, at para. 1. The license gives no indication that it seeks to incorporate the meaning ascribed to “derivative works” under the US Copyright Act.
223 Id. at para. 4. Programmers must, however, offer royalty-free licenses on any software which they donate to the Apache Foundation itself. See id. at 5.
225 See id.
modified versions of the underlying OS software commercially. Commercial versions of BSD include the “Cryptix General License”\(^\text{226}\) and the “Intel Open Source License.”\(^\text{227}\)

The popularity of Apache and BSD licenses shows that many OS collaborations can be stabilized with minimal viral restrictions. That said, there may be cases where weakly viral, Mozilla-type licenses are necessary. The reason is that the Apache and BSD licenses usually appear in markets where one-man businesses deliver custom software to individual clients. Here, customers find it impractical to reuse or share software no matter what their license says. It would be more informative to find Apache or BSD-style licenses in mass markets. Apple’s decision to make many of its software products available under Apache instead of the weakly viral Apple Public Source license\(^\text{228}\) suggests that such markets do, in fact, exist.

D. Viral Licenses in a Commercial Age

We have argued that viral terms—and especially strong, GPL-type licenses—needlessly extend cartel effects and limit the amount that companies invest in OS. Historically, these terms have usually been motivated by ideology.\(^\text{229}\) However, we have seen that companies can also use such licenses to reinforce the cartel effect. The appeal of this strategy will normally depend on the type of company involved:

**Software Companies.** We have already emphasized that OS modules are most profitable when CS competition is weak. GPL terms provide a natural way to suppress CS and make these modules even more profitable. The main constraint, as we have seen, is that most software companies would prefer not to cartelize the Killer App market. GPL licenses that do this are unlikely to be adopted.

**Hardware and Service Providers.** Hardware and service providers face a more ambiguous choice. On the one hand, strong viral terms increase the supply of free software. This lets consumers spend more of their IT budgets on hardware and services. On the other hand, we have seen that strong viral terms suppress CS competition. This will normally reduce


\(^{229}\) But see Josh Lerner & Jean Tirole, The Scope of Open Source Licensing, 21 J.L. ECON. & ORG. 20, 53–55 (2005) (arguing that commercial firms may sometimes adopt more restrictive OS licenses to attract volunteer programmers who would not otherwise participate).
the total supply of software so that consumers receive less value from whatever hardware and services they do buy. For this reason, the decision to adopt strong, GPL-style licenses will usually be highly fact dependent.

For most businesses, then, viral terms offer both benefits and costs. This may explain why most commercial OS collaborations use weaker licenses.\footnote{See supra notes 196–220 and accompanying text.} At the same time, deliberate cartelization remains a plausible threat. This provides yet another reason for courts and policymakers to scrutinize GPL-style licenses.

VI. GOVERNMENT INNOVATION POLICY: TAXATION, SUBSIDIES, GRANTS, AND PROCUREMENT POLICY

Part III has argued that the OS cartel effect suppresses Year One competition and leads to a systematic imbalance between OS and CS companies. On the other hand, Part IV suggests that these costs will often be acceptable under the rule of reason. At this point, nothing the Sherman Act can do to protect Year Two competition is likely to fix the imbalance. Relief, if it comes at all, will require government intervention.

The idea of government intervention has been widely debated for most of the past decade.\footnote{Compare LAWRENCE LESSIG, THE FUTURE OF IDEAS 247–49 (2001) (arguing for more government encouragement of OS solutions), with David S. Evans & Bernard J. Reddy, Government Preferences for Promoting Open-Source Software: A Solution in Search of a Problem, 9 MICH. TELECOMM. & TECH. L. REV. 313, 393–94 (2003) (disagreeing, and arguing that governments should not intervene in favor of OS solutions where there is “no general market failure . . . in the provision of commercial software”).} At first, pro-OS scholars assumed a simple objective: if free software was good, then more software would be better.\footnote{See LESSIG, supra note 231, at 174–76 (arguing that the benefits of freer Internet infrastructure and code substantially outweigh the benefit of corporate controlled infrastructures).} This led to many different proposals for promoting OS through procurement preferences, taxation, subsidies, and other interventions.\footnote{See, e.g., id. at 247; LERNER & SCHANKERMAN, supra note 16, at 199 (“Incentives for software development or adoption can take a variety of forms, including monetary incentives such as direct subsidies and tax credits. But they can also be more indirect, such as when governments attach specific software conditions to procurement decisions, regulatory policy and even informal pressure and ‘moral suasion.’’”). Other scholars have called for “introducing a National Software Foundation . . . that will fund software development projects on condition that the fruits be licensed as free software, and the adoption of a . . . policy that would require that software written under government contract be released as free software.” Yochai Benkler, Freedom in the Commons: Towards a Political Economy of Information, 52 DUKE L.J. 1245, 1275 (2003).} Significantly, scholars who opposed these suggestions seldom argued with the assumption that more OS was always desirable. Instead, they raised a practical objection. Given that OS volunteers were
hardly ever motivated by money, 234 traditional policy levers based on taxation or spending were unlikely to change behavior. 235 This academic discussion was followed by a similarly inconclusive debate in governments around the world. 236 While some promotion schemes were enacted, most were abandoned and new proposals dropped off rapidly after 2005. 237 This does not, however, mean that governments lost interest. Instead, the original idea that there could never be enough OS became more nuanced. In the words of one leading observer, schemes to promote OS have become “subsumed” in a broader “search for business models that can profitably blend open and proprietary processes and products.” 238 But what these OS/CS blends might look like was seldom, if ever, specified.

The rise of commercial OS changes—and significantly clarifies—this debate. On the one hand, the old objection that government cannot possibly influence an activity motivated by “fun” no longer holds. Modern commercial OS clearly responds to financial incentives. 239 On the other hand, the cartel effect provides a much clearer analytical framework for deciding when the “right” OS/CS blend has been achieved. This section asks how government can use the various options at its disposal to suppress the cartel effect and help OS sharing reach its full potential.

---

234 Evans & Reddy, supra note 231, at 341, 344 (OS has “primarily been developed by individuals who donate their time to work on projects that interest them” and “[t]he circumstances under which a for-profit firm has incentives to invest in open source, particularly under the GPL, may be limited.”); Jyh-An Lee, New Perspectives on Public Goods Production: Policy Implications of Open Source Software, 9 VAND. J. ENT. & TECH. L. 45, 47 (2006) (profit incentive “does not exist in the open source community”); Klaus M. Schmidt & Monika Schnitzer, Public Subsidies for Open Source? Some Economic Policy Issues of the Software Market, 16 HARV. J.L. & TECH. 473, 481–84 (2003) (noting that most OS is motivated by nonmonetary incentives and asserting that commercial OS investments “are likely to remain limited” because of free-riding).


236 Politicians around the globe have given various reasons for promoting OS including favoring local industry (Argentina, Australia, Brazil, China, Costa Rica, Germany), putting pressure on CS companies to compete (Germany, Denmark, Taiwan, Thailand), becoming less dependent on American multinationals (Costa Rica, Japan, Russia, Thailand), improving computer security (Costa Rica, Argentina, Iran) and reducing copyright piracy (Iran, Thailand). Government Open Source Policies, CENTER FOR STRATEGIC & INT’L STUD. (July 2008), www.csis.org/files/media/csis/pubs/0807218_government_opensource_policies.pdf; see also Lee, supra note 234, at 101, 105.

237 E.g., Government Open Source Policies, supra note 236.

238 Id. at 1.

239 See supra notes 12–15 and accompanying text.
A. Purchasing Preferences

Governments already spend large amounts of money purchasing software.\textsuperscript{240} For this reason, politicians often argue that redirecting these resources to promote OS is essentially costless.\textsuperscript{241} Since the early 2000s, at least sixteen countries have considered legislation or regulation that would require government agencies (and in some cases state-owned companies) to adopt OS solutions when such products exist.\textsuperscript{242} Furthermore, at least four countries have actually adopted such measures\textsuperscript{243} and six more have adopted “preferences” for OS where its performance is comparable to CS alternatives.\textsuperscript{244}

Do these policies make sense? It depends. We have seen that market imperfections can (a) prevent all-OS markets from evolving into mixed OS/CS states,\textsuperscript{245} (b) prevent all-CS markets from evolving into mixed OS/CS states,\textsuperscript{246} and (c) systematically over-supply OS companies where mixed OS/CS states already exist.\textsuperscript{247} Using pro-OS purchasing preferences to encourage new OS entry would be irrelevant in case (a) but potentially helpful in case (b). On the other hand, we have seen that there are already too many OS companies in mixed markets.\textsuperscript{248} This

\begin{itemize}
\item \textsuperscript{240} Professors Lerner and Schankerman’s survey of 1,894 companies operating in fifteen countries found that government purchases accounted for 21 percent of total revenue. \textit{Lerner & Schankerman, supra} note 16, at 177. This figure was “about equally divided” among federal, local, and municipal governments. \textit{Id.}
\item \textsuperscript{241} See, e.g., \textit{Government Open Source Policies, supra} note 236, at 19.
\item \textsuperscript{242} Argentina, Australia, Belgium, Brazil, Bulgaria, Chile, Colombia, Finland, Italy, Malaysia, The Netherlands, Peru, Portugal, Spain, Ukraine, and Venezuela have all considered bills or regulations to implement mandatory open source procurement. \textit{Id.} at 3–6, 8, 11–16, 19.
\item \textsuperscript{243} The Netherlands, Peru, South Africa, and Venezuela have all adopted some version of mandatory preferences at the national level. Many more countries have adopted mandatory adoption requirements at the state and local level. For example, many Brazilian state and municipal governments have established requirements or preferences in favor of OS. \textit{Id.} at 4, 13–14, 16, 19.
\item \textsuperscript{244} Australia, Belgium, Brazil, China, Malaysia and Spain. \textit{Id.} at 3–5, 12, 16. This probably understates the number of countries that have adopted preferences. For example, France claims that it has merely “encouraged” OS use in government. However, its 96 percent adoption rate suggests that something more systematic is at work. Gijs Hillenius, \textit{FR: Almost the Entire Public Sector is Using Open Source}, \textsc{European Commission Joinup} (Sept. 30, 2009), https://joinup.ec.europa.eu/news/fr-almost-entire-public-sector-using-open-source.
\item \textsuperscript{245} See \textit{supra} text accompanying notes 77–88.
\item \textsuperscript{246} See \textit{supra} text accompanying notes 86–91.
\item \textsuperscript{247} See \textit{supra} text accompanying notes 93–98.
\item \textsuperscript{248} See \textit{supra} text accompanying notes 93–98.
\end{itemize}
suggests that government procurement preferences for OS software would make problem (c) even worse than it is today.249

But if pro-OS preferences are a bad idea, why not adopt pro-CS preferences instead? Currently, the idea is politically unlikely. But this could change over time as the image of OS becomes more corporate. In the meantime, it is important to know whether a pro-CS procurement policy makes sense. Here the good news is that preferences would help CS companies win more business and become more profitable. This would eventually lead to more CS entrants and dilute the cartel effect. However these gains would come at a price:

1. **Fine-Tuning**

   We have argued that government should try to achieve the right “blend” of OS and CS companies. But how will government know when it achieves this mix? Certainly, any idea of fine-tuning seems hopelessly optimistic. On the other hand, this may not matter if—as Figure 4 suggests—the current blend is far from ideal. Several scholars have pointed out that government purchases are too small to have more than a minor impact on software markets.250 This could be a good thing if it means that there is no chance of overshooting the desired mix.

2. **Crowding Out**

   By definition, procurement preferences let CS companies sell more software with identical effort or sell the same amount of software with less effort. In practice, we expect a mix of both. This suggests that government preferences will cause CS companies to increase their investments, but not as fast as the value of the preference.

3. **Costs to Government**

   Government procurement preferences only matter to the extent that they force government to choose software that it would not otherwise use. We should therefore assume that government will receive less value (and fewer capabilities) for every dollar it spends on software.

---

249 For a different argument that pro-OS subsidies are counterproductive, see LERNER & SCHANKERMAN, supra note 16, at 170 (arguing that OS code is not priced at its true cost and this artificially suppresses CS markets).

250 Lee, supra note 234, at 60 n.86 (estimating $17 billion government procurement market worldwide as of 2002). But see LERNER & SCHANKERMAN, supra note 16, at 177 (suggesting that government purchases account for one-fifth of all software revenue).
B. Subsidies

Governments have launched a variety of initiatives to subsidize OS methods including grants and cash incentives to private sector OS adopters,251 government-funded projects and alliances to promote OS methods,252 government-funded training, support, and demonstration centers,253 and government-funded OS workshops.254 Since government sometimes promotes CS models as well—particularly in the case of start-ups—it is hard to know how much these policies favor OS on net.255 Nevertheless, it is at least reasonable to think that OS firms receive substantially more support than their CS competitors.

One nice feature of these policies is that they are disproportionately aimed at helping OS companies penetrate new markets. This suggests that they may sometimes help all-CS markets avoid lock-in. The case is more difficult for subsidies in mixed markets where OS is already firmly established. Here, subsidies—like procurement policy—are likely to crowd out significant amounts of private investment.

Once again, the deepest problem relates to fine-turning—that is, knowing when to stop the subsidies once a mixed OS/CS market exists. Part III has argued that OS collaborations should follow something like the five effort rule—the rule that the number of OS companies should be comparatively small.256 Politicians, on the other hand, may see the fact that “only” one in five companies practices OS as an invitation for continued support. This tendency is especially likely given government’s traditional reluctance to dismantle “infant industries” programs that have served their purpose.257

C. Tax Policy

Economics textbooks routinely praise the nondistortionary benefits of tax policies that use lump sum taxes and tax breaks to influence behavior. While no government seems to have considered the idea so far, tax policy provides a natural lever for making OS companies less profitable and CS companies more profitable.

251 Singapore, Hong Kong, and Israel. Israel also offers funds for OS start-up companies. Government Open Source Policies, supra note 236, at 10–11, 15.

252 Cambodia, China, Czech Republic, Israel, South Africa, South Korea, Japan, Netherlands, Philippines, Thailand, and Vietnam. Id. at 5–6, 11–20.

253 Brazil, Malaysia, Pakistan, Western Australia, and Spain. Id. at 4, 12–14, 16, 21.

254 Brazil. Id. at 4.

255 Despite a very large data set, Professors Lerner and Schankerman were unable to determine whether government subsidies favor OS or CS on net. LERNER & SCHANKERMAN, supra note 16, at 199–200.

256 See supra text accompanying notes 135–138.

This would almost certainly shift the OS/CS ratio toward the five effort rule’s 15–20 percent target.

The great advantage of lump sum tax policies is that—unlike purchasing preferences or subsidies—they do not change companies’ incentives.258 Indeed, the only way for companies to evade such taxes is to go out of business. This means that we can confidently expect industry to go on making the same investment decisions as if government had never intervened. For this reason, a well-designed tax policy would eliminate our “crowding out” and “costs to government” objections. Only “fine-tuning”—that is, knowing when the right OS/CS ratio has been reached—remains.

D. Direct Investment in OS

Each of the foregoing policy levers is designed to strengthen CS companies so that increased competition forces OS firms to develop more software. But this seems terribly indirect. If the goal is to develop more OS code, why not purchase it directly? To some extent, government grants do this already.259 This usually happens informally when grant agencies let individual faculty—most of whom are ideologically predisposed to OS—decide how to license their government-funded software. Furthermore, many American grant agencies—notably NIH—now require grant applicants to provide “dissemination strategies” for sharing their data and discoveries with the wider world.260 OS licenses provide a particularly simple way to meet these requirements. Indeed, the United Kingdom has taken this logic one step further by adopting interim regulations that make OS licenses the “default position” for government-funded software.261 Finally, many governments fund projects, institutes, alliances, and consortia whose missions include writing OS software.262

Not surprisingly, purchasing still more OS from the private sector aggravates the usual imbalance in mixed OS/CS industries. On the one hand, government contracts make OS companies more profitable. This increases their numbers and

261 The regulations provide: “[I]f no exploitation route is specified for government-funded R&D software outputs, the default position of the government should be ‘to adopt an open source software license which complies with the OSI definition (which includes the GPL and Berkeley style licenses) or a United Kingdom-specific analogue of it’ [and] ‘all government-funded software should be accompanied by appropriate documentation which will assist the exploitation via the open source software license.’” Government Open Source Policies, supra note 236, at 18.
262 China, Finland, Japan, South Korea, France, India, Slovakia, Spain, Thailand, Venezuela, and Vietnam. Id. at 5, 8–10, 12, 15–20.
makes the ratio of OS-to-CS companies even less favorable than it was before.\textsuperscript{263} On the other hand, each dollar that government spends to develop OS code dilutes companies’ incentives to invest their own money. This aggravates the cartel effect. Instead of closing the gap in Figure 4, then, direct government investment widens it still further. Despite this, the glass is at least half-full. Cartel effect or not, direct investment increases the amount of OS that society can use and enjoy. Does it really matter whether this investment is made by taxpayers instead of shareholders?

Conservatives will object that this kind of government-funded R&D is bound to cost more than commercial OS. Furthermore, direct support could easily substitute government design choices for market signals.\textsuperscript{264} And indeed, poorly designed funding schemes—for example, paying a computer science professor to create her personal vision of “ideal” cell phone software—could easily end in disaster. At the same time, these objections are not really fundamental. The trick will be to design schemes that encourage companies to work cost-effectively on software projects that the market actually wants. This could be done by, for example, inviting commercial OS collaborations to tender competing bids that promise to implement a specific software idea at a guaranteed price. Government would then select whichever software/cost pair promised the most value. Because OS members expect new software to make their complementary goods more desirable, winning bids would usually come in well below cost. In order to make a profit companies would have to propose software that consumers actually wanted.\textsuperscript{265}

VII. CONCLUSION

The new commercial OS provides a potentially powerful vehicle for shared software development. At the same time, sharing also creates a de facto cartel that limits investment. This can stop OS output far short of its theoretical potential.\textsuperscript{266} Antitrust doctrine can only do so much to fix this problem. Because OS sharing delivers important benefits, the rule of reason will usually justify significant cartelization. The most judges can do—and it is a great deal—is to make sure that this cartelization does not spread unnecessarily. Judges should be particularly wary of OS collaborations that fall outside (a) operating systems, and (b) the five effort rule. Viral licenses should also receive scrutiny. We have argued

\textsuperscript{263} See supra text accompanying notes 71–76, 93–98.

\textsuperscript{264} Schmidt & Schnitzer, supra note 234, at 495 (“[A] government-sponsored research lab does not face the constraints of the market and has much less incentive to focus on customer needs and cost efficiency”).

\textsuperscript{265} Stephen M. Maurer & Suzanne Scotchmer, Procuring Knowledge, in 15 Advances in the Study of Entrepreneurship, Innovation and Growth 1, 26–27 (Gary Libecap ed., 2004).

\textsuperscript{266} See supra text accompanying notes 58–107.
that GPL-style licenses should normally be struck down and that even weakly viral, Mozilla-type licenses may not always be necessary.\footnote{See supra text accompanying notes 145–230.}

In the long run, only government intervention can eliminate the cartel effect. Initiatives to reset the OS/CS balance by enacting tax policies that make CS more profitable provide an elegant—if politically unsightly—solution to this problem. Direct government funding of OS projects would also work and is politically more palatable. Here, the principal danger is that poorly designed schemes could replace market signals of consumer need with government’s own, top-down vision.

The shock of traditional OS—that people both can and do produce valuable software for non-monetary reasons—has never entirely worn off. For this reason, many observers still see OS as a fragile bloom that must be protected and encouraged. This view is increasingly out of touch: today’s OS is commercial, hardheaded, and durable. For this reason, it is time to shift the goal from simply “promoting OS” to getting the OS/CS balance right—and then knowing when to stop. This new viewpoint will make our legal and policy choices harder but also much more interesting.
Table 1: Top 10 Eclipse Contributors\textsuperscript{268}

<table>
<thead>
<tr>
<th>Company (Membership Status)</th>
<th>Total (Share) of Commits 2001–2010</th>
<th>Total (Share) of Commits in 2010</th>
<th>Number of employee programmers since 2001</th>
<th>Company Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. IBM (Strategic Member)</td>
<td>3,079,053 (42.6%)</td>
<td>173,171 (16.7%)</td>
<td>388</td>
<td>Hardware, software, and consulting services.</td>
</tr>
<tr>
<td>2. Oracle (Strategic Member)</td>
<td>375,810 (5.2%)</td>
<td>84,880 (8.2%)</td>
<td>45</td>
<td>Business management software specializing in information management.</td>
</tr>
<tr>
<td>3. Intalio Inc. (Supplier Member)</td>
<td>368,686 (5.1%)</td>
<td>27,824 (2.7%)</td>
<td>10</td>
<td>Business management software and services including Intalio</td>
</tr>
<tr>
<td>4. Cloudsmith, Inc. (Strategic Member)</td>
<td>289,534 (4.0%)</td>
<td>76,047 (7.3%)</td>
<td>12</td>
<td>Web Services that help developers find, share, and use OS software\textsuperscript{270}</td>
</tr>
<tr>
<td>5. Actuate Corp. (Strategic Member)</td>
<td>204,828 (2.8%)</td>
<td>0 (0.0%)</td>
<td>58</td>
<td>Business management and reporting software</td>
</tr>
<tr>
<td>6. itemis AG (Strategic Member)</td>
<td>143,073 (2.0%)</td>
<td>49,678 (4.8%)</td>
<td>15</td>
<td>Model driven software development (MDSD) and Eclipse based tool chains.</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>#</th>
<th>Company</th>
<th>Membership</th>
<th>Shares</th>
<th># of Employees</th>
<th>Industry Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Borland Software Corp.</td>
<td>Strategic Member</td>
<td>139,308</td>
<td>5,895</td>
<td>Mass-market software.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.9%)</td>
<td>(0.6%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Red Hat Inc.</td>
<td>Solutions Member</td>
<td>117,280</td>
<td>14,884</td>
<td>Pre-packaged mass-market software.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.6%)</td>
<td>(1.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Tasktop</td>
<td>Solutions Member</td>
<td>116,734</td>
<td>10,391</td>
<td>Task management solutions for Eclipse.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.6%)</td>
<td>(1.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Soyatec</td>
<td>Solutions Member</td>
<td>125,371</td>
<td>50,070</td>
<td>Software and Eclipse solutions.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.5%)</td>
<td>(4.8%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
You can’t accuse class action commentary or case law of ignoring the Due Process Clauses. But you can accuse them of selective attention. Commentary and cases have focused almost exclusively on the due process rights of absent class members. But class defendants are entitled to due process, too. And in countless briefs filed at the state and federal levels, class action defendants complain that trial courts are ignoring their due process right to a fair hearing.

This Article undertakes a historical examination of one version of this argument, which goes like this: In civil proceedings, due process guarantees defendants a right to mount a full defense based on presentation of any probative “rebuttal evidence” that they “choose.” Trial courts, the argument goes, violate this right when they ration defendants’ opportunities to offer evidence rebutting class claims.

It’s easy to see why a court might do so in a class proceeding. When defendants want to rely on individualized evidence—that is, evidence unique to individual claims—it’s simply impossible to lump together large numbers of those claims into a class action and, at the same time, respect defendant’s rights to present that evidence. Take an employment discrimination suit alleging that a very
large class of employees was denied promotions because of its members’ gender. After investigating the unique features of each class member’s employment history, defendants might find evidence that a number of class members were refused promotions because they were bad employees, not because they were women. And if just one of these women had sued the defendant, the defendant would certainly be allowed to develop that evidence. Yet, if the class is very large—comprising, say, hundreds of thousands, or even millions, of women—providing defendants the opportunity to rebut each claim based on particularized evidence would be utterly infeasible. Hundreds of thousands of minihearings, after all, could take years or even decades to complete.5

Trial courts, accordingly, sometimes modify the way the claims can be proven so that aggregate proof of the claim is easier to manage.6 Consider, again, an employment discrimination suit. To smooth the way for class proceedings, courts might decide that class members meeting criteria easily gleaned from defendants’ records are entitled to an irrebuttable presumption that they were victims of discrimination.7 In this way, proof of the claims is simplified and easy to administer in an aggregate proceeding.8 And when courts do this, or something like it, class defendants often argue that courts thereby deny defendants a right to assert a “full defense” against the individual claims comprising the class action—violating defendants’ due process rights.

A few lower federal courts have invoked due process to reverse certification decisions modifying evidentiary rules in this way.9 Even so, this success comes in the face of an uncomfortable fact: class action defendants’ arguments are not rooted in the historical meaning of the Due Process Clause. Their arguments, instead, feed off intuitions about good policy—namely, growing (and, I think, reasonable) concerns that plaintiffs and courts are using the class action device to pressure institutional defendants to provide relief for undeserving claims.

Indeed, in their disinterest in history, if nothing else, class action defendants’ arguments echo the due process claims made by a very different group of litigants: welfare recipients. In the 1970s, welfare-rights advocates engineered their own procedural due process revolution, by persuading the Court that due process

---

5 See, e.g., In re Visa Check/Mastermoney Antitrust Litig., 280 F.3d 124, 149 (2d Cir. 2001) (Jacobs, J., dissenting) (“[T]here is good reason for considering the manageability of defenses in this [class] case [encompassing millions of claims]: even if each merchant’s claim took no more than a half a day to sort out, the damages phase of trial would last as long as the whole course of Western civilization from Ur.”).

6 See infra Part I.A.


9 See, e.g., W. Elec. Co. v. Stern, 544 F.2d 1196, 1199 (3d Cir. 1976) (noting that defendants must be afforded an opportunity to offer “any rebuttal evidence they choose” but declining to order the district court to reverse certification); see also Tager, supra note 3 (collecting cases).
required federal agencies to provide welfare recipients with certain procedural protections before taking away their benefits (the so-called “new property.”)\(^\text{10}\) Like today’s class action defendants, welfare advocates’ arguments were not based on the historical meaning of due process, but on intuitions about good policy—specifically, protecting vulnerable welfare recipients from welfare state bureaucrats.\(^\text{11}\) Indeed, there were no historical arguments to make. Up until that point, welfare benefits had been thought of as a statutory privilege that could be taken away without any process.\(^\text{12}\)

But the ground has shifted since the 1970s. Originalism has come into its own. There were, of course, originalists in the 1970s. And they complained (loudly) that the Brennan-era Court had constitutionalized a “view of desirable policy” with no roots in the historical understanding of the Due Process Clause.\(^\text{13}\) Then, though, originalists were a small, marginalized band. Today, originalism has grown in sophistication and influence, as the Court’s decision in District of Columbia v. Heller\(^\text{14}\) underscores. Unlike welfare rights advocates of the 1970s, class action defendants can’t afford to ignore originalism.

Today’s new originalists differ from those of the previous generation in some important ways.\(^\text{15}\) Even so, many (although not all) agree with the basic claim made by their precursors in the 1970s and 1980s: that when it comes to the vague and ambiguous Due Process Clause, early historical practice, as Frank Easterbrook put it in 1982, “lays down the baseline against which . . . arguments are measured . . . “\(^\text{16}\)

What are originalists to make of class action defendants’ due process arguments? This Article is the first to examine the historical record with that question in mind. And, for class action defendants, the verdict is a bad one.

The gist of class defendants’ due process claim is that the Due Process Clause guarantees defendants an adversarial presentation of any probative rebuttal evidence they chose. Yet, there is no long historical tradition protecting this “choice” in civil proceedings. True, throughout our history, due process has been

---


\(^\text{11}\) See, e.g., Richard A. Epstein, No New Property, 56 Brook. L. Rev. 747, 747–48 (1990) (“Goldberg represents one of those rare efforts to transform a set of academic and philosophical insights about the nature of property into the imperative language of constitutional law.”).


\(^\text{13}\) Easterbrook, supra note 12, at 125.

\(^\text{14}\) 554 U.S. 570, 625 (2008).


\(^\text{16}\) Easterbrook, supra note 12, at 92.
understood to guarantee a right to be heard before a judge “in... defence.” 17 But for most of the nineteenth century, the right to a hearing was not conceived as right to present a certain amount of evidence. It was conceived as a structural right to an independent judicial hearing. 18 Due process, in effect, selected the institution that should enter a final judgment in individual cases and controversies—a deliberative judiciary, acting free of legislative influence or control.

But as long as Congress left room for independent judicial judgment about the evidence that parties could present, the requirements of due process in the field of evidence gave out. Courts could, and did, bar defendants from presenting probative evidence when economy or public policy demanded. 19 In effect, judicial regulation of parties’ opportunities to offer evidence was seen as a subconstitutional matter left to courts’ reasoned discretion, based entirely on “equity, reason, and good sense.” 20

That’s not to say there is no historical support for class action defendants’ arguments. There is. But it is found in an awkward period: the Lochner era. The Lochner era’s biggest innovation—the recognition of substantive due process rights to freedom of contract and property—is its most (in)famous. 21 But in tandem with its development of substantive due process, the Lochner Court developed a new conception of procedural due process. Due process, the Court held,

---

17 See, e.g., Hurtado v. California, 110 U.S. 516, 532 (1884) (acknowledging due process requires means for “bringing the party against whom the proceeding is had before the court and notifying him of the case he is required to meet; for giving him an opportunity to be heard in his defence; for the deliberation and judgment of the court; for an appeal from this judgment to the highest court of the State, and for hearing and judgment there.”).


19 See infra Part III.


guarantees a right to present a “full defense” to a civil liability judgment, based on “every [probative] fact and circumstance” bearing on liability.22

The two innovations were related: the *Lochner* era’s substantive due process cases treated property—including intangible property or wealth—as a fundamental interest that deserved heightened constitutional protection.23 The procedural right to present “all the facts,” in turn, protected against the erroneous taking, through civil damages, of that interest. In effect, the *Lochner* Court’s recognition of a right to present “every fact and circumstance” bearing on liability was the Court’s first gesture toward its current recognition that fundamental interests demand heightened procedural protections.24 (And, recovering this *Lochner*-era contribution is an important side payoff of investigating the historical bona fides of class defendants’ claims).

*Lochner* has had some lasting impact.25 But *Lochner*’s procedural due process cases suffered the same dire fate as *Lochner*’s economic substantive due process cases with which they were linked. The post-New Deal Court has rejected the idea defendants resisting civil liability must be afforded a fixed right to present all the facts.26 Regulation of opportunities to present evidence in civil proceedings is restrained only by the loose requirement, articulated in *Connecticut v. Doehr*27 and *Mathews v. Eldridge*,28 that the regulation reflect a judicious trade-off between accuracy and efficiency.29

---

23 See, e.g., Truax v. Corrigan, 257 U.S. 312, 328 (1921) (holding that deprivation of intangible property without due process “can not be held valid under the Fourteenth Amendment”).
24 See, e.g., Santosky v. Kramer, 455 U.S. 745, 753-54 (1982) (holding that depriving someone of a fundamental liberty interest). The *Lochner*-era cases captured brief attention in the 1970s, during the mini-boomlet in legal scholarship on the Brennan-era “irrebuttable presumption doctrine,” which drew on the *Lochner* era decisions. Unfortunately, commentary on the historical roots of these cases was often cursory and commentators uniformly misinterpreted them in substantive due process terms. See, e.g., Note, The Irrebuttable Presumption Doctrine in the Supreme Court, 87 HARV. L. REV. 1534, 1539–40 (1974) (characterizing the *Lochner* cases as substantive due process decisions). As this Article shows, they grew out of, and transformed, nineteenth century decisions dealing with the implications of the due process hearing right for irrebuttable presumptions. See infra notes 217–31, 284–93.
25 See, e.g., Bernstein, supra note 21, at 60 (“Griswold, Roe, and their progeny can be traced back to *Lochner*.”).
26 See, e.g., Weinberger v. Salfi, 422 U.S. 749, 785 (1975) (rejecting a due process right to an individualized evidentiary hearing, including with respect to defenses to civil liability).
29 See, e.g., *Doehr*, 501 U.S. at 11 (applying the *Mathews* test and holding that due process, in the civil context, requires courts to weigh accuracy, the parties’ relative
In effect, modern procedural due process cases and the nineteenth century tradition converge on essentials: Neither construes due process as a fixed limit on the type or quantity of evidence presented in ordinary civil proceedings. Then and now, due process leaves a great deal of room for courts to regulate parties’ opportunities to present relevant evidence in civil proceedings in the service of equity and convenience. Class action defendants’ arguments are rooted in a brief, and brief-lived, deviation from this tradition—the *Lochner* era. If history provides the “baseline” against which constructions of due process should be tested, class action defendants’ claims are losers.30

This is not to say that defendants’ objections to proof rationing might not have merit on other grounds. For example, even if due process doesn’t require a particularized evidentiary hearing in every case, Congress might require one in claims arising under a particular statutory scheme. If it does, federal courts are presumably required to respect that entitlement.31 This Article’s narrow focus is on the cases where Congress (or in the case of state law, the state legislature) has not clearly displaced courts’ discretion to modify the way that claims can be proven in a class proceeding by statute. In those cases, due process, at least, does not categorically forbid courts from exercising that discretion in favor of rationing interests, and the “ancillary interest the government may have in providing the procedure or forgoing the added burden of providing greater protections”).

30 Easterbrook, *supra* note 12, at 92. As a result, I no longer subscribe to the view articulated in an early piece written for the Cato Institute before I became an academic. See Mark Moller, *The Rule of Law Problem: Unconstitutional Class Actions and Options for Reform*, 28 HARV. J.L. & PUB. POL’Y 855, 857 (2005) (sketching an intuition-based, nonoriginalist argument that class-specific evidentiary shortcuts are inconsistent with due process). That article’s focus was on the procedural avenues for removing state-filed class actions into federal court if defendants’ due process argument were taken seriously. *Id.* at 886. My investigation of the historical evidence examined herein has, however, convinced me that defendants’ due process argument is without merit.

31 On appeal to the Supreme Court in *Wal-Mart Stores, Inc. v. Dukes*, Wal-Mart successfully emphasized this type of due process argument, abandoning the broader formulation of its due process rights that it had advanced in the lower courts. It no longer suggested, as it had below, that due process generally restrains courts’ discretion to restrict defendants’ “choice” of probative “rebuttal evidence” for reasons of fairness or economy. Instead, Wal-Mart claimed that Title VII required courts to give defendants individualized rebuttal hearings—Congress, in other words, had affirmatively displaced courts’ discretion to exclude relevant evidence by statute; and because due process requires courts to follow the law, denying defendants this right therefore not only violates Title VII but denies Wal-Mart due process. *See, e.g.*, Brief for Petitioner at 42–43, *Wal-Mart Stores, Inc. v. Dukes*, No. 10-277 (U.S. Jan. 20, 2011) (arguing that “[u]nder Title VII, Wal-Mart is entitled to prove that an individual employment action is taken for non-discriminatory reasons” and to “rebut the claims of each individual class member”; the trial plan therefore “refashioned” Title VII by eliminating “statutory defenses” in violation of the Rules Enabling Act and due process). Questions can be raised about this argument, particularly its understanding of Congress’s intent is correct, its due process conclusion is unexceptional.
parties’ opportunities to present probative proof when doing so is necessary to facilitate class proceedings.

The argument proceeds in five parts. Part I summarizes class defendants’ due process defense. After an overview of originalism in Part II, Parts III and IV turn to examine the historical evidence. Part V concludes.

I. THE DUE PROCESS DEFENSE

A. Evidentiary Shortcuts

Class actions are more than a device for aggregating claims. They are a vehicle for amplifying the deterrent effect of the substantive law. And, they are a platform that enables self-appointed representatives of large groups (e.g., consumers or employees) to force corporate entities to the bargaining table to negotiate changes in the way they do business or treat their employees via settlement. To capitalize on these uses of the class action, courts must sometimes ration factual investigation and presentation of evidence to make class treatment possible.

Changes in the way that parties prove their claims is an inevitable artifact of the Rule 23 requirement that class claims coalesce into a cohesive core of evidence. Rule 23(b)(3), which authorizes class treatment of claims seeking monetary relief, makes this “proof cohesion” requirement explicit by requiring common issues to predominate. It is implicit in the text of Rule 23(b)(2), which authorizes class treatment for claims seeking group-wide injunctive relief and assumes “a homogenous and cohesive group with few conflicting interests among its members.”

The proof cohesion requirement ensures that a class proceeding will conserve enough administrative costs to justify sacrificing class members’ control over litigation of their own claims. It also ensures that representation of the class will not disadvantage some subset of the class—since, when evidence necessary to prove the claims coheres into a shared evidentiary unit, the interests of class members are much more likely to be aligned.

Yet, satisfying the proof cohesion requirement is often complicated by the default rules governing proof of most civil claims. In ordinary, nonclass claims, defendants are entitled to produce any probative evidence tending to rebut plaintiff’s claims. And, of course, the ultimate burden of production and persuasion usually lies with the plaintiff.

33 See FED. R. CIV. P. 23(a).
34 FED. R. CIV. P. 23(b)(3); Amchem Prods. Inc. v. Windsor, 521 U.S. 591, 623 (1997) (“The Rule 23(b)(3) predominance inquiry tests whether proposed classes are sufficiently cohesive to warrant adjudication by representation.”).
Yet, in classes aggregating a very large number of claims, slavish adherence to these default rules makes class proceedings far too costly and unwieldy. Take a simple case: an individual fraud suit, in which plaintiffs must prove that the defendant consciously misrepresented a fact and that the claimant reasonably relied on the representation, suffering an injury as a result. If, in a large class suit, plaintiffs were forced to prove these elements in individual hearings, aggregation of the claims would be utterly, absurdly infeasible.

Courts can solve this problem by modifying the evidentiary ground rules that apply—providing substitutes for usual forms of evidence in order to accommodate the scarcity of time and resources parties have to investigate massive inventories of claims. Alterations take two forms: legal presumptions and limits on defendants’ opportunity to rebut these presumptions with individualized evidence.

Presumptions facilitate class treatment of claims in a straightforward fashion. If proof of facts common to the class creates a presumption in favor of liability, class plaintiffs can prove their claims based on a cohesive core of common proof. By itself, though, this doesn’t solve the management dilemma. If defendants were nonetheless entitled to rebut the presumption with respect to individual class members, cases involving an aggregation of very large numbers of claims would remain utterly infeasible.

Courts deal with the problem posed by affirmative defenses by “devis[ing] . . . evidentiary shortcuts around management problems” they pose. These shortcuts, in effect, ration defendants’ opportunities to litigate these defenses, and they take at least three different forms.

“Irrebuttable Presumptions.” First, Allan Erbsen argues that some courts seem to dispense with a meaningful opportunity to raise defenses entirely—certifying classes based on a determination that plaintiffs, after proving easily discoverable facts common to the class, are entitled to what he terms an “irrebuttable evidentiary presumption” that the defendant is liable. Consumer fraud class actions provide the simplest example: there “plaintiffs often propose that when liability is premised on a consumer not knowing a certain fact, or relying on a given representation, the court should [irrebuttably] presume that all class members who acted in a specified manner”—say purchasing the product—“must have . . . been relying on a misleading statement.”

---

36 Erbsen, supra note 7, at 1012.
37 Id. at 1012–13.
38 Id. at 1013 n.24. There are many examples of federal and state courts certifying classes based on this kind of presumption. Singer v. AT&T Corp., 185 F.R.D. 681, 691 (S.D. Fla. 1998) (in a civil RICO action alleging fraud in the sale of phone lines, a “uniform written price representation . . . provide[d] a sufficient basis upon which reliance may be presumed”); Smith v. MCI Telecomm. Corp., 124 F.R.D. 665, 679 (D. Kan. 1989) (where salespeople signed written commissions containing common misleading terms, plaintiffs were entitled to a presumption that they relied on the misrepresentation, making the issue of their reliance “an objective inquiry common to the entire proposed class”); Liberty Lending Servs., Inc. v. Canada, 668 S.E.2d 3, 12 (Ga. Ct. App. 2008) (“In claims of fraud based upon written representations, the reliance element may sometimes be
do so, the claims are eminently manageable as a class, since discovery and trial will be limited to discrete, readily verifiable questions: Did the defendant send class members uniformly misleading representations?

To be sure, it's hard to find class certification decisions that expressly claim they are establishing irrebuttable presumptions. But, as Erbsen says, it's also hard to interpret at least some of these decisions otherwise. If the presumptions “were rebuttable the individual issues would remain in the case (subject to a flipped burden of proof) and would still present obstacles to adjudicating class actions.”

Some courts seem to characterize their decisions as rulings of substantive law—that parties lack a legal right to rebut the presumption. Others, though, do not purport to be interpreting the content of defendant’s legal defenses. They purport, instead, to ration the defendant’s procedural opportunity to litigate a defense—usually based on a threshold determination that defendants are unlikely to find widespread evidentiary support for their defense during merits discovery and that particularized hearings based on individualized evidence are therefore unlikely to achieve gains in accuracy significant enough to outweigh the cost of the

presumed." (citation omitted)); Varacallo v. Mass. Mut. Life Ins. Co, 752 A.2d 807, 816 (N.J. Super. App. Div. 2000) (defendant may be held liable upon proof defendant made uniform false written misrepresentations “with the intent that they be communicated to others for the purpose of inducing the others to rely upon them” (citation omitted)); Cope v. Metro Life Ins. Co., 696 N.E.2d 1001, 1008 (Ohio 1998) (“proof of reliance in this case may be sufficiently established by inference or presumption” that uniform written misrepresentations “influenced plaintiffs’ decision to borrow money from . . . defendant banks” (citation omitted)).

39 Erbsen, supra note 7, at 1013 n.23. This is not to say that all cases allowing the plaintiff class to utilize presumptions can be characterized in this way. Some dodge how to handle individualized defenses entirely by authorizing a class trial on the question of liability to the class (based on presumptive proof) and deferring a decision on whether class or individualized hearings are more appropriate for determining causation or damages. See, e.g., In re Visa Check/Mastermoney Antitrust Litig., 280 F.3d 124, 139–41 (2d Cir. 2001) (punting on whether individualized hearings might be necessary for determining damages). Here, the decision punts to the parties, via settlement negotiations, the difficult decision about how to fashion standards and an administrative mechanism for determining which members of the class are actually entitled to relief. Id. at 45. Because a number of circuits have forbidden issue classing unless proof of every issue necessary to establish liability satisfies Rule 23’s proof coherence requirement, this approach is used only in some circuits. See, e.g., Castano v. Am. Tobacco Co., 84 F.3d 734, 745 n.21 (5th Cir. 1996) (noting that Rule 23(c)(4) does not give courts the power to certify classes on discrete issues if the claim as a whole doesn’t satisfy the test for certification).

40 Varacallo v. Mass. Mut. Life Ins. Co. adopted an irrebuttable presumption of reliance as a matter of law in a consumer fraud action, while suggesting that in nonaggregated claims the presumption would be unavailable. 762 A.2d 807, 817–18 (N.J. Super. App. Div. 2000); see also Erbsen, supra note 7, at 1012–13 & n.24 (citing Varacallo as an example of this use of irrebuttable presumptions).
hearings. The result is a class proceeding that is litigated as though plaintiffs were entitled to an irrebuttable presumption as a matter of law.

Sampling. Second, some courts give defendants a right to individualized hearings in a representative subset of class claims and then apportion liability for damages to the remaining class members through sampling techniques. Here, in the paradigmatic case, trial is bifurcated. After an initial proceeding, in which prima facie case for defendants’ liability to the class is assessed by one jury, a randomly selected representative set of individual claims are litigated in a series of trials. Class damages are then calculated based on the average recovery in the litigated set of claims.

In effect, sampling is a hybrid of individualized hearings and “irrebuttable presumptions.” By litigating a sample of the class claims, defendants are given a restricted opportunity to litigate some of their defenses, after which an irrebuttable presumption arises that the rest of the class has suffered damages equivalent to the average claimant in the test sample.

“Formulaic Proof.” A third set of cases combines features of statistical evidence and irrebuttable presumptions, by resorting to what is sometimes termed “formulaic proof.” Courts in these cases dispense with some opportunities to rebut the presumption of liability established in the first phase of the proceeding. In

---

41 See, e.g., Waste Mgmt. Holdings, Inc. v. Mowbray, 208 F.3d 288, 297 (1st Cir. 2000) (declining to fault trial court because the possibility of widespread meritorious affirmative defenses was too “speculative” to warrant individualized discovery and evidentiary hearings); Blackie v. Barrack, 524 F.2d 891, 906 n.22 (9th Cir. 1975) (noting that where plaintiff’s evidence will “undoubtedly be conclusive as to most of the class,” and defendants may be able to defeat a showing of causation with respect to “a few individual class members,” the district court can exercise its discretion to limit discovery and opportunities for evidentiary hearings relating to defendants’ affirmative defenses in order to make the claims manageable). Some include a disclaimer: “If . . . evidence later shows that an affirmative defense is likely to bar claims against at least some class members, then a court has available adequate procedural mechanisms[,]” including decertifying the class or excluding class members who are likely subject to affirmative defenses. Smilow v. Sw. Bell Mobile Sys., Inc., 323 F.3d 32, 39–40 (1st Cir. 2003).


43 See, e.g., Dukes v. Wal-Mart Stores, Inc., 222 F.R.D. 137, 183 (N.D. Cal. 2004) (adopting a “formula approach” in lieu of individualized hearings in the context of promotions claim, but not adopting a similar approach in the context of equal pay claims); see also EEOC v. O&G Spring and Wire Forms Speciality Co., 38 F.3d 872, 874–75 (7th Cir. 1994); Domingo v. New Eng. Fish Co., 727 F.2d 1429, 1444 (9th Cir. 1984); Hameed v. Int’l Ass’n of Bridge, Structural & Ornamental Ironworkers, 637 F.2d 506, 518–22 (8th Cir. 1980).
this way, they, too, bear some resemblance to those that adopt “irrebuttable presumptions.”

But, rather than apportioning liability for damages based on a sample of tried cases, they apportion damages based on a formula derived from econometric analysis and “objective” (i.e., easily discoverable) data about class members, located in defendants’ records. The data serves as a rough predictor of the likelihood that the class member was victimized by defendant’s conduct and the extent of injury. As a result, this approach also bears some resemblance to trials that employ sampling. In both, the damages are calculated based on an approximating method, devised by experts, that may not perfectly capture the real variance in merit and value among individual claims.

Wal-Mart Stores, Inc. v. Dukes, a mammoth Title VII employment discrimination class action on behalf of an estimated 1.5 million female Wal-Mart employees, is a prime example of this last sort of case. Dukes involved a disparate treatment claim under Title VII—one seeking not only back pay, but also punitive damages governed by standards promulgated in the Civil Rights Act of 1991. And in International Brotherhood of Teamsters v. United States, the Court set out ground rules for proving a disparate treatment claim. In the absence of direct evidence of an intentional policy of discrimination, plaintiffs can rely on evidence of a pattern or practice of discrimination in order to show discrimination was the company’s “standard operating procedure.” If the plaintiffs carry that burden, a rebuttable presumption arises that each class member is entitled to appropriate monetary relief. In a second stage, a “district court must usually” give the employer an opportunity “to demonstrate that the individual . . . was denied an employment opportunity for lawful reasons.”

The key word here is “usually.” Defendants usually have an opportunity to demonstrate lawful reasons for an adverse employment action in a mini-hearing in which the defendant offers (if it can) evidence that an individual employee lost a job or promotion for nondiscriminatory reasons (absenteeism or substance abuse,

---

44 See, e.g., Erbsen, supra note 7, at 1013 n.24
45 See, e.g., id. (noting that courts allow “objective evidence (such as the defendant’s business records) of how the class member acted, which is generally easy to present in a class action, . . . [to] substitute for subjective proof . . . which is generally difficult to present in a class action”).
46 This Article was written before the Supreme Court’s recent decision in the case and my focus below is on the arguments presented in the Ninth Circuit. See, e.g., 603 F.3d 571 (9th Cir. 2010) (en banc).
47 Compare id. at 578 n.3, with id. at 627 (Ikuta, J., dissenting).
50 Id. at 336–43.
51 Id.
52 See id. at 361–62.
53 Id. (emphasis added).
say). Yet, as the Dukes trial court noted, with a class of 1.5 million women, “the traditional Teamsters mini-hearing approach is not feasible here.”

Instead, the court crafted a “formula approach” for determining the back pay award for class members, in which “economic models” crafted by a special master based on expert testimony would employ “performance evaluation scores, tenure, and positions held[,]” as well as “objective applicant data documenting which class members were interested in each . . . promotion,” as “rough” bases for identifying the women who lost promotions because of gender discrimination.

Under this approach, expert analysis of data drawn from Wal-Mart’s business records replaces the usual Teamsters minihearing. Class members who fit the “formula” are presumed to have been denied promotions based on their gender. Wal-Mart would have no opportunity to offer testimony about an employee’s actual performance or a manager’s motivations for denying the promotion. It would be limited to contesting the statistical “pattern or practice” evidence and the lost pay formula.

B. Enter the Due Process Clause

Judicial modification of evidentiary rules raises an obvious question about courts’ authority to do so. And it can be analyzed in different ways. It can be analyzed as a problem implicating Rule 23’s text, as a problem implicating courts’ statutory authority, or as a problem implicating the due process clause.

Unfortunately, Rule 23’s text, on this point, quickly gives out. Rule 23 requires only that proof parties would independently choose to present coheres, but parties’ choices are limited by the kinds of evidence the law allows them to present. And the evidence that parties can present is dynamic not static—rules of evidence change over time, and sometimes differ based on the kind of claim asserted or the procedural context in which it is asserted.

Here, a second concern that takes us beyond Rule 23 swims into view. What if Congress modifies the evidence that parties may present in a class proceeding—eliminating parties’ privilege to present certain kinds of evidence? Alternatively,
what if Congress delegates to courts the power to fashion new or different rules of evidence, or adjust opportunities for presenting evidence, in class proceedings?

Consider Wal-Mart’s argument to lower courts in Wal-Mart Stores, Inc. v. Dukes. There, Wal-Mart complained that the court’s certification order dispensed with opportunities for presenting individualized proof of defenses to back-pay liability (in other words, that one or more members of the class were not victimized by discrimination). But what if Congress decreed that, in gender discrimination claims proceeding as a class action, defendants are simply not entitled to present all the evidence in support of that defense—so long as plaintiff makes out a prima facie case of disparate treatment, defendants, where equity demands, may be held liable to individual class members based on less accurate formulaic proof? If so, plaintiffs surely can’t complain that the claims raise individualized questions of proof that defeat a predominance finding. That is proof that courts tasked with formulating the evidentiary rules in this area have decided, pursuant to legislative authorization, that defendants simply aren’t entitled to present in large class actions.

This point was made by an amicus brief submitted to the Ninth Circuit in Dukes by a collection of nonprofit legal advocacy groups, including the Lawyer’s Committee for Civil Rights under the Law and the NAACP Legal Defense and Education Fund. Title VII, they argued, vests federal courts with “broad discretion to devise comprehensive relief ‘in light of the large objectives of the Act.’” That discretion, which they contended hadn’t been displaced by amendments contained in the Civil Rights Act of 1991, includes the power to fashion “formulaic [monetary] relief”—that is, awards of back pay based on formulaic proof of discrimination that is less sensitive to individualized facts about each class members’ employment history—“when the class is so large that it would not be economical to have individual hearings.”

These amici, in other words, framed the problem as one of legislative delegation. Had Congress delegated to courts the power to fashion second-best substitutes for particularized evidentiary hearings when necessary to accommodate large-scale class proceedings? They thought it was apparent Congress, through Title VII, had done so.

In an attempt to leapfrog over thorny questions of legislative authorization entirely, the defendant’s argument in the lower courts pinned their objection to

60 Dukes v. Wal-Mart Stores, Inc., 509 F.3d 1168, 1175 (9th Cir. 2007).
62 Id. at 19 n.7 (citation omitted).
63 Id. at 5, 19.
64 Id. at 4–5.
class-specific judicial alteration of evidentiary burdens on the due process clause.  

At the historical core of procedural due process is the idea that parties have a right to a “day in court.” And Wal-Mart, like many class defendants, read into this right an implied corollary. To the extent existing law provides defendants with a defense to liability (here, the defense that one or more members of the class were not victimized by discrimination), defendants, to have a meaningful hearing, must have the opportunity to present a “full defense” based on “any rebuttal evidence they choose,” including all probative individualized evidence. In this view, due process fixes a set of evidence that defendants can present at trial in support of their defense—“all probative evidence” relating to their legal entitlement—and guarantees defendants a procedural right to investigate and present that evidence to a fact finder in a particularized hearing.

C. Beyond Mathews v. Eldridge

The phenomenon of “evidentiary shortcuts” in class action practice invites more than one vein of analysis. It raises difficult questions about the extent to which due process constrains courts power to modify defendants’ entitlement to present certain kinds of proof. But it also raises statutory questions. Assuming, say, evidentiary shortcuts satisfy due process, does a statute like Title VII implicitly authorize courts to make these sorts of modifications? To what extent?

These are separate questions and exploring each would occupy a standalone article. This Article’s focus is on the first and broadest question: the extent to which evidentiary shortcuts, in general, implicate due process.

Even here, the analysis requires picking and choosing among different ways to think about due process. For example, should we confine analysis to recent procedural due process precedents, such as Mathews v. Eldridge and Connecticut v. Doehr? Plaintiffs and their amici in Dukes v. Wal-Mart Stores, Inc. thought so. Mathews, of course, announced a set of factors that must be balanced when deciding the procedures due when agencies deprive welfare recipients of their entitlement to benefits. Doehr extended Mathews to procedures adopted in civil proceedings. Together both cases require courts to strike a fair balance between

---

65 Wal-Mart narrowed its due process argument considerably once it was before the Supreme Court, see supra note 32 and accompanying text, and that very different due process argument is beyond the scope of this Article.


67 Principal Brief for Appellant, supra note 4, at 45 (citation omitted).


70 See Principal Brief for Appellant, supra note 4, at 45 (characterizing plaintiff’s approach as violating due process).

71 424 U.S. at 335.

72 Doehr, 501 U.S. at 9–11.
avoiding an erroneous deprivation of the parties’ interests and respecting the fiscal
or administrative interests of the state.73

Class plaintiffs argued that, in a class of such gargantuan size, these factors
cut in favor of an evidentiary shortcut74 and an en banc panel of the Ninth Circuit
agreed. While the formulaic approach to proof may prove somewhat less accurate
than individual hearings, said the Ninth Circuit, “adversarial resolution of each
class member’s claim would pose insurmountable practical hurdles . . . since
[millions of] individual adversarial determinations of claim validity would clog the
docket of the district court for years.”75

While it is tempting to dismiss defendants’ claims out of hand based entirely
on Mathews and Doehr, as the Ninth Circuit did, an exclusive focus on these cases
is too narrow. The Court has never held that the Mathews/Doehr test is the
universal measure of procedural due process. Indeed, the Court’s recent procedural
due process cases strongly suggest that it doesn’t think this is the case. In Taylor v.
Sturgell,76 the Court, for example, resisted employing the Mathews/Doehr test
when analyzing the extent to which prior litigants can be deemed to “virtually
represent” nonparties, precluding them from litigating related claims. Instead the
Court invoked the “deep-rooted historic tradition that everyone should have his
own day in court,”77 a tradition the Court implicitly tied to the Due Process Clause.
In cases like these, the Court hints that procedural due process is not endlessly
elastic. It has some fixed core content.

The obvious source of this fixed content is the original meaning of the Due
Process Clause. Originally understood, due process may “lock in” certain
procedural entitlements, insulating them from the Mathews/Doehr weighing
approach. Justice Scalia, for example, made just such a claim in his dissent in
Hamdi v. Rumsfeld: “The gist of the Due Process Clause, as understood at the
founding and since,” he argued, “was to force the Government to follow those
common-law procedures traditionally deemed necessary before depriving a person
of life, liberty, or property.”78 “Whatever the merits of th[e] [Mathews balancing]
technique when newly recognized property rights [such as welfare benefits] are at

73 See, e.g., id. at 11 (“[A]ny burden that increasing procedural safeguards entails
primarily affects not the government, but the party seeking control of the other’s property
. . . . [N]onetheless, due regard [must be given] for any ancillary interest the government
may have in providing the procedure or forgoing the added burden of providing greater
protections.”).
74 See, e.g., Opening Brief for Appellees at 57, Dukes v. Wal-Mart Stores, Inc., 509
F.3d 1168 (9th Cir. 2007) (No. 04-16688), 2004 U.S. 9th Cir. Briefs LEXIS 675120 (“Wal-
Mart makes no effort to apply . . . the standards for evaluating potential due process
violations.”).
75 See, e.g., Dukes v. Wal-Mart Stores, Inc., 603 F.3d 571, 627 (9th Cir. 2010) (en
banc) (quoting Hilao v. Estate of Marcos, 103 F.3d 767, 786–87 (9th Cir. 1996)).
77 Id. at 892–93 (quoting Richards v. Jefferson Cnty., 517 U.S. 793, 798 (1996)).
issue (and even there they are questionable), it has no place where the Constitution and the common law already supply an answer.”

Might the original meaning of the Due Process Clause include a fixed right to present all probative evidence in defense of civil liability? The idea is surprising. But the premise for this claim shouldn’t raise too many eyebrows—not any more. Originalism is having a heyday on the current Court—its rescue of the Second Amendment from desuetude is ample evidence of that. Arguments against class action defendants’ due process claims accordingly ignore the possibility that the original understanding of due process supports class action defendants at their peril. If plaintiffs want to advance a rich and deeply satisfying account of courts’ power to devise evidentiary shortcuts in the class context, they have to engage with the history of the Due Process Clause.

The rest of this Article turns to investigate that history.

II. DUE PROCESS AND CONSTITUTIONAL CONSTRUCTION

Before turning to the history, it’s important to be clear why, and when, history matters for originalists. The Due Process Clauses are deeply ambiguous—they have multiple possible meanings, as Section A underscores. Many originalists, in turn, think that early historical practice has a strong claim to guide ambiguous texts, for reasons canvassed in Section B.

Together these two sections review why, and how, history serves an interpretive guide for many originalists. Parts III and IV then turn to the history.

A. The Ambiguous Due Process Clause

Originalists think the Constitution’s meaning is fixed by the conventional preratification semantic meaning of the Constitution’s text, revealed by sources, like dictionaries, evidencing contemporary usage. Yet, some constitutional texts are “ambiguous”—like the word “bank,” they were conventionally associated with two or more different meanings. Other texts are “vague”—their meaning takes the form of an abstract concept or principle whose specification in individual circumstances admits of difficult borderline cases and therefore demands a significant degree of judicial judgment or “construction” to put into practice.

79 Id. at 575–76.
83 Id.
Some, finally, are both. The text might be construed to enact either one principle or another, and some or all of the principles the text might enact are vague.84

The Due Process Clause is a quintessential example of a constitutional text that is both ambiguous85 and vague.86 Its ambiguity is rooted in its antecedents. Lifted from New York’s statutory Bill of Rights, its text is a mishmash that evokes three precursors, each of which had a different tradition of eighteenth-century interpretation.87

The first is a 1354 statute, which provided that no person may be evicted from lands or leases, disinherited, imprisoned, or executed, “without being brought in Answer by due Process of the Law.”88 As Frank Easterbrook says, this “had a plain enough meaning.”89 “[C]ourts could not proceed in any important civil or criminal case without ‘process,’ that is, without service of a writ on the defendant giving him an opportunity to appear . . . .”90

84 See, e.g., Solum, supra note 81, at 73. The distinction between “interpretation” and “construction” is among the most hotly debated terminological questions among originalists. For a discussion of the technical distinction that has grown up around the terms in originalist scholarship, see Lawrence B. Solum, The Interpretation-Construction Distinction, 27 CONST. COMMENT. 95, 100–09 (2010). In this Article, in the interest of avoiding getting bogged down in a lengthy, technical discussion of the evolving interpretation/construction distinction in originalist scholarship, I use the terms interpretation and construction interchangeably in a loose, colloquial sense, to mean the process of assigning legal content to constitutional text when preratification evidence of contemporary meaning (drawn from dictionaries, treatises, or other evidence of preratification usage) and intratextual clues available on the face of the constitution are insufficient to settle debates about the principles, practices, and rules that the text is meant to enact. This sense is, by the way, similar to my understanding of Keith Whittington’s use of the term “construction.” See, e.g., Keith Whittington, Constructing a New American Constitution, 27 CONST. COMMENT. 119, 120 (2010) (“The process of constitutional construction is concerned with fleshing out constitutional principles, practices and rules that are not visible on the face of the constitutional text and that are not readily implicit in the terms of the constitution.”); see also John O. McGinnis & Michael Rappaport, Original Methods Originalism: A New Theory of Interpretation and the Case Against Construction, 103 Nw. U. L. REV. 751, 780 (2009) (noting Whittington seems to use the term “construction” when he takes an “external” view of constitutional interpretation, and does not clearly use the term in the “positive and normative” sense employed by originalists like Lawrence Solum).


87 Easterbrook, supra note 12, at 95–100 (discussing the genesis of the Fifth Amendment’s due process clause); Rosenthal, supra note 85, at 29–34 (same).

88 28 Edw. III, c. 3 (1354) (Eng.).

89 Easterbrook, supra note 12, at 96.

90 Id.
The second statutory antecedent was Parliament’s declaration, following the dissolution of the court of Star Chamber,

[t]hat neither his Majestie, nor his Privie Councell, have or ought to have any Jurisdiction power or authority . . . to examine or drawe into question determine or dispose of the Lands Tenements Herditaments Goods or Chattels of any the Subjects of this Kingdome But that the same ought to be tried and determined in the ordinary Courts of Justice and by the ordinary course of the Law.\(^91\)

The import of the statute, Blackstone thought, was also plain. “The king,” he wrote, “may erect new courts of justice; but then they must proceed according to the old established forms of the common law,” including “[n]ot only the substantial part . . . of the law, but also the formal part, or method of proceeding.”\(^92\) Modes of proceeding, he said, “cannot be altered but by parliament.”\(^93\)

The third antecedent was Magna Carta’s guarantee that “[n]o free man shall be arrested or imprisoned, or disseised or outlawed . . . or in any way victimised . . . except by the lawful judgment of his peers or by the law of the land,”\(^94\) which the Framers almost uniformly equated with the Due Process Clause.\(^95\) Some, including many pamphleteers in the ferment preceding the American Revolution, thought Magna Carta’s “law of the land” provision locked in a host of natural law rights, rooted in the common law.\(^96\) Others have interpreted it as a very narrow guarantee to a few particular procedures rooted in longstanding legal practice, including “indictment and jury trial.”\(^97\)

But many in the framing generation understood it in more specific terms than the first camp, but in more general terms than the second: as a right to a hearing before an independent judiciary prior to the deprivation of private rights.\(^98\) As

\(^91\) 16 Car., ch. 10, § 3 (1640) (Eng.).
\(^92\) 1 WILLIAM BLACKSTONE, COMMENTARIES ON THE LAW OF ENGLAND 142 (George Sharswood ed., Lippencott Co. 1893).
\(^93\) Id.
\(^94\) Magna Carta (1215 & 1225), reprinted in R ALPH V. TURNER, MAGNA CARTA THROUGH THE AGES, app. at 226, 231 (2003).
\(^96\) On some readings, Lord Coke held this view, and the phrase was interpreted this way by many pamphleteers in lead-up to the American Revolution. Id. at 614–21, 662 (noting Lord Coke’s influence on colonial law).
\(^97\) Easterbrook, supra note 12, at 97 (“Coke’s natural law was a rather tame creature, satisfied with the inalienable rights to indictment and jury trial”).
\(^98\) For one preratification association between the “law of the land” and judicial independence (and impartiality), see JOHN ADAMS, INSTRUCTIONS TO THE TOWN OF BRAINTREE: TO THEIR REPRESENTATIVE, 1765, in 3 THE PAPERS OF JOHN ADAMS 465, 466–67 (Charles Francis Adams ed., 1851) (expansion of admiralty jurisdiction is “directly repugnant to the
authorities put it in the early decades after ratification, deprivation of rights according to the “law of the land” requires a proceeding “in its nature judicial”\(^9\)—that is, a “forensic” proceeding, \(^10\) which, said Daniel Webster, “hears before it condemns; . . . proceeds upon inquiry, and renders judgment only after trial.”\(^11\)

In effect, according to this construction of Magna Carta, it and the 1354 ban on ex parte proceedings collapse into different historical statements of a single principle. Parties have a right to a judicial hearing and independent judicial judgment prior to deprivation of core individual rights. Due process, in other words, identified the institution that must do the depriving. And that institution must be a deliberative, independent judiciary.

This reading not only had some warrant in the conventional construction of the clause’s antecedents, but, as James Thayer noted, accorded with the medieval semantic meaning of the phrase “due process of law.” In the “older days,” he said, rights were associated with custom, or with prepolitical principles of natural

---

Great Charter itself” because “[n]o juries have any concern there!—The Law, and the Fact, are both to be decided by the same single Judge, whose commission is only during Pleasure, and with whom, as we are told, the most mischievous of all Customs has become established, that of taking Commissions on all Condemnations . . . .”) (emphasis added).

This construction is also evident in Hamilton’s (often misunderstood) speech on the Fifth Amendment’s Due Process Clause’s immediate precursor—the statutory Due Process Clause adopted by the New York legislature in 1787. Hamilton equated this Clause with Magna Carta’s law of the land provision. And, he argued, the New York Assembly’s deprivation of loyalists’ right to hold public office violated New York's statutory due process clause because the deprivation was secured legislatively, rather than through the “proceedings of the courts of justice.” Alexander Hamilton, New York Assembly: Remarks on an Act Regulating Elections, in 4 THE PAPERS OF ALEXANDER HAMILTON 35–36 (Syrett & Cooke eds., 1962) (the “law of the land” does not “include an act of the legislature”; rather, deprivation of a civil right to hold public office requires “presentment and indictment, and process of outlawry” in “courts of justice”). For a discussion of Hamilton’s speech, which is sometimes misunderstood by modern readers, see Rosenthal, supra note 85, at 30–31.

Hamilton’s view is consistent with his vision of separation of powers, in which an independent judiciary would serve as an aristocratic check on popular assemblies. See PAUL CARRESE, THE CLOAKING OF POWER: MONTESQUIEU, BLACKSTONE, AND THE RISE OF JUDICIAL ACTIVISM 197–207 (2003).

\(^9\) The phrase is Justice Marshall’s in Marbury v. Madison, where he famously noted that “[t]he question whether a right has vested or not, is, in its nature, judicial, and must be tried by the judicial authority.” 5 U.S. (1 Cranch) 137, 167 (1803). Later courts connected the idea with the Due Process Clause. See, e.g., Taylor v. Porter & Ford, 4 Hill 140, 146 (N.Y. 1843) (holding deprivation by “law of the land” requires that it be “ascertained judicially that he has forfeited his privileges, or that some one else has a superior title to the property he possesses, before either . . . can be taken from him. It cannot be done by mere legislation.”). For additional antebellum authorities linking “due process” with deprivation of rights through a judicial judgment, see infra notes 206–32 and accompanying text.

\(^10\) Taylor, 4 Hill at 147.

justice, and not, in our modern, positivist sense, with commands of a sovereign.\textsuperscript{102} By contrast, “[l]eges,” he says, were “what is instituted by rulers and kept up by the people . . . for settling particular controversies” and were closely associated with the “judicature” or modes of “trial.”\textsuperscript{103} Our due process of law phrase, said Thayer, “comes down out of the midst of all of this,” as a phrase whose original, ancient meaning required the settlement of disputes through the process of the popular “Courts of Justice.”\textsuperscript{104}

In the end, based on its historical antecedents, four competing possible meanings of the phrase “due process” swim into view. At its narrowest, it might simply ban ex parte judicial proceedings. Slightly more broadly, it might guarantee a narrow set of core procedural rights—like a right to criminal proceedings by indictment and jury trial—associated with the most restrained interpretations of Magna Carta. More broadly still (and more vaguely) it might guarantee a hearing according to settled law and procedures fixed by common law tradition. At its broadest and most vague, it might protect a variety of fundamental common law liberties.

Or, last, it might do something somewhat different than any of these three constructions. It might, consistent with some eighteenth century interpretations of Magna Carta and the most ancient semantic meaning of the words “due process,” stand instead for an important structural principle—that people’s rights to life, liberty, and property require an \emph{independent} judicial hearing free from the control of the political branches.\textsuperscript{105}

\textbf{B. Construing “Due Process”}

Some ambiguous texts can be clarified by looking to intratextual evidence and surrounding context in which the ambiguous phrase is used.\textsuperscript{106} The Due Process Clause is not one of these. As a result, due process presents one of the thorniest of interpretive problems of any constitutional provision.


\textsuperscript{103} \textit{Id.} at 199–201.

\textsuperscript{104} \textit{Id.} at 200–01. When the Constitution was ratified, the medieval meaning of the term existed alongside other, newer meanings, some of which were more expansive. See, e.g., Gedicks, \textit{supra} note 95, at 614–21. The association between due process and a vague prohibition on “non-arbitrary” law came later after the Fifth Amendment’s ratification. See David P. Currie, \textit{The Constitution and the Supreme Court: The First Hundred Years} 112 n.43 (1985) (tracing the introduction of the “non-arbitrariness” concept to Justice Johnson’s opinions in \textit{Bank of Columbia v. Okely}, 17 U.S. (4 Wheat.) 235 (1819) and \textit{Livingston v. Moore}, 32 U.S. 469 (1833)).

\textsuperscript{105} One can see some of the basic procedural requirements everyone agrees is at the heart of “due process”—the right to notice, jury trial, habeas—as different applications of the larger principle that an independent judicial hearing must be interposed between the legislative and executive branches and deprivations of a person’s life, liberty, and property.

\textsuperscript{106} Barnett, \textit{supra} note 86, at 268 (“Most terms are not ambiguous in context.”).
Many originalists attempt to construe what the Due Process Clause means by referring to postratification judicial interpretations of the Due Process Clause.\textsuperscript{107} For these originalists, history serves as a plausible fallback interpretive guide for several reasons. First, when a text could reflect one of several different principles, evidence of contemporaneous expectations about the applications of the text revealed in treatises and postratification case law “help us determine which . . . principle was adopted.”\textsuperscript{108}

At the same time, there is evidence that the Framers thought that courts should rely on an early practice of interpretation to resolve uncertainty about the meaning of a constitutional text. “James Madison and other prominent founders,” writes Caleb Nelson,

\begin{quote}
\textit{did not consider the Constitution’s meaning to be fully settled at the moment it was written. They recognized that it contained ambiguities and that subsequent interpreters would help ‘fix’ its meaning . . . . Once practice had settled upon one of the possible interpretations of a disputed provision, they expected that interpretation to persist.}\textsuperscript{109}
\end{quote}

\textsuperscript{107} See, e.g., Easterbrook, \textit{supra} note 12, at 100–09 (using historical judicial interpretations as evidence of the meaning of due process); Rosenthal, \textit{supra} note 85, at 37–41 (examining the meaning of the Fourteenth Amendment’s Due Process Clause based on the “state of decisional law” at the time of its ratification).

\textsuperscript{108} See John O. McGinnis & Michael Rappaport, \textit{Original Interpretive Principles as the Core of Originalism}, 24 CONST. COMMENT. 371, 379 (2007). Interpreters can eliminate some principles as candidates for the “meaning” of due process by focusing not only on what early courts said, but also on what they didn’t say. Arguments from silence appear as early as Justice Bradley’s decision in \textit{Davidson v. New Orleans}, 96 U.S. 97, 103–04 (1877): “It is not a little remarkable, that while this provision has been in the Constitution of the United States, as a restraint upon the authority of the Federal government, for nearly a century, and while, during all that time, the manner in which the powers of that government have been exercised has been watched with jealousy, and subjected to the most rigid criticism in all its branches, [the Due Process Clause] has rarely been invoked in the judicial forum or the more enlarged theatre of public discussion.” Given that history, Justice Bradley concluded, litigants proposing aggressive interpretations of due process must have “some strange misconception of the scope of this provision . . . .” \textit{Id.} at 103–04.

When the historical record is devoid of a particular interpretation of the due process clause, despite widespread practices that would seem to \textit{violate} that interpretation, that’s fairly powerful evidence that the legal community of the time did not associate due process with that interpretation. The argument is even stronger when courts are \textit{not} silent with respect to other practices that offend \textit{competing} interpretations of due process. Where courts are willing to strike down a set of practices that offend one interpretation, but do nothing with respect to practices that offend a second, the natural inference is that courts associated due process with the first interpretation, not the second.

Not all originalists, though, agree that early precedent should limit the process of construing vague or ambiguous texts. Randy Barnett, for example, argues that abstract and vague constitutional provisions “delegate discretion to judges” to formulate constructions that “enhance[] the legitimacy of the Constitution.”110 Yet, even assuming textual vagueness should be treated as a delegation of interpretive discretion, reference to historical practice can be useful—not simply as evidence of what the framers intended when they enacted the clause, but as evidence of the best, or most wise, interpretation.

The idea that history can reveal the most wise interpretation is associated with Edmund Burke, who believed “traditions are likely to be wise” because they “represent the judgment of not just a single person, but of countless people over” time.111 As such they mimic “some of the advantages of markets, reflecting the assessments of many rather than few.”112 This is not to say that the imprimatur of tradition is conclusive evidence of wise practices—long practice can reflect prejudices, enduring errors about the state of the world, or thinking that has become outmoded by fundamental changes in the way the world works.113 But the sanction of age is at least strong presumptive evidence that a practice serves important “social interests.”114

As a result, originalists of many stripes can agree about the value of history as an interpretive source. A longstanding tradition of interpretation, with roots in early practice, can reveal the way framers understood ambiguous texts, when those texts cannot be disambiguated by reference to context. Construing texts in light of

---

112 Id. See generally A.C. Pritchard & Todd J. Zywicki, Finding the Constitution: An Economic Analysis of Tradition’s Role in Constitutional Interpretation, 77 N.C. L. REV. 409 (1999) (stating courts should use common law traditions to construe ambiguous constitutional clauses because common law traditions are likely to be efficient).
113 See Sunstein, supra note 111, at 2106.
114 Id. See, e.g., infra notes 390–394 and accompanying text.

Below I characterize traditionalist arguments against class defendants’ due process defense as “originalist” arguments. See, e.g., infra notes 395–97 and accompanying text. Some originalists, who limit the term “originalism” to the practice of interpreting the original semantic content of the Constitution’s text, may resist this characterization. See, e.g., Lawrence B. Solum, District of Columbia v. Heller and Originalism, 103 NW. U.L. REV. 923, 967 (2009) (characterizing arguments that do not focus on the semantic meaning of the text as beyond the “domain” of originalism). However, the term “originalism” is conventionally used much more broadly. It is used to encompass the family of fellow-travelling interpretative approaches that many self-described originalists tend to rely on, including text, history, and—when text gives out—tradition to make sense of the Constitution. See Michael W. McConnell, Active Liberty: A Progressive Alternative to Textualism and Originalism?, 119 HARV. L. REV. 2387, 2415 (2006) (reviewing Stephen Breyer, Interpreting Our Democratic Constitution (2005)) (characterizing originalism as a cluster of interpretative commitments—to text (first and foremost), as well as history and “tradition”—driven by an overarching normative commitment to judicial restraint).
early historical practices also accords with the way many framers thought the Constitution should be interpreted. And, regardless of the framers’ expectations, the fact a modern interpretation finds support in historical interpretations may suggest it is a wise interpretation that responds to important social needs. As such, it can appeal even to many moderate originalists who concede an inevitable role for judicial choice and judgment in the construction of unclear constitutional provisions.

In the next two parts, I turn to examine the history of interpretation of the Due Process Clause, with a focus on its implications for judicial power to ration opportunities to present proof in the service of judicial economy. That examination will encompass both the background practices that implicate competing constructions of due process in the evidentiary field, and inferences about original understanding that can be drawn from judicial reactions to those practices.

As we will see, that history reveals two traditions: an early tradition, which lasted for a century after the ratification of the Constitution and decisively rejected the idea that due process guaranteed an opportunity to present any specific quantum of evidence in civil proceedings; and a brief-lived departure from that tradition, during the Lochner era, where the Court recognized that civil defendants have a due process right to present all probative rebuttal evidence.

Because the earlier tradition arose shortly after the ratification of the Fifth Amendment, it has a much more powerful warrant to inform construction of the Fifth Amendment’s Due Process Clause. And because the early tradition is very close to the way modern courts approach due process, and the Lochner-era approach is a brief-lived deviation from that tradition, the early tradition also has the strongest Burkean claim to inform construction of what due process, under both the Fifth and Fourteenth Amendments, requires.\textsuperscript{115}

\textsuperscript{115} For Burkeans, inconsistent interpretations of a clause do not render history inconclusive, so long as one interpretation has “stood the test of time” while others have not. After all, if the record shows that early courts gravitated toward one interpretation, which persists today, and others adopted a different interpretation that reigned for a brief period of time only to be decisively abandoned, that’s powerful Burkean evidence that the first interpretation, not the second, serves important “social interests.”

It bears noting that while this argument treats the Fifth and Fourteenth Amendment Due Process Clauses together, there is a growing body of literature that suggests the two Clauses should be interpreted differently. See, e.g., Ryan C. Williams, The One and Only Substantive Due Process Clause, 120 YALE L.J. 408 (2010). Even so, the view that the original “meaning” of the Fourteenth Amendment’s Due Process Clause was just as ambiguous as the Fifth’s when it was ratified remains dominant and, on that prevalent view, the evidence presented below ought to bear on the construction of both clauses. See Lawrence Rosenthal, The New Originalism Meets the Fourteenth Amendment: Original Public Meaning and the Problem of Incorporation, 18 J. CONTEMP. LEGAL ISSUES 361, 407–08 (2009) (noting the continuing ambiguity of the meaning of “due process” at the time of the Fourteenth Amendment’s ratification, as well as evidence that the ratifiers expected its ambiguity would be construed in light of methods other than reference to its original public meaning at the time of ratification).
III. THE EARLY TRADITION

The idea that due process restrains the judicial power to modify, or restrict, parties’ opportunities to present probative evidence never seemed to have occurred to antebellum courts. It’s hard to see why it would. The common law of evidence inherited by American courts, as Section A shows, reflected a variety of judge-made restrictions on parties’ ability to present probative evidence—all in the service of promoting judicial economy and limiting the power of the jury. While early American courts modified the common law of evidence in various ways, they didn’t change their conception of judicial power. Like their immediate English precursors, as Section B shows, they continued to adapt the common law of evidence in the service of equity and convenience—sometimes expanding, and sometimes restricting, parties’ rights to present evidence. And they gave no inkling that the Due Process Clause imposed any restraints on their power to do so.

Instead, as Section C shows, courts focused their attention entirely on legislative interference in the law of evidence, by limiting legislative efforts to dictate the effect of certain kinds of proof. Of the four possible constructions of the Due Process Clause competing for judicial attention following ratification of the Fifth Amendment, Section D concludes, the pattern and reasoning of these early decisions is most consistent with one. Due process, nineteenth-century courts thought, guaranteed a right to a judicial hearing. But that right was not conceived as a personal right to present certain kinds of evidence. It was a structural right, which selected the appropriate institution that should make the final judgment in individual case and controversies: a deliberative judiciary, acting free from legislative influence or control.

And at the core of the independence secured by the Due Process Clause, nineteenth century courts thought, was their flexibility to regulate the weight and effect of evidence in the service of “good sense, equity and convenience.”

A. The Law of Evidence on the Eve of Ratification

On the eve of the Bill of Rights’ ratification, formal rules of evidence were so few that Edmund Burke could declare that the law of evidence was “comprised in so small a compass that a parrot he had known might get them by rote in one half hour, and repeat them in five minutes.”

As a result, it’s easy to think the eighteenth century system of evidence placed few restraints on parties’ opportunities to introduce evidence. Initial appearances, though, are deceptive. Driven by a concern for controlling the costs of litigation and the risks of juror error, eighteenth-century evidence law had a deeply

116 See infra note 245 and accompanying text.
restrictive ethos, limiting jury’s ability to consider probative evidence in a host of cases.118

First, the eighteenth-century common law of evidence arrayed types of probative evidence in a strict hierarchy, forbidding jury reliance on relatively less probative forms of evidence when more probative forms were available.119 And at the top of that hierarchy was written proof. In line with a “centuries-long proclivity for suppressing resort to oral evidence at jury trial in civil matters,” courts barred the jury’s consideration of otherwise probative testimonial evidence when better evidence—written evidence—was available.120

The preference for written over oral evidence “protect[ed] against the shortcomings of jury trial,” which “[d]espite its merits, . . . has always been fraught with . . . the risks of error and partiality . . . .”121 The common law of evidence cabined that risk by narrowing the range of cases that would turn on a jury’s fallible assessment of witness credibility. “Testimonial credibility,” notes Stephan Landsman, “was only to be scrutinized . . . when there was no other guide,” because “definitive [written] evidence was lacking.”122

The common law “powerfully reinforced the . . . preference for written evidence” by limiting the testimonial evidence that could be admitted to the jury.123 True, the common law took a more lenient approach to some testimony we exclude today, like hearsay, which was admitted with reduced credit when definitive written evidence wasn’t available.124 But the common law also barred testimony central to most cases today: the testimony of interested parties, including the parties to the suit. In practice, this rule “greatly narrowed the range of potential witness testimony,” severely curtailing the role for the jury assessments of credibility.125 In transactional settings, for example, testimony of the parties about their knowledge of the terms of the transaction was not admissible if the meaning of the terms fell into contention, creating further incentive to channel agreements into writing.126 And in the as yet-undeveloped realm of tort, the rule also

118 See Stephan Landsman, From Gilbert to Bentham: The Reconceptualization of Evidence Theory, 36 WAYNE L. REV. 1149, 1154 (1990) (explaining the common law of evidence ensured “a significant number of questions could be resolved by exclusion rather than examination”).
119 See id. at 1153.
120 Landsman, supra note 118, at 1153–54 (explaining the common law’s best evidence rule “organize[d] the different sorts of available evidence in a hierarchical fashion” and the “preeminent category in this hierarchy was written evidence”; the common law’s exclusions on oral evidence, in favor of written evidence, “extended the best evidence approach”); Langbein, supra note 117, at 1183.
121 Langbein, supra note 117, at 1194.
122 Landsman, supra note 118, at 1158.
123 Langbein, supra note 117, at 1185.
124 Id. at 1189–90.
125 See id. at 1185.
126 Id. (“Especially in a transactional setting, such as contract or conveyancing . . . the parties would not be allowed to testify about the transaction if it fell into contention.”).
“prevented the testimony of the victim and injurer, testimony that would often [be] indispensable to prove the case.” 127 Indeed, tort law would not develop into its modern form until lawmakers lifted the ban on party testimony in the middle of the nineteenth century.

Eighteenth-century law not only sharply limited the jury’s ability to consider otherwise probative oral witness testimony, when better written evidence was available, but also heavily regulated, and rationed, parties’ proof opportunities by policing parties’ burdens of proof and the inferences juries could draw from evidence presented.

This was accomplished through “presumptive evidence,” which could be “disputable” or “conclusive.” 128 The disputable kind is a familiar concept: the production of certain evidence shifts the burden of proof to the other party, as in the Teamsters framework in Dukes. “Conclusive evidence,” though, went farther: it gave rise to an irrebuttable presumption that the event at issue had occurred.129 The most famous example is the concept of prescriptive title, a lynchpin of medieval property law, which assumed that an “uninterrupted enjoyment of an incorporeal hereditament . . . beyond the memory of man . . . furnish[ed] a conclusive presumption of a prior grant . . . .”130

Although conclusive evidence played a significant role in eighteenth-century property, criminal, and contract law, 131 many modern lawyers will resist classifying conclusive presumptions as rules regulating opportunities to present proof. We divide the law into separate categories—the law of evidence, regulating fact-finding, on the one hand, and law defining “substantive rights,” on the other. And modern lawyers often treat conclusive presumptions not as rules of evidence, but as rules defining the content of “substantive rights.” The very concept of an “irrebuttable or conclusive presumption [is a] classic misnomer,” declares Weinstein’s evidence treatise: “[a] so-called ‘irrebuttable presumption’ is a rule of substantive law.” 132

Applied to the late eighteenth and early nineteenth century, though, our way of thinking is anachronistic. The divide between substantive and procedure, law and fact, is notoriously hazy, and lawyers of the period conceptualized the divide differently. In their view, it would be nonsensical to say that a rule specifying that proof of fact X creates a conclusive presumption that legal violation Y occurred is anything other than a rule regulating opportunities to present evidence. The presumption has not changed the definition of the parties’ substantive rights and obligations, which, lawyers of the period would say, still forbids Y. It simply reflects a judgment to bar the presentation of evidence rebutting the presumption that Y occurred.

---

127 Id. at 1179.
128 1 SIMON GREENLEAF, A TREATISE ON THE LAW OF EVIDENCE 21 (9th ed. 1858).
129 Id.
130 Id. at 23.
131 Id. at 21–43 (discussing the role presumptions played in each of these areas).
132 1-301 WEINSTEIN'S FEDERAL EVIDENCE § 301.02[1].
This way of conceptualizing presumptions—as evidentiary rules that regulate, or even dispense with, consideration of further proof, not as rules defining the content of substantive rights—is evident in early cases’ and commentators’ uniform classification of conclusive presumptions either as “rules of evidence” or (using the nineteenth century term of art for rules of practice and procedure) as rules relating to “remedies” rather than “rights.” Treatment of conclusive presumptions as a species of evidentiary rule was, indeed, characteristic of every major treatment of the law of evidence prior to, and in the decades following, the Constitution’s ratification. This includes Bracton’s in the medieval period, Coke’s in the seventeenth century, Porthier’s and Gilbert’s in the eighteenth, and...

---

133 See, e.g., Webb v. Den, 58 U.S. (17 How.) 576, 578 (1854) (holding that legislation codifying the doctrine of prescriptive title relates to “what should . . . be received in courts as legal evidence”); Ogden v. Saunders, 25 U.S. (12 Wheat.) 213, 349 (1827) (stating that rules prescribing both the admissibility and “effect of” evidence relate to “the remedy and mode of proceeding” in court); Hobbs v. Bibb, 2 Stew. 54, 61 (Ala. 1829) (judicially-adopted conclusive presumptions “determin[e] the facts”); Rhinehart v. Schuyler, 7 Ill. 473 (1845) (syllabus) (argument of counsel) (conclusive presumptions relate to the “ascertainment” and “investigation” of adjudicative facts); Allen v. Armstrong, 16 Iowa 508, 512–13 (1864) (conclusive presumptions are “rules of evidence” and relate to the “mode and manner” by which rights are enforced); Groesbeck v. Seeley, 13 Mich. 329 (1865) (statute giving conclusive effect to evidence regulates the scope of evidentiary “hearing”); Quinlon v. Rogers, 12 Mich. 168 (1863) (argument of counsel) (legislation specifying that evidence should have conclusive effect enacts a “new rule of evidence”); Paschal v. Perez, 7 Tex. 348, 361–62 (1851) (rules governing the presumptive effect of evidence are part of the “rules of evidence,” which relate to the “remedy”); see also THOMAS M. COOLEY, A TREATISE ON THE CONSTITUTIONAL LIMITATIONS WHICH REST UPON THE LEGISLATIVE POWER OF THE STATES OF THE AMERICAN UNION 367 (1868) (discussing irrebuttable presumptions as part of his treatment of due process implications for the “rules of evidence”); id. at 367 (classifying evidentiary rules, including irrebuttable presumptions, as rules related to “remedies,” which “are not regarded . . . as being of the essence of a right”); LYSANDER SPOONER, AN ESSAY ON THE TRIAL BY JURY 10 (1852) (conclusive presumptions, which Spooner calls “conclusive proof,” are rules of the “law of evidence” that “regulate[[]] presentation of proof by “shut[ting] out” further evidence.

The first major, influential American critique of the traditional association between presumptions and rules of “evidence” came in a famous 1889 essay by James Thayer; and the modern concept of the law of evidence, as something that excludes conclusive presumptions, is part of a conceptual world that Thayer, in effect, helped to create. See James B. Thayer, Presumptions and the Law of Evidence, 3 HARV. L. REV. 141 (1889).


135 Id. at 207–08.

136 Id. at 208–09 (Coke’s brief discussion of presumptive evidence was, for a century, “the most authoritative statements of presumption and circumstantial evidence in English law texts”).
and Simon Greenleaf’s in the mid-nineteenth century. Each, to varying degrees, conceived of the law of evidence as a system for regulating opportunities for proof through a series of presumptions, each treated conclusive presumptions as an evidentiary judgment to, as Greenleaf put it, “dispense” with “corroborating evidence” and forbid “opposing evidence.” Indeed, as Stephan Landsman notes, Baron Gilbert even thought the common law’s best evidence rule, which Gilbert thought to be the organizing idea behind the entire law of evidence, was based on a conclusive presumption. If a party cannot produce written evidence, and the other can, “the very not producing it” raised a conclusive presumption against the former party, dispensing with the need for considering other proof.

Presumptions were largely judge-made—and courts adopted conclusive presumptions based on an array of considerations. One was the probative strength of the evidence given conclusive effect. “Just as in natural science,” Greenleaf

137 Id. at 222–26 (discussing Porthier and Gilbert); see also 2 WILLIAM BLACKSTONE, COMMENTARIES 370 (Philadelphia 1893) (1768) (discussing the law of circumstantial evidence in terms of a series of evidentiary presumptions).
138 GREENLEAF, supra note 128, at 21–62 (treating presumptive evidence under the “nature and principles of evidence”).
139 See SHAPIRO, supra note 134, at 206–09, 220–41 (discussing how the major treatise writers in the seventeenth, eighteenth, and early nineteenth century borrowed from and refined the canonists and civilians’ use of presumptions in their treatment of circumstantial evidence); see also Landsman, supra note 118, at 1158–59 (explaining that presumptions informed the eighteenth-century rules governing not only inferences and proof, but admissibility. For example, rules against admitting documents with certain flaws were based on the idea that these documents were presumptively inauthentic).
140 See, e.g., GREENLEAF, supra note 128, at 22 (characterizing presumptions as decisions to “dispense[] with,” or “forbid[],” opposing or corroborating evidence).
141 Landsman, supra note 118, at 1153 (“Th[e] best evidence principle governed the whole of Gilbert’s work [on evidence].”).
142 Id. at 1158 (quoting G. GILBERT, THE LAW OF EVIDENCE 4 (1754)).
143 As a result, presumptions effectively transferred fact-finding to the judge. SHAPIRO, supra note 134, at 242; see also Anne Woolhandler & Michael G. Collins, The Article III Jury, 87 VA. L. REV. 587, 657–58 (2001) (noting that it “was the understood duty of the federal judiciary . . . to expand the realm of reason by expanding the realm of law, specifying its particular applications, and narrowing the more chaotic realm of fact”); THAYER, supra note 102, at 212 (common law courts used presumptions to “modify[] the jury’s action in dealing with questions of fact” in order to promote “consistency in administration” and as a “sharp and short way of bridling the jury”). Presumptions were a basis not only for granting a new trial or special verdict, when the jury found the facts inconsistently with a presumption; but for excluding evidence related to defenses foreclosed by a conclusive presumption. See, e.g., Thayer, supra note 133, at 161–63 (discussing new trials and special verdicts); GREENLEAF, supra note 128, at 35 (noting that no averment or evidence may be “received” to controvert a conclusive presumption); see also Thayer, supra note 133, at 146–47 (suggesting that courts’ reliance on presumptions as a basis for excluding evidence, coupled with courts’ statement of presumptions in “evidential” terms, lead to what Thayer insisted was a mistaken view of presumptions as rules of evidence, rather than as rules of “substantive law”).
saying, “universality of experience” leads to “the presumption of aquatic habits of an animal found with webbed feet,” the same “universality of experience” supports legal presumptions, like a “presumption of a malicious intent to kill, from the deliberate use of a deadly weapon.” In either case, “one fact being proved or ascertained, the other, its uniform concomitant, is . . . safely presumed.”\textsuperscript{144} The other considerations were predictability and judicial economy. Conclusive presumptions, he says, “consist [mainly] of those cases” in which the universal experience of correlation between facts was so powerful that it “render[ed] it expedient for the common good”—that is, “for the sake of greater certainty,” and the “promotion of peace and quiet in the community”—that “all corroborating evidence [be] dispensed with, and all opposing evidence . . . forbidden.”\textsuperscript{145} (Similar considerations, Greenleaf’s English contemporary W.M. Best noted, could also lead courts to convert irrebuttable presumptions into rebuttable presumptions—in light of “enlarged experience” about the probative value of presumptive evidence or a changing understanding of policy trade-offs implicated by a truncated evidentiary hearing.\textsuperscript{146})

In effect, conclusively “presumptive evidence,” as lawyers of the period conceptualized it, begins to look a bit like a primitive, brutish ancestor of the more evolved forms of evidence so central to, and so controversial in, many class cases: statistical and formulaic proof. Like these forms of proof, proof that gives rise to a conclusive presumption was thought to have such strong, empirically grounded predictive power that it could substitute for a more fact-intensive evaluation of the facts at issue, in the service of conserving on the unpredictability, cost, and inconvenience of fact-finding in a world of imperfect institutions and process scarcity.

The bottom line: like rules governing admissibility of evidence, conclusive presumptions were treated as \textit{procedural} rules defining the proper scope of an evidentiary hearing. The point bears special emphasis. When we turn to postratification evidence of the original understanding of the Due Process Clause, our modern habit of conceptualizing conclusive presumptions as rules of

\textsuperscript{144}Greenleaf, \textit{supra} note 128, at 21.

\textsuperscript{145}Id. at 21–22; see also 1 Thomas Starkie, \textit{A Practical Treatise of the Law of Evidence, and Digest of Proofs, in Civil and Criminal Proceedings}, at iii-iv (London, J. & W.T. Clarke 1833) (characterizing evidential presumptions of law as technical and “positive rules” that “exclude evidence” based on “utility and convenience”). This view was not so different from Gilbert’s a century earlier. Landsman, \textit{supra} note 118, at 1158 (explaining that Gilbert “advocated presumptions because they could be precisely fashioned and applied in a fixed and mechanical manner,” resolving “doubtful questions without resort to an [imprecise and unpredictable] in-court testing of credibility”).

\textsuperscript{146}W. M. Best, \textit{A Treatise on Presumptions of Law and Fact: With the Theory and Rules of Presumptive or Circumstantial Proof in Criminal Cases} 21–22 (Philadelphia, T. & J.W. Johnson 1845) (“Many presumptions, which, in earlier times, were deemed absolute and irrebuttable, have, by the opinion of . . . judges, acting on more enlarged experience, either been ranged among presumptiones \textit{juris tantum} [rebuttable presumption] or considered . . . to be made at the discretion of a jury.”).
substantive law can blind us to the import of some nineteenth century due process
cases. Because lawyers of the period conceptualized these presumptions as
procedural rules regulating the scope of an evidentiary hearing, their treatment in
the due process law of the period sheds light on early American lawyers’
understanding of procedural, not substantive, due process.\textsuperscript{147}

Our focus, in this Section, however, is on the preratification law of evidence.
And the preratification prevalence of conclusive presumptions, coupled with the
way they were conceptualized, underscores the thoroughgoing degree to which
eighteenth century lawyers thought that courts could ration parties’ opportunities to
present probative evidence. In the old common law, there was simply no absolute
right to present any particular quantum of evidence—everything depended on
judicial judgment, seasoned by experience and sensitive to the need for
predictability and conservation of judicial resources.\textsuperscript{148} And, in certain
circumstances, these considerations led courts to dispense with corroborating
evidence and forbid, not just restrict, the defendant’s power to submit any
probative rebuttal evidence.\textsuperscript{149}

\textbf{B. Post Ratification Judicial Practice}

The key question for us is the extent to which the Constitution altered this
understanding. Antebellum courts had ample opportunity to confront that question.
A tide of reform reshaped the American law of evidence between 1780 and 1850—
offering litigants disadvantaged by the change an opportunity to cry a
constitutional foul by invoking the protection of the Due Process Clause.

Lord Mansfield was, arguably, the godfather of this era of reform. Consistent
with his general “unwillingness ‘to be bound . . . by strict rules and precedents,’”\textsuperscript{150}
Mansfield had “played down the authority of precedents in his . . . decisions.”\textsuperscript{151}
The common law, including the law of evidence, he thought, “did not consist of
particular cases, but in general principles”—reflecting “reason, equity, and

\begin{footnotesize}
\begin{enumerate}
\item See also infra Part IV.B.2–3 (elaborating on this point). For discussion of
postratification cases involving conclusive presumptions, see infra notes 176–88, 217–23,
231 and accompanying text.
\item See supra notes 119–47 and accompanying text.
\item See supra notes 119–47 and accompanying text.
\item James Oldham, \textit{Judicial Activism in Eighteenth-Century English Common Law in
the Time of the Founders: The More Things Change, the More They Stay the Same}, 8
\textit{Green Bag 2D} 269, 269 (2005); Wood, supra note 20, at 800 (“Mansfield played down
the authority of precedents in his . . . decisions and instead emphasized reason, equity, and
convenience . . . .”); see also \textit{Christopher J.W. Allen, The Law of Evidence in
Victorian England} 16 (1997) (Mansfield thought “[t]he law did not consist of particular
cases, but in general principles that ran through the cases and governed the decision of
them”).
\item Wood, supra note 20, at 800.
\end{enumerate}
\end{footnotesize}
convenience”—that “ran through the cases and governed the decision of them.”

His views were shared by many late eighteenth-century lawyers, who believed the “principles of nature and reason—morality and logic—. . . were part of the common law and enabled it to operate as an instrument of reform.” It was a view stated forcefully in Blackstone’s Commentaries. Even though Blackstone had suggested common law courts must adhere to settled law, including “the formal part, or method of proceeding,” he celebrated the role that common law courts had played in reforming the adjective law, which in aggregate, he thought, had ensured “speedy and substantial justice, much better than could . . . be effected by any great fundamental alterations” by Parliament.

American courts took a page from Mansfield and Blackstone, the proponent of judicial discretion, not Blackstone the proponent of judicial stasis. State courts adopted new rules affecting admissibility, many of which expanded parties’ opportunities for proof. Federal courts happily applied them. Take *Hinde v. Vattier*. The defendant insisted that federal courts follow the “settled rule” that required the plaintiff to establish his chain of title by “produce[ing] . . . the [original] grant[s], or an official or sworn copy” of those grants. Ohio courts, however, relented in cases where those grants had gone missing, allowing recordings in Ohio’s “book of land laws . . . to be taken as sufficient evidence of the grant . . . .” And the federal trial court followed that rule, with the Supreme Court’s approval. It is “fully [in the power] of courts to depart from the settled law of evidence, said Justice Baldwin, so long as new rules are “reasonable” and “conducive to the convenience of suitors.”

*Hinde* reflects a small-bore relaxation in the rules governing admissibility; other changes—which evolved in response to the rise the adversarial system in the last two decades of the eighteenth century—were much more momentous. For example, as “a more aggressive attitude on the part of lawyers striving to keep

---

152 Id. at 800.
153 ALLEN, supra note 150, at 16.
154 Id. at 15.
155 BLACKSTONE, supra note 137, at 267–68. For an illuminating discussion of this passage, see CARRESE, supra note 98, at 163–67 (Blackstone’s remarks “advocating liberal reform by judges amply indicates a project to cloak British politics and law with the moderating power of the judiciary”). As Christopher Allen notes, Blackstone’s views were common among lawyers of the period. ALLEN, supra note 150, at 19 (“Lawyers in the seventeenth and eighteenth centuries had preferred common law to statute as an instrument of reform because it was thought that statutes . . . were unable to deal efficiently with the variety of situations giving rise to legal disputes.”).
156 See Wood, supra note 20, at 801 (“In the decades following the Revolution, many Americans . . . took th[e] heightened interpretive power of English common law jurisprudence and ran with it[,] . . . sacrificing precedents for the sake of principle.”).
158 Id. at 399 (syllabus).
159 Id. at 401.
160 Id.
problematic testimony from the ears of the jury . . . manifested itself first in the
criminal courts and spread from there into civil litigation” in the last quarter of the
eighteenth century,”161 courts softened the rigor of the best evidence rule,
admitting, in fits and starts, oral testimony to contravene written evidence, with
the expectation that vigorous cross-examination would discipline jurors.162

This is not to say that antebellum change was a one-way ratchet. Courts also
began to take a more exclusionary approach to more suspect oral testimony—
abandoning, for example, their once-permissive approach to the admissibility of
hearsay. Just twenty-five years before the American Revolution, John Langbein
finds that, in civil trials, there is “scant indication that anything resembling the . . .
hearsay rule was in force . . . .”163 Then, hearsay seems to have been admitted, he
says, with reduced “weight or credit,” not excluded entirely, as we do today;
judges policed against its misuse by telling juries they shouldn’t place much
weight on it.164 The hearsay rule familiar to us—a formal exclusionary rule—
began to take hold in civil trials in the 1780s165 as judges, in face of an increasingly
aggressive bar, lapsed into passivity, giving up their old power to comment on the
evidence.166 Rather than police against misuse of hearsay by commenting on the
evidence, they barred lawyers from introducing it.167 Courts in the first quarter of
the nineteenth century continued to work out and refine the system of exceptions
with which we are familiar today.168

Even as they adjusted the strictures of the best evidence rule and tamped
down on the admissibility of hearsay, American courts, like their prerevolutionary

(1999).

162 See Landsman, supra note 118, at 1160–75.

163 Langbein, supra note 117, at 1187; see also id. at 1188 (noting “episodic
indications of official disquiet about hearsay” in the mid-eighteenth century); Gallanis,
supra note 161, at 512 (in 1755, hearsay “was accepted almost without comment.”).

164 Langbein, supra note 117, at 1189–90; see also Gallanis, supra note 161, at 512
(discussing the eighteenth century trial judge’s “discretionary power” over “areas of
evidentiary practice governed today by strict rules,” including the law of hearsay).

165 Gallanis, supra note 161, at 551 (while very little development in the growth of
exclusionary rules occurred “between 1754 and 1780,” the 1780s “were a period of
considerable activity” and “by 1800 much of the modern approach to hearsay was already
in place”).

166 Langbein, supra note 117, at 1198.

167 See id. (suggesting that “[a]dversary procedure pressured the judge toward
passivity and broke up the older working relationship of judge and jury.”); see also ALLEN,
supra note 150, at 186 (stating that as the adversarial system became entrenched by the
middle of the nineteenth century, the common law of evidence, applied to testimonial
evidence, “became increasingly exclusionary”).

168 Gallanis, supra note 161, at 502 (“[t]he period crucial to [the hearsay] rules’ full
development and consistent application” spanned “late eighteenth and early nineteenth
centuries”); id. at 535 (adding that in “the first quarter of the nineteenth century,” courts
“reinforced and refined” the major exceptions to the hearsay rule, while adding a few new
ones—notably the exception for necessity).
English precursors, continued to employ evidentiary presumptions to regulate, and sometimes restrict, opportunities for proof. Sometimes, courts adopted new rebuttable presumptions. For example, in *The Luminary*,\(^{169}\) in which the United States sought to condemn a ship that the government claimed had travelled under fraudulent papers to evade certain duties, Justice Story—over a typically vigorous dissent by Justice Johnson—adopted a presumption to ease the government’s burden of proof. The United States had raised a reasonable suspicion the papers were fraudulent, he argued, and so was entitled to a presumption of fraud, shifting the burden of proof to the parties opposing condemnation, because they had better access to relevant documentation establishing the chain of conveyance.\(^{170}\)

Similarly, in *American Fur Company v. United States*,\(^{171}\) a case in which the government sought forfeiture of a trader’s goods for engaging in proscribed trade in “ardent spirits” with Native Americans, the Court affirmed the trial court’s decision to allow the government a rebuttable presumption that the defendant had acted with the proscribed intent so long as the spirits were found mixed with goods intended for sale.\(^{172}\) This obviously furthered the deterrent policies of the forfeiture statute. And later courts saw these decisions as pragmatic innovations that assigned greater effect to plaintiff’s evidence “than it [had] possessed at common law.”\(^{173}\)

In prize cases, courts sitting in admiralty employed a mixed form of evidentiary presumption that combined rebuttable and irrebuttable features. Ships and cargoes found in the enemy’s possession were presumed to be property of the enemy, a presumption that could be rebutted by a neutral party claiming ownership rights to the prize. But admiralty courts limited rebuttal evidence, as a default

---

\(^{169}\) 21 U.S. (8 Wheat.) 407 (1823).

\(^{170}\) *Id.* at 411.

\(^{171}\) 27 U.S. (2 Pet.) 358 (1829).

\(^{172}\) *Id.* at 367–68.

\(^{173}\) See, e.g., State v. Cunningham, 25 Conn. 195, 203 (1856). *Cunningham* attributed the evidentiary alteration to Congress, not courts. *Id.* (attributing the evidentiary alteration in each case to the “legislature[,]”). Yet, while both *The Luminary* and *American Fur Company* involved federal statutory violations—a violation of section 27 of a 1792 act concerning registration of ships (*The Luminary*) and a violation of an act regulating trade with Native Americans (*American Fur Company*)—the presumptions adopted in each case were not commanded by the statutes at issue, which were silent with respect to how the statutory violations at issue could be proven. See, e.g., An Act concerning the registering and recording of ships or vessels, ch. 1, § 27, 1 Stat. 287 (1792) (providing that the punishment for use of a “fraudulent[]” certificate of registration was forfeiture of the ship, without providing how fraud may be proven); An act to regulate trade and intercourse with Indian Tribes, ch. 13, § 21, 3 Stat. 682 (1802) (barring transportation of “spirrituous liquors” into Indian territory for the purpose of “vending or distributing” them, without specifying how that intent might be proven). The presumptions adopted in each case were, rather, a judicial gloss intended to further the statutes’ purpose. In effect, in these cases, early courts were acting much like modern courts in employment discrimination suits, which adopt evidentiary presumptions to further the goals of Title VII. Cf. *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 416–17 (1975) (federal courts have discretion to fashion rules governing award of back pay, which must be guided by the “large objectives” of Title VII).
matter, to documentary evidence found on the ship and the testimony of the captured crew. In effect, the absence of onboard documentary evidence favoring the claimant gave rise to a default conclusive presumption against the claimant, who it was said, had no absolute right to introduce “farther proof”—for example, probative documents located on other ships. The right to a more particularized evidentiary hearing was, instead, committed to judicial discretion.

Chief Justice Marshall put his imprimatur behind the continuing use of irrebuttable presumptions in *Hamilton v. Russell*, a once controversial but now forgotten decision, in which the Chief Justice addressed how litigants could prove violations of Virginia’s 1779 fraudulent conveyance act. The act voided debtors’ sales of their property to third parties when the sale was made with the intent to avoid the reach of creditors. The statute, like most states’ fraudulent conveyance statutes of the time, reproduced the text of the 1571 English fraudulent conveyance act, which didn’t specify how fraudulent intent might be proven. And, at the time of the Virginia statute’s enactment in 1779, English authorities hadn’t resolved that question.

The conveyance at issue in *Hamilton* certainly looked suspicious. The debtor, Thomas Hamilton, had sold his brother property sought by Hamilton’s creditors, but his brother conveniently allowed him to keep the property. The creditor, in turn, argued the Supreme Court should follow the interpretation in *Edwards v. Harben*, a 1788 King’s Bench decision construing the English fraudulent conveyance act (a decision that postdated the Virginia statute’s enactment by a decade). In it, Justice Buller had held that possession of property should be treated as “clear and conclusive evidence” of the seller’s fraudulent intent. Hamilton, in turn, urged the Supreme Court to disregard *Edwards* and give him an opportunity to prove the sale of the property to his brother was not fraudulent. “Fraud or no fraud,” he argued, is for the jury.

---

174 3 SIMON GREENLEAF, A TREATISE ON THE LAW OF EVIDENCE 458, 461–65 (1846).
175 *Id.*
176 5 U.S. (1 Cranch) 309 (1803).
177 Virginia’s statute read: “Every gift, grant, or conveyance of lands, tenements, hereditaments, goods or chattels, . . . contrived of malice, fraud, covin, collusion or guile, to the intent or purpose to delay, hinder, or defraud creditors of their just and lawful actions . . . or to defraud or deceive those who shall purchase the same lands, tenements, or hereditaments, or any rent, profit, or commodity out of them, shall be . . . deemed . . . clearly and utterly void.” An Act to Prevent Frauds and Perjuries, Rev. Bills of Va., Ch. XXV (1779).
178 The King’s Bench would not resolve the question until 1788. See infra note 180 and accompanying text.
179 *Hamilton*, 5 U.S. at 310.
181 *Hamilton*, 5 U.S. at 312 (transcript) (noting the King’s Bench had held that “[i]f the possession be inconsistent with the deed, it is clear and conclusive evidence of fraud.”).
182 *Id.* at 314 (transcript) (“Fraud or no fraud, is a point to be decided by the jury and not by the court. . . . The possession of the vendor is not . . . a fraud, but only a
The creditor’s position was controversial. North Carolina courts, three years earlier, had adopted the debtor’s position. We “cannot agree,” said Justice Taylor in *Vick v. Kegs*, 183 “with [Edwards] that the property going otherwise as to its possession than the deed points out is absolutely fraud”; it is only a “mark of fraud” to be considered by the fact-finder. 184 Other state courts would stake out similar positions, some in unusually heated opinions. The Alabama Supreme Court complained, for example, of the “evil” occasioned by deference to the high-toned English construction, which usurped the fact-finding “genius” of the jury. 185 And that construction warred, the court said, with the text of the fraudulent conveyance act, which, by hinging the validity of the conveyance of the conveyer’s “intent” suggested the need for a particularized factual inquiry into the conveyor’s state of mind.186

Even so, in *Hamilton*, Marshall sided with the creditor. The best construction, said Marshall, furthers the statute’s overarching purpose—preventing fraudulent conveyances.187 And those conveyances, he said, are more likely to be deterred if a sale unaccompanied by a transfer of possession is treated as conclusive evidence of fraud.188 The reasons are obvious. Opening the trial to a more fact-intensive

circumstance from which, connected with others, the jury may presume the fact of a fraudulent intent.”).

183 3 N.C. (Cam. & Nor.) 121 (1800).
184 Id. at 121–22.
185 Hobbs v. Bibb, 2 Stew. 54, 61 (Ala. 1829) (“How, then, can the Court . . . say that if the possession remains with the vendor, it is conclusive evidence of fraudulent intent, not to be controverted by any testimony to shew the fairness of the transaction? It is contrary to the genius of our government . . . that Courts should encroach on the peculiar privileges of the jury, in determining on the facts. . . . . Such assumptions of authority render the boasted trial by jury a mere farce. This is an evil that has, to some extent, crept into the judicial tribunals of our own country, from too close an adherence to English authority.”).
186 Id. at 60 (“On reading this statute it does seem that the unsophisticated mind would be much at a loss to imagine, by what possible artificial rule of construction invented by the ingenuity of man, a contract entered into with good faith, and for a fair and valuable consideration, could be brought within its proscriptive influence. He would at once say that the statute forbids no honest transaction, it only proscribes fraud. The intention of the parties to the contract is . . . clearly one of fact to be determined by the jury”); id. at 62 (“If sound policy requires that he who has once owned a chattel shall never have possession of it after he has sold it, by hire, loan or otherwise, it is a proper subject for the attention of another branch of the government. Courts should never undertake to distort the words of a statute, to effect any object, however salutary.”).
187 *Hamilton*, 5 U.S. at 318 (the best construction “best promote[s]” the “intent of the statute” by “most effectually prevent[ing]” fraudulent conveyances).
188 Id. (“[F]raudulent conveyances, which are made to secure to a debtor a beneficial interest while his property is protected from creditors, will be most effectually prevented by declaring that an absolute bill of sale is itself a fraud, unless possession ‘accompanies and follows the deed.’”).

Some might read the decision as an interpretation of the substantive elements of the cause of action. This, though, anachronistically reads modern concepts into the decision. Because contemporary sources treated presumptions as “evidentiary” judgments committed
investigation of the debtor’s intent would increase the cost of litigation for creditors, increase the risk of erroneous or biased decisions against creditors by juries, and therefore increase the incentives for debtors to attempt fraudulent conveyances to third parties.

The upshot: In the antebellum period, parties’ opportunities to present proof expanded along some dimensions. They constricted along others. And, in the interest of economy and public policy, American courts continued to recognize “presumptive evidence” as means of regulating, and sometimes limiting severely, parties’ opportunities to present evidence.189

For our purposes, the crucial point of consistency with the old common law was courts’ view of their own power over evidence. Lawyers of the period were conscious they were living through a dynamic period in the law of evidence,190 and they were equally conscious that judges were responsible for that dynamism. Early nineteenth-century treatise writers, for example, acknowledged that evidence law was a “judicial development,” and much of it was of recent vintage.191 Yet, federal and state courts gave no hint that their respective due process guarantees either limited, or compelled, courts’ departures from the common law baseline. The legal weight and effect given to evidence was said to be a matter for reasoned “judicial discretion,”192 in light of the “great principles” of reason, equity and convenience that had always informed the common law of evidence.193

to judicial discretion subject to the loosely guiding principles of the common law, and the oral argument in the case argued the case for a presumption on these terms (by characterizing discrepancy between possession and the deed as “conclusive evidence” of fraud, see infra note 183), Marshall’s opinion is better seen as an exercise of traditional judicial judgment in the field of evidence, albeit one guided by his understanding of the goals of the statute, rather than what we would characterize as a construction of the proper definition of the “substantive” right created by the statute.

189 Antebellum courts, said the English commentator W. B. Best, were “slow[er] to recognize presumptions as irrebuttable” than their seventeenth and eighteenth century precursors. BEST, supra note 146, at 22 (“On the whole, modern courts are . . . disposed rather to restrict, than to extend, their number.”).

190 See, e.g., THOMAS PEAKE, A COMPRENDIUM OF THE LAW OF EVIDENCE, at vii (2d ed. London 1804) (the “rules of evidence [concerning oral testimony] have been so much altered, and so much light . . . thrown on them by modern decisions, that, comparatively, little is to be collected from ancient books that is satisfactory on the subject”); see also ALLEN, supra note 150, at 25 (the proliferation of evidence treatises in the early nineteenth century reflected a “change in the conception of the case law.”).

191 See ALLEN, supra note 150, at 24 (nineteenth century commentators “acknowledged and approved of judicial development of evidence law” and recognized much of that law was a “modern development”); see, e.g., BEST, supra note 146, at 36 (“[O]ur system of judicial evidence is . . . spoken of as something altogether modern.”).

192 See United States v. Lyon, 25 F. Cas. 249 (D. Mich. 1840) (No. 15,651) (finding Congress is without power to limit “the exercise of judicial discretion” concerning the weight and effect of evidence).

193 Wayman v. Southard, 23 U.S. (10 Wheat.) 1, 46–47 (1825) (absent legislation, courts must exercise discretion to develop “modes of proceeding” based on “great
Indeed, to the extent courts and commentators acknowledged the Due Process Clause in the field of evidence, it was in the course of denying that the clause compelled courts to retain, add to, or take away traditional opportunities to present proof: There was, as Thomas Cooley summarized the state of the law at the close of the antebellum period, no “vested right” to be governed by any particular rules of evidence.

C. Legislatures and Evidence

Even if antebellum courts didn’t seem to think due process restrained their own common lawmaking in the field of evidence, they didn’t think the Due Process Clause was inert in that field. It just had bite on a different branch—the legislature.

principles” of the common law); see also Odgen v. Saunders, 25 U.S. (12 Wheat.) 213, 349 (1827) (Marshall, C.J., concurring) (Congress may regulate “the remedy and mode of proceeding” in federal courts, including by “prescribing the evidence [that] shall be received . . . and the effect of that evidence”). For other statements equating the common law with principles of reason or common sense, see Hinde v. Vattier, 30 U.S. (5 Pet.) 398, 401 (1831) (noting state courts’ power to adopt new rules of evidence consistent with “reason[[]” and the “convenience of suitors”); Greenleaf, supra note 128, at 18–19 (principles of presumptive evidence reflect the “first principles of justice; or the laws of nature; or the experienced course of human conduct and affairs”); William David Evans, Appendix to 2 M. Pothier, A Treatise on the Law of Obligations 142 (W. Evans trans., 1806) (advocating a pragmatic approach to the best evidence rule, and noting that the excellence of the law of evidence consists “in requiring as much certainty and regularity as is consistent with general convenience, and in admitting as much latitude to private convenience as is consistent with general certainty and regularity”); see also Allen, supra note 150, at 16 (for many eighteenth-century English lawyers, “[t]he law did not consist in particular cases, but in general principles that ran through the cases and governed the decision of them”).

194 Cooley, supra note 133, at 367 (“It appears also that a right to be governed by existing rules of evidence is not a vested right.”). The striking absence of any concern for due process in Hamilton suggests this was Chief Marshall’s view, as well. To be sure, Hamilton involved a Virginia statute, and the Virginia Constitution’s due process clause, at the time, guaranteed due process only in cases that threatened deprivations of liberty and therefore had no applicability to a civil case involving property rights. See Siegel, supra note 22, at 33 n.160. Even so, the litigants were proceeding in federal court and appealing to federal judicial power, and so one would expect someone would have invoked the Fifth Amendment if there was any reason to think the federal Due Process Clause might plausibly restrain federal courts’ recognition of such presumptions. The fact that no one ever broached the issue suggests that neither Marshall nor the parties in the case viewed the federal Due Process Clause to incorporate any such restraint. Compare Livingston v. Moore, 32 U.S. 469, 549–53 (1833) (noting the parties had raised numerous state and federal constitutional arguments against a state statute directing that an executive settlement of accounts established a lien on private property in favor the state, before engaging in a choice-of-constitutional-law analysis).
Due process, antebellum courts uniformly held, guarantees a judicial, not a legislative, hearing—that is, an opportunity for exercise of “judicial discretion,” and judicial judgment, with respect to what happened and what the law required given the facts of the case before depriving someone of her property. The due process right to a judicial hearing, in turn, was thought to dictate certain features of judicial proceedings. A judicial hearing required the presence of adverse parties—an actor and reus—brought before the court through service of process. Ex parte proceedings, accordingly, were not properly “judicial.” A judicial hearing also required a judex, or an impartial judge excising duly constituted judicial power, who “hears before [he] condemns.”

But the right to a judicial hearing had its greatest bite as a qualified limit on the legislative power. Legislatures were not courts. And they could not usurp the judicial function by declaring private rights in an individual case through a legislative judgment. The point appears early. In an illustrative case, Demises

195 See, e.g., Lyon, 26 F. Cas. at 1037 (“I would say that congress have not the power, by an act of legislation, to take away the exercise of that discretion by a court [regarding how to decide a case, and the sufficiency of evidence] which is essential to the attainment of justice.”); Rhinehart v. Schuyler, 7 Ill. (2 Gilm.) 473, 482, 485 (1845) (argument of counsel that due process requires the judicial exercise of judicial judgment or “conformation and sanction” of the evidence, after a “judicial investigation”); Taylor v. Porter, 4 Hill 140, 146 (N.Y. 1843) (due process requires that it be “ascertained judicially that he has forfeited his privileges, or that some one [sic] else has a superior title to the property he possesses, before either of them can be taken from him. It cannot be done by mere legislation.”). For a discussion of other cases, see infra notes 215–31 and accompanying text.

196 See, e.g., Nelson, supra note 195, at 574 (observing that in the state and federal systems “form generally required the presence (actual or constructive) of adverse parties” and that “[p]roceedings that did not satisfy these minimal requirements were often said to be . . . not properly ‘judicial’”).

197 3 BLACKSTONE, supra note 137, at 24.

198 United States v. Klein, 80 U.S. (9 Wall.) 128, 147 (1871) (the power to “give the effect to evidence which, in its own judgment, such evidence should have” is an essential attribute of the “judicial” power); Trs. of Dartmouth Coll. v. Woodward, 17 U.S. (4 Wheat.) 518, 581 (1819) (argument of Daniel Webster).

199 See 2 JAMES KENT, COMMENTARIES ON AMERICAN LAW 13 (14th ed. 1896) (due process means law “in its regular course of administration, through courts of justice”); THOMAS MCINTYRE COOLEY, THE GENERAL PRINCIPLES OF CONSTITUTIONAL LAW 319 (1880) (“The legislature makes the laws, but cannot pass judgments or decrees, or make a law that is such in substance.”); THEODORE SEDGWICK, A TREATISE ON THE RULES WHICH GOVERN THE INTERPRETATION AND APPLICATION OF STATUTORY AND CONSTITUTIONAL LAW 676–77 (1857) (“by the right to the law of the land is meant the right to judicial procedure, investigation, and determination, whenever life, liberty, or property is
of Robinson v. Barfield (1818), a woman and her husband had executed a deed conveying their property. But a North Carolina statute dictated that the conveyance of a married woman is valid only if she is examined in a private examination by a judge, to ensure her part in the conveyance was voluntary. Yet, the woman in Barfield had died before she could be subject to an in camera judicial examination. The grantee and her husband successfully petitioned the North Carolina legislature to fix this problem by passing a special bill declaring the deed valid and quieting title in the grantee.

The woman’s heirs sued, arguing that the special bill was an impermissible legislative decree. The North Carolina Supreme Court agreed. “The Legislature,” wrote Judge Daniels in the North Carolina Supreme Court’s decision upholding their challenge, “has not the power to take the lands of A. and give them to B.” “The transfer of property from one individual, who is the owner [under generally applicable legal principles], to another individual, is a judicial and not a legislative act.” As the New York Supreme Court would later put the same point in a different case, the transfer of property requires a proceeding in which it is “ascertained judicially that [a party] has forfeited his privileges. . . . [i]t cannot be done by mere legislation.”


200 See, e.g., Bayard v. Singleton, 1 N.C. (Mart.) 5, 7 (1787) (the legislature cannot require a person “to stand condemned in his property without a trial”).

201 6 N.C. (2 Mur.) 391 (1818).


204 Barfield, 6 N.C. at 421.

205 Id. at 420.

206 Taylor v. Porter, 4 Hill 140, 146 (1843). For a sample of similar statements before and after Barfield, see Wynehamer v. People, 13 N.Y. 378, 433 (1856) (“[B]oth courts and commentators in this country have held that [the due process clause] . . . secure[s] to every citizen a judicial trial, before he can be deprived of life, liberty or property.”); Univ. of N.C. v. Foy, 5 N.C. (1 Hayw.) 58, 87–89 (1805) (due process protects vested rights to property until “the judiciary of the country in the usual and common form pronounce them guilty of such acts as will . . . amount to a forfeiture”); Norman v. Heist, 5 Watts & Serg. 171, 173 (Pa. 1843) (due process requires deprivation “by the judgment of his peers,” not “an ex post facto . . . [legislative] decree made for the occasion”); State v. Heyward, 37 S.C.L. (3 Rich.) 389, 410–12 (1832) (“a corporation can only be deprived of its powers, rights, privileges, and immunities, by a judgment of forfeiture, obtained according to the law of the land,” meaning “a trial had, and a judgment pronounced, in the court of law of this State”).
As the antebellum period wore on, courts expanded on this idea, relying on the Due Process Clause and its state analogues to bar not only special bills declaring special rights in individual cases, but, more broadly, retroactive alteration of laws in a way that provided that “past acts meeting the new requirements had already effected a transfer of property.” The rule against retroactive divestment of “vested rights” in effect acted as a prophylactic rule that prevented legislatures from retroactively manipulating generally applicable substantive standards in a way that would foreseeably change the outcome in identifiable disputes.

Initially, however, some courts refused to recognize due process restrictions on legislative power to regulate remedies, including evidence. While legislatures could not enter legislative decrees by dictating the results in individual cases, legislatures had, said Justice Marshall, the “acknowledged power” to regulate “the remedy and mode of proceeding” in federal courts, including by “prescribing the evidence which shall be received . . . and the effect of that evidence.” Congress and state legislatures “may,” as one court put it, “so change the rules of evidence as to make it difficult to establish our rights.” Indeed, legislatures could

---

207 Mendelson, supra note 199, at 126–27 (the antiretroactivity principle was an organizing principle of pre-Lochner due process cases, rooted in a separation of powers-influenced understanding of the right to a hearing); Woolhandler, supra note 199, at 1018–25 (same); Siegel, supra note 21, at 56 (between 1818 and 1829, “most, if not all, courts used the concept of [the distinction between judicial and] legislative power to build a textually based constitutional prohibition of retrospective laws”); Kainen, supra note 199, at 88–89 (“the principle of non-retroactivity” was the primary organizing principle in the Court’s pre-Lochner due process cases).

208 See, e.g., Mendelson, supra note 199, at 127 (the Court’s vested-rights retroactivity cases were aimed at preventing legislatures from affecting the outcome in “particular cases”); see also Gordon S. Wood, The Origins of Vested Rights in the Early Republic, 85 VA. L. REV. 1421, 1439 (1999) (the vested rights doctrine reflected a concern with preventing state legislatures from “meddling” in “disputes between contending parties”); Woolhandler, supra note 199, at 1025 (the vested rights doctrine was wrapped up in separation of powers concerns that retroactive legislation usurped the judicial role over cases and controversies). Courts of the period also thought retroactive confiscation of property was a form of “punishment,” which could only be imposed after a judicial hearing. EDWARD S. CORWIN, LIBERTY AGAINST GOVERNMENT: THE RISE, FLOWERING AND DECLINE OF A FAMOUS JUDICIAL CONCEPT 93 (1948).

209 Ogden v. Saunders, 25 U.S. (12 Wheat.) 213, 349 (1827); see also Gibbs v. Gale, 7 Md. 76, 87 (1854) (the legislature has the power “to abolish or modify [a rule of evidence] in every case, or to modify and annul it under particular circumstances”).

210 Edwards v. Pope, 4 Ill. (3 Scam.) 465, 468 (1842); Oriental Bank v. Freeze, 18 Me. 109, 112 (1841) (“When a person, by the existing laws, becomes entitled to recover a judgment . . . he is apt to regard the privilege, which the law affords him, as a vested right, not considering that it has its foundation only in the remedy, which may be changed, and the privilege thereby destroyed.”).
retroactively impose new rules of evidence even in pending cases.211 “It is not our province,” one judge put it bluntly, “to determine the expediency of such legislation [dealing with adjective law]; that is a matter confided to another and independent branch.”212

The latitude was often explained by the notion that remedies were separate from rights. While legislatures could not usurp a “judicial” disposition of rights by retroactively declaring that A has a right to B’s property, changes in evidence were said to affect the “remedy,” which was distinct from the “rights” in controversy.213 Changes in the law of evidence, even retrospective ones, therefore did not impinge on a “judicial” disposition of private rights.214

Even so, this deference principle was in tension with the principle that legislatures could not deprive people of property through “legislative decrees.” After all, case-specific legislative changes in the rules of evidence could certainly dictate the outcome of cases. Indeed, it was hard to distinguish the special bill found unconstitutional in Barfield from a statute proposing case-specific rules of evidence, as counsel for the grantees defending the bill noted.

True, the North Carolina bill had transferred property rights from A to B, by validating the deed and overruling the operation of the ordinary law. But, because the bill included legislative findings that the wife’s conveyance had been attested by two witnesses out of court and these witnesses supported the conveyance’s validity, it might also have been characterized as a declaration of special, “for this case only” evidentiary rule. “Cannot the legislature,” argued counsel for the grantees,

say, that other proofs than those required by existing statutes, shall be good as to the execution of deeds by feme-coverts? If so, could not the Legislature of 1788, say, that the acknowledgment of Mrs. Brown to the two witnesses to the deed, ‘of her having executed it freely and voluntarily,’ should be deemed good evidence of the execution of the deed? And having said so, shall not Courts of Justice be bound by it? . . . The Legislature cannot exercise judicial powers, properly speaking; but they can . . . establish the rules of evidence in all cases whatsoever.215

211 Webb v. Den, 58 U.S. (17 How.) 576 (1854); see also Bartlett v. Lang, 2 Ala. 401, 406 (1841) (“A statute which merely changes the remedy, it has been often holden . . . may, where such seems to have been the intention of the legislature, operate retrospectively.”).

212 Gibbs, 7 Md. at 88.

213 See Sturges v. Crowninshield, 17 U.S. (4 Wheat.) 122, 202 (1819) (“The distinction between the obligation of a contract, and the remedy given by the legislature . . . exists in the nature of things. Without impairing the obligation of the contract, the remedy may certainly be modified as the wisdom of the nation shall direct”).

214 See COOLEY, supra note 133, at 367 (rules of evidence “pertain to the remedies which the State gives to its citizens, and are not regarded as . . . being of the essence of a right” and “are therefore at all times subject to modification and control by the legislature”).

215 Barfield, 6 N.C. at 401 (syllabus).
Courts eventually resolved the tension by imposing some limits—with fuzzy outer edges—on legislative interference with judicial judgments about the inferences juries could draw from evidence. The first move came in the bluntest sort of case. Legislatures, rather than explicitly decreeing that A receive the property of B, as in Barfield, declared the existence of facts that would entitle A to relief from B. This, in turn, invited courts to treat the legislative findings as presumptive evidence of the truth of those facts in an ensuing lawsuit. As the Court of Appeals of Kentucky said, in the course of rejecting this type of gambit: “Once adopt the principle that such facts are conclusive or even prima facie evidence against private rights, and many individual controversies may be . . . drawn from the functions of the judiciary into the vortex of legislative usurpation.”

It was a short step from there to reject not only legislative efforts to “declare” facts in particular cases, but also legislation mandating that courts give conclusive weight to certain classes of evidence. The problem was presented in a series of cases challenging federal and state statutes making various official documents created by executive officials “conclusive evidence” of rights to property. For example, statutes treating tax deeds prepared by a public auditor and given to buyers of property seized to satisfy a tax delinquency as “conclusive evidence” of valid title or of the property’s value. The obvious problem with these statutes was that the legislature had substituted the judgment of executive officials for judicial judgment about adjudicative facts. In doing so, the legislature allowed executive officials to dictate the outcome of individual cases.

As counsel put it in Rhinehart v. Schuyler, an illustrative Illinois case, “due process of law” requires “judicial process” and “judicial judgment” prior to the deprivation of property—which means that an opportunity has to be afforded the owner of the seized property to appear and “be heard in his defense”—that is, to appeal to judicial judgment, after a “judicial investigation” of the facts. “[T]he exercise of a power of inquiry into, and a determination of facts between debtor and creditor . . . conclusive upon the rights of all persons affected by it . . . makes it judicial,” not an executive or legislative power. Accordingly, that power must remain with the judiciary. Decisions by public auditors therefore cannot be made conclusive—they can receive their effect only “from judicial confirmation and

216 Elmondorff v. Carmichael, 13 Ky. (3 Litt.) 472, 480 (1823); see also Parmelee v. Thomson, 7 Hill 77, 80 (N.Y. 1845) (citing Elmondorff with approval and noting “the legislature has no jurisdiction to determine facts touching the rights of individuals.”).
217 See, e.g., Wantlan v. White, 19 Ind. 470, 472 (1862) (statute making evidence conclusive is void because a declaration “would seem to be a judicial act”); Groesbeck v. Seeley, 13 Mich. 329, 332 (1865) (making auditors’ deeds conclusive evidence of title unconstitutionally deprived the claimant of a judicial “hearing”); see also Cooley, supra note 133, at 367–69 (discussing cases).
218 7 Ill. (2 Gilm.) 473 (1845).
219 Id. at 485 (syllabus).
220 Id.
221 Id.
sanction.”

If it were otherwise, “[t]he judiciary would have nothing to do but to register legislative edicts.”

In the two decades bookending the Civil War, the Supreme Court extracted from these cases some general principles with fuzzy edges. With respect to deprivation of private rights, it declared in 

* Murray’s Lessee v. Hoboken Land and Improvement Company,*

and 

* Hurtado v. California,*
due process requires a “judicial” hearing. A hearing, to be properly “judicial,” the Court suggested in 

* Hurtado* and 

* United States v. Klein,* requires that courts exercise their own “deliberation and judgment” about the “effect of . . . evidence” concerning adjudicative facts in individual cases and controversies.

In *Hurtado,* the Supreme Court summed up the basic points of agreement in this century-long train of decisions. Legislatures cannot declare special law, or special remedial rules, for “particular case[s]” through “legislative judgments and decrees,” bills of attainder, “acts reversing judgments,” or “acts directly

---

222 *Id.*

223 *Id.* The court in *Rhinehart* dealt with the problem by finding that the auditor’s deed created a rebuttable presumption that the tax sale was lawful, which the former owner could rebut in an ejectment action. *Id.* at 528. The Supreme Court, a decade later, “acknowledged the traditional framework,” see Nelson, *supra* note 195, at 586, and solved the due process problem by holding that framework applied in cases involving a “private right,” but not “public rights,” such as proceedings involving the collection of tax revenue. *Den ex dem Murray v. Hoboken Land & Improvement Co.,* 59 U.S. (18 How.) 272, 284 (1855). As Caleb Nelson notes, this “deviation from the traditional framework” proved to be “quite limited.” Nelson, *supra* note 195, at 585.

224 59 U.S. (18 How.) 272 (1855).

225 110 U.S. 516 (1883).

226 *Murray,* 59 U.S. at 280 (“[D]ue process of law’ generally . . . [requires] a trial according to some settled course of judicial proceedings.”); *Hurtado,* 110 U.S. at 533 (suggesting due process is generally satisfied by a “ trial . . . according to the settled course of judicial proceedings,” in which “provision has been made . . . for giving [the party] an opportunity to be heard in his defence [sic]; for the deliberation and judgment of the court”).

227 80 U.S. (13 Wall.) 128 (1871).

228 *Hurtado,* 110 U.S. at 533, 535 (due process requires a “judicial” trial, which “hears before it condemns which proceeds upon inquiry,” and provides room for judicial “deliberation and judgment”); *Klein,* 80 U.S. at 147 (the power to “give the effect to evidence which, in its own judgment, such evidence should have” is an essential attribute of the “judicial” power). *Klein,* in particular, built on the antebellum state decisions invalidating statutes legislating conclusive presumptions: It struck down a postwar federal statute that instructed courts to treat a presidential pardon as “conclusive evidence” that the recipient had aided the Confederacy, emphasizing, in the process, that decisions about the “effect of . . . evidence” were an attribute of the judicial power, which must be left to the judiciary. *Id.* For further discussion, see also Nelson, *supra* note 195, at 590–91 (according to the nineteenth century understanding of due process and separation of powers “[w]hen core private rights were at stake, courts had to be able to exercise their own judgment . . . [about] relevant details of their individual interaction”; as a result “the legislature could not conclusively determine what have come to be called ‘adjudicative facts’”).
transferring one man’s estate to another.”229 And the Court suggested a general principle that flowed from, and explained, these points of agreement. Congress must define rights and regulate remedies in a way that leaves courts room to “hear” before “condemning”230—that is (as the Court had emphasized in Klein) to independently assess the weight and “effect of evidence” in individual cases and controversies.231

D. Due Process and Judicial Independence

In the end, viewed against the pattern of early history of construction of the Due Process Clause, the idea pushed by defense counsel in Dukes v. Wal-Mart Stores, Inc., that due process guarantees an individual right to a deeply contextual, especially fact-intensive evaluation of the underlying claim is anachronistic. Due process did guarantee civil defendants a right to a right to be “heard” before a

229 Hurtado, 110 U.S. at 536.
230 Id. (while each state “prescribes its own mode of judicial proceeding,” due process forbids states from adopting “special rule[s] for a particular person or a particular case”; rather they must regulate rights and remedies in a “general way,” conducive to the “public good,” ensuring space for courts to “hear” before “condemning”).
231 Klein, 80 U.S. at 146–47 (insisting Congress must leave decisions about the “effect of . . . evidence” to the judiciary). This principle, applied to conclusive presumptions, had fuzzy edges. Rhinehart and its federal companion, Klein, never specified whether the problem with legislatively enacted conclusive presumptions was a general one or was confined to the facts of either case.

Portions of each decision suggest the courts viewed the problem as a general one, stemming from the presumptions’ thoroughgoing intrusion on the judicial power to investigate and weigh the evidence. See, e.g., id. at 147 (stating flatly that legislation giving a “conclusive” effect to evidence categorically infringes on the “judicial” power). However, both courts may have been particularly troubled for a different, narrower reason: the presumption’s operation in each case effectively allowed the political branches to dictate the effect of individual cases in advance, making an end-run around the due process-derived rule forbidding executive or legislative “decrees.”

In Rhinehart, for example, the presumption gave an executive agent, a public auditor, the final word about the validity of particular tax seizures. In Klein, the Court considered the constitutionality of a legislatively mandated conclusive presumption that persons who had received a post-Civil War presidential pardon were guilty of aiding the Confederacy and therefore not entitled to the return of or compensation for seizure of their property. Indeed, in Klein, the statute was enacted after the respondent, Klein (the administrator of the estate of a pardon recipient), had prevailed against the United States in the Court of Claims on just such a property-recovery claim; the statute dictated the outcome of the United States’ appeal in the United States’ favor. See, e.g., DAVID P. CURRIE, THE CONSTITUTION IN THE SUPREME COURT: THE FIRST HUNDRED YEARS 311 n.173 (1985) (noting one basis for Klein, although in his view not the best, was that the “United States was attempting to decide its own controversy” by “set[ting] aside a decision already rendered”). In either case, the presumption not only deprived courts of an opportunity to make an independent judgment about the effect of evidence, but made the political branches the ultimate arbiter of their own controversies. Id. at 311 n.173.
judge “in defense.” But that right was not conceived as a personal right to a maximally fact-intensive evidentiary hearing. It was a structural right, which selected the institution that should make the final judgment in individual case and controversies: a deliberative judiciary, acting free from legislative influence or control.

In this, the Due Process Clause was construed in a way that paralleled the focus of other provisions in the Bill of Rights, which, Akhil Amar has shown, were “more structural than not.” Amar also shows that the Bill of Rights was in many respects more majoritarian than countermajoritarian. But the antebellum construction of the Fifth Amendment’s Due Process Clause was countermajoritarian—it envisioned the judiciary as a buffer between individuals and the popular elected branches.

This construction was rooted in antecedents of the Due Process Clause, like Magna Carta. But an account that explains the antebellum construction solely by reference to the clause’s formal antecedents obscures the deeper reasons for its appeal to antebellum lawyers. That appeal was rooted in a larger vision of the judiciary’s role in a system of separated powers with wide currency in the legal profession during and after the framing period.

John Adams and Alexander Hamilton were the framers most associated with this theory. “[L]ittle that Adams ever wrote,” David McCullough notes, “ever had such effect as his Thoughts on Government,” his tract that made the plea for an independent judiciary. In it, Adams made the usual argument for judicial independence: it was essential to judicial review of the constitutionality of legislative and executive actions. But, echoing Blackstone, he also stressed that an independent judiciary “experience[d] in the laws, of exemplary morals, invincible patience, unruffled calmness and indefatigable application” would ensure political “stability” and “upright and skillful administration of justice.”

In the Federalist Papers, Hamilton elaborated it is “not with a view to infractions of the constitution only that the independence of the judges may be an

---

232 See, e.g., Hurtado, 110 U.S. at 533.
234 Id. at xiii. 137–294.
237 Id. at 207. For a discussion of Blackstone’s views, see supra note 155 and accompanying text.
essential safeguard,” he wrote in Federalist 78.\textsuperscript{239} The “firmness of the judicial magistracy,” he said, “is [also] of vast importance in mitigating the severity and confining” the effect of “unjust and partial laws” enacted under the influence of popular passions.\textsuperscript{240} At the same time, an independent judiciary is best fit to exercise the “discretion” to construe the law in line with “reason” and “propriety” and to conduct the “administration of justice” toward the end of “utility.”\textsuperscript{241}

The judiciary, in effect, would guard against popular whims and legislative “ill humor” by “confining” and “modera[ting]” unjust laws.\textsuperscript{242} And, in line with Blackstone’s account of the English judiciary, it would also be a mildly progressive influence: The judiciary would take the lead in improving the administration of justice through its independent power to administer the judicial system.\textsuperscript{243} It would mold the foundation on which legislatures might build.

This benign view of the judiciary wasn’t universally shared among the framers. But it was Adams and Hamilton who won the day, as Jefferson, who came to oppose their vision of a quietly dynamic judiciary, conceded late in his life: “[A]ll the young brood of lawyers,” Jefferson complained to Madison in 1826, subscribe to Blackstone’s “honied Mansfieldism”\textsuperscript{244}—that is, the view associated with Adams and Hamilton in the framing era colonies, but with Mansfield and Blackstone in England, favoring a quietly innovating judiciary that exercises “discretion” to improve law “in accord with equity, good sense, and convenience.”\textsuperscript{245}

\begin{thebibliography}{99}
\bibitem{239} The Federalist No. 78 (Alexander Hamilton), reprinted in 1 The Papers of Alexander Hamilton 662 (Harold C. Syrett & Jacob E. Cooke eds., 1962) (1788).
\bibitem{240} Id.
\bibitem{241} Id. at 439, 442. For an insightful discussion of Hamilton’s defense of an independent judiciary, see Carrese, supra note 98, at 197–207. Adams and Hamilton echoed the establishment view of the preratification English legal profession, and anticipated the view of the American legal profession postratification. See Allen, supra note 150, at 19 (“Lawyers in the seventeenth and eighteenth centuries had preferred common law to statute as an instrument of reform because it had been thought that statutes . . . were unable to deal efficiently as with the variety of situations giving rise to legal disputes.”); Wood, supra note 20, at 800–01 (American lawyers of the period thought that “judges in their multiplicity of piecemeal decisions could control and transform the law more rationally” than legislatures).
\bibitem{242} See Hamilton, supra note 239, at 440.
\bibitem{243} Alexander Hamilton, Farmer Refuted, reprinted in 1 The Papers of Alexander Hamilton, supra note 239, at 137 (“All Lawyers agree that the spirit and reason of a law, is one of the principle rules of interpretation . . . ”). For discussion of Blackstone’s influence on Hamilton, see Carrese, supra note 98, at 187–90 (at the founding, “Blackstone’s Commentaries was replacing Coke-Littleton as the basic law text, and Hamilton seems to have discerned the quiet emphasis that both English jurists placed on an effectively independent, moderating judiciary”).
\bibitem{245} See, e.g., Posting of Gordon S. Wood re: “Mansfieldism,” H-Net Discussion Network (Aug. 17, 2005), http://h-net.msu.edu/cgi-bin/logbrowse.pl?trx=vx&list=h-
The antebellum construction of due process reflects the degree to which this view of the judiciary’s role captured the imaginations of antebellum American lawyers. It also illustrates one way in which antebellum lawyers envisioned courts would take on this role. Some modern commentators suggest courts should boldly “improve” the law by updating statutes to reflect contemporary policy preferences. The antebellum due process cases, instead, emphasized courts’ power to renovate and mitigate the law’s severity in a more cautious, less nakedly legislative, way—through their power over evidence, including their power to limit opportunities for proof by adjusting parties’ evidentiary burdens or adopting legal presumptions when necessary to promote predictability, protect against biased fact-finding, and conserve on the costs of jury trial, much as Marshall did in *Hamilton v. Russell*.

This vision was countermajoritarian, but not aggressively so. The legislature had the acknowledged authority to regulate rules of evidence. It might build on—or even overrule—judicial innovations. But its authority needed tempering, lest legislatures rashly predetermine cases against unpopular parties by dictating the effect of evidence in individual cases and controversies.

By construing due process in a way that imposed most of its bite on legislatures’ exertions in the fields of evidence and adjective law, antebellum courts put this vision into effect. They asserted the power to incrementally improve the administration of justice by “minute,” sometimes fictional, “contrivances” in the fields of procedure and evidence. They acknowledged legislatures’ concurrent power to legislate in the field of evidence. Through the Due Process Clause, though, they tempered that power by barring dangerous legislative micromanagement of adjudicative fact-finding in individual cases. In the process,

---

246 See, e.g., *Eskridge, Dynamic Statutory Interpretation* (1994) (discussing the practice of dynamic statutory interpretation).

247 See supra notes 176–89 and accompanying text. The claim that due process presupposes judicial flexibility in the field of evidence is not that different from claims made by many originalists today (although, these claims have been made without the postratification historical support developed here); see, e.g., Rosenthal, *supra* note 85, at 44 (the “original meaning” of the Due Process Clause was “non-originalist,” meaning it delegated discretion to judges to adopt fair procedures). It also parallels Frank Easterbrook’s claims about statutory interpretation. He thinks statutes should be interpreted against a background presumption against disturbing judicially devised evidentiary and procedural rules. Frank Easterbrook, *The Case of the Speluncean Explorers: Revisited*, 112 Harv. L. Rev. 1876, 1913 (1999) (suggesting statutes must be interpreted against “the legal system’s [judicially] accepted procedures, evidentiary rules, [and] burdens of persuasion”). The claim here is that the same background assumption in favor of judicial discretion in this field informed the early understanding of the Due Process Clause.

248 *3 Blackstone, supra* note 137, at 268.
by requiring legislatures to legislate in a way that left room for judges to “hear” before “condemning,” and decide the “effect of evidence,” they also ensured themselves room to mitigate the effect of unwise or unjust laws through their power to define the weight and sufficiency of evidence necessary to prove violations of rights on a case-by-case basis.

But as long as Congress left some room for independent judicial judgment about the weight and sufficiency of evidence, the requirements of due process in the field of evidence gave out. Judicial regulation of parties’ opportunities to offer evidence was, instead, a subconstitutional matter: a question properly left to courts’ reasoned discretion, based entirely on judicious weighing of the principles of “good sense, equity and convenience” undergirding the common law of evidence.

IV. THE RISE AND FALL OF LOCHNERIAN PROCEDURAL DUE PROCESS

As the last Part showed, class action defendants’ argument—that due process guarantees civil defendants a fixed right to present all probative rebuttal evidence—finds little support in the early-nineteenth-century liquidation of the meaning of due process. Through most of the nineteenth century, due process was understood in structural terms: as a limit on legislative interference with courts’ traditional discretion to determine “the effect of evidence.” What courts did with that discretion, however, was a subconstitutional concern.

That doesn’t mean class defendants’ view of due process lacks historical support. As Section A explains below, for a three-decade period—the so-called Lochner era—the Court enforced a version of due process indistinguishable from class action defendants’. It recognized that due process guarantees civil defendants a right to fact-intensive evidentiary hearing prior to a civil liability judgment.

Even so, the Court’s approval of this theory was short-lived. It was rooted in the Lochner era’s solicitude for property and economic rights and collapsed shortly after the collapse of its economic rights jurisprudence in the late 1930s. Since then, as Section B shows, courts have reverted to a stance close to the old due process tradition that Lochner had supplanted—today’s courts, like their nineteenth-century precursors, recognize that, as a matter of due process, courts have broad discretion to regulate opportunities to present evidence in civil cases.

A. The Right to a Hearing in the Lochner Era

The idea that due process guaranteed civil defendants a right to present all probative evidence was driven by interrelated conceptual shifts, which fused together to support recognition of a right to a maximally fact-intensive evidentiary hearing. The first, most profound of these is also the most well known. In the closing decade of the nineteenth century, courts abandoned the old structural

249 See infra Part III.
conception of due process—concerned with protecting the right to an independent judicial hearing—and replaced with it substantive due process.

The sources of this shift have been exhaustively examined elsewhere, and don’t need to be reviewed at great length here beyond a brief summary. The first tremor came in *Mugler v. Kansas*, an 1889 decision upholding a Kansas statutory provision declaring breweries “common nuisance[s],” and authorizing the destruction of liquor found on their premises, as well as injunctions to “abate and perpetually enjoin” their operation. The petitioners had challenged the statute by employing a creative extension of antebellum due process analysis. They argued that the legislature had “by the mere exercise of its arbitrary caprice,” without “notice, trial, or hearing,” declared a brewery to be a common nuisance, and then “prescribe[d] the consequences which are to follow inevitably by judicial mandate,” effectively turning courts into “legislative agent[s].” The statute amounted to an extrajudicial “legislative decree”—depriving the defendant, in turn, of his due process right to an independent exercise of “judicial discretion or judgment” about whether breweries were common nuisances.

The Court rejected this argument. In the process it brushed aside the old structural way of thinking about due process on which it was premised, linking “the extent of protection accorded to [the petitioner’s] property with the scope of the police power, which it judged in a manner characteristic of *Lochner*.” In the succeeding decades, the old equation of due process with a right to an independent judicial hearing was demoted to a subordinate strand of the Court’s due process analysis, as the Court, starting in the late 1890s, focused its due process jurisprudence on protecting fundamental substantive rights, at the center of which

---

250 See, e.g., Bernstein, supra note 21; Siegel, supra note 21.
251 123 U.S. 623, 671 (1887).
252 Id. at 677.
253 Id. at 671; see also id. at 639 (argument of petitioners that the statute “without notice, trial or hearing . . . declared [the brewery] to be a common nuisance”).
254 Id. at 642 (argument of petitioners).
255 Id. at 672 (describing petitioners’ argument).
256 Id. (argument of petitioners) (the statute violated due process because it permits “the exercise of no judicial discretion or judgment”; rather “[t]he brewery being found in operation, the court is not to determine whether it is a common nuisance, but under the strict behest of the statute is to find it to be one”).
257 Id. at 264.
258 Kainen, supra note 199, at 133–41. But see Mendelson, supra note 199, at 135 (it is not until 1896 in *Missouri Pacific RR. v. Nebraska* that one finds “a Supreme Court majority invalidating legislation . . . without reliance upon some element of separation [of powers]”).
259 By the end of the *Lochner* era, the equation of due process with an evidentiary hearing in the judicial branch was jettisoned entirely. See, e.g., Crowell v. Benson, 285 U.S. 22, 56 (1931).
lay rights to property and freedom of contract—subjecting legislation affecting those rights, in turn, to heightened scrutiny.\textsuperscript{260}

The second shift is the collapse of the old right/remedy distinction. In the latter half of the nineteenth century, courts progressively abandoned the idea that remedies are separate from, and subordinate to, rights.\textsuperscript{261} Instead, lawyers came to recognize, as we do, that the value of a right depends on the remedies provided for protecting it.\textsuperscript{262} At some point, remedial frameworks may create such a large risk of error that the framework is tantamount to a deprivation, or “spoliation,” of the right itself.\textsuperscript{263}

Third, the conception of the value of a “hearing” changed. The old construction of due process had treated the right to a judicial hearing in classical separation of powers terms, with roots in Enlightenment political theory: as a right to a decision by the right institution, an independent judiciary. By the closing decades of the nineteenth century, courts began to explain the value of the judicial hearing in new ways, by emphasizing the truth-seeking value of robust adversarial presentation of evidence.\textsuperscript{264}

The change reflected the normalization of the adversarial system, which was slow in coming. Into the 1840s, lawyers and lay people on both sides of the Atlantic complained that the “extent to which it [had become] possible for counsel . . . to control proceedings” in civil cases had increased the cost of litigation and perverted the trial’s truth seeking function.\textsuperscript{265} Indeed, it took nearly a half-century

\textsuperscript{260} The scrutiny imposed was a precursor to modern substantive due process “strict scrutiny.” Barry Friedman, The History of the Countermajoritarian Difficulty, Part Three: The Lesson of Lochner, 76 N.Y.U. L. REV. 1383, 1416 (2001).

\textsuperscript{261} Kainen, supra note 199, at 140–41 (detailing the rise of the modern notion that “the existence or nonexistence of a remedy implied the existence or nonexistence of a right”).

\textsuperscript{262} Skepticism about the distinction between rights and remedies was evident in the Supreme Court’s late antebellum period Contract Clause cases, when the Court, in the 1840s, qualified the usual rule that altering remedies does not impair the underlying contract right with the proviso that legislatures may not eliminate remedies for violations of preexisting contracts. See, e.g., Bronson v. Kinzie, 42 U.S. (1 How.) 311, 317–19 (1843).

\textsuperscript{263} See, e.g., Truax v. Corrigan, 257 U.S. 312, 329–30 (1921) (withdrawal of “real remedy” for violation of a legal entitlement respecting property rights is arbitrary and capricious infringement of those rights); see also infra note 279 and accompanying text.

\textsuperscript{264} See, e.g., Rosen v. United States, 245 U.S. 467, 471 (1918) (noting the “dominance of the conviction of our time that the truth is more likely to be arrived at by hearing the testimony of all persons . . . who may seem to have knowledge of the facts involved in a case, leaving the credit and weight of such testimony to be determined by the jury or by the court”).

after the founding for the adversarial system to make significant inroads on American equity practice. Federal equity rules and leading state courts of equity, like New York’s, abandoned their judge-dominated inquisitorial features—including judicial control of discovery and ex parte judicial examination of witnesses—in favor of adversarial in-court presentation of evidence only after the 1840s.266 And as Amalia Kessler has recounted, from the 1840s until the 1870s, some prominent American procedural reformers, concerned with the “evils of lawyers and legalism” in which civil trials had become enmeshed, pushed reforms designed to undo the growth of the adversarial process, by shunting civil litigation into nonadversarial alternative dispute resolution mechanisms designed to “pacify” litigants and promote out-of-court settlements.267 Indeed, versions of these proposals made their way into a number of state constitutions and in drafts of the Field Code in the decade preceding the Civil War.268

In the antebellum period, in effect, robust adversarial proceedings were an accomplished fact in civil trials at common law but one in search of a legitimizing procedural theory.269 “[W]hile there was a long tradition of associating the common law with English constitutional liberties . . . [this] tradition . . . tended to focus piecemeal on particular . . . institutions and devices, such as habeas corpus.”270 The popular explanation of the adversarial system in terms of a larger instrumental theory, one equating robust adversarial presentation of evidence itself

English bar and among the lay public over the virtues of the adversarial system, with a focus on criminal trials).

266 Amalia D. Kessler, Our Inquisitorial Tradition: Equity Procedure, Due Process, and a Search for an Alternative to the Adversarial, 90 CORNELL L. REV. 1181, 1231–33 (2005) (identifying 1842 as the year in which the equitable inquisitorial tradition began to collapse in federal equity proceedings, although noting that American equity practice only completely assimilated to the common law adversarial system around the turn of the twentieth century).

267 The proposals centered on “tribunals of conciliation,” nonadversarial tribunals based on European models. Delegates to the 1846 New York State Constitutional Convention—“a key defining moment” in the mid century period of procedural reform, given the “leading role that New York (and its legal system) had long played in national life”—advocated a constitutional provision authorizing the establishment of these tribunals, which the New York Convention adopted (while leaving their structure to the New York legislature). Kessler, supra note 265, at 444, 446–64. Similar proposals made it into other state constitutions and drafts of the Field Code circulated in the 1840s and 1850s, and were implemented during the Reconstruction era by the Freedman’s Bureau, which, influenced by drafts of the Field Code, adopted “conciliation tribunals” for disputes between freed slaves and white landowners. Id.

268 Id. at 427, 442, 466–688.

269 As she says, while “many of the constitutive techniques of adversarial civil procedure, including, most importantly, representation by counsel and cross-examination, well predated the mid-nineteenth century (at least within the common-law courts, if not in equity), . . . individual techniques do not in themselves constitute a procedural theory.” Id. at 479–80.

270 Id. at 480.
with truth seeking and the accurate enforcement of legal rights, was a late
nineteenth-century development. 271 By the 1890s the conceptual shift had taken place. 272 The adversarial system
was an entrenched feature of both common law and equity practice, 273 reflecting a
new professional orthodoxy: a robust adversarial presentation of the evidence was
essential to the accurate determination and enforcement of individual rights. 274 By
the end of the nineteenth century, courts’ and litigants’ accounts of the value of a
hearing began accordingly to deemphasize the importance of a judicial hearing and
emphasize instead the value of a robust adversarial evidentiary hearing that
maximized the chances of determining people’s individual rights and obligations
accurately. 275 Decades after the Civil War, the three conceptual shifts—the rise of
substantive due process, concern about the effect of remedies on rights, and the

271 As Kessler shows, that theory grew out of the countrywide debates about the
virtues of conciliation courts between 1840 and 1870. Id. at 464–78, 79 ("[I]t was in the
process of opting against conciliation courts that nineteenth-century Americans opted in
favor of an ideal of formal, adversarial legal process."). And it reflected the “broader legal
context”—the collapse of the old writ system, and the ensuing attempt to forge a coherent
concept of substantive and procedural law as distinct systems, which “raise[d] for the first
time the problem of what precise functions procedure was supposed to serve.” Id. at 480.
272 It was not until the mid to late nineteenth century that American courts and
lawyers began to conceptualize procedure as a coherent system—and to think about the
defining characteristics of a good procedural system. Kessler, supra note 265, at 481–82
(collecting linguistic evidence of the conceptual shift toward thinking about procedure in
systemic terms drawn from legal authorities in the 1880s and 1890s).
273 Kessler, supra note 266, at 1232–33 (discussing gradual shift to the adversarial
model in federal equity proceedings in the post-Civil War period, which wasn’t completed
until around 1912).
274 Kessler, supra note 265, at 476–77 (in the late nineteenth century, the idea
emerges that, in disputes between free people, “conflict was a value to be promoted and the
judge’s role was not to exert personal influence in an attempt to generate compromise, but
simply to serve as a neutral umpire, ensuring that each litigant adhered to the rules
regulating combat”); id. at 480 (this view associated the adversarial system with accurate
enforcement of individual legal rights—particularly “those rights necessary to promote a
society based on both freedom and free enterprise”).
275 See, e.g., Rosen v. United States, 245 U.S. 467, 471 (1918) (associating a
“hearing” with getting at “the truth,” which “is more likely to be arrived at by hearing the
testimony of all persons of competent understanding who may seem to have knowledge of
the facts involved in a case”); Twining v. New Jersey, 211 U.S. 78, 110–11 (1908)
(associating the due process right to hearing with the right to be heard “upon evidence”);
Laing v. Ringey, 160 U.S. 531, 538 (1896) (argument of counsel) (counsel for defendant
arguing that due process requires an “opportunity to controvert the evidence on the part of
the plaintiff”); McClatchy v. Superior Court of Sacramento Cnty., 51 P. 696, 698 (Cal.
1897) (associating due process right to hearing with the right to be heard “upon evidence”);
O’Brien v. Bonfield, 72 N.E. 1090, 1092 (Ill. 1904) (equating due process right to a
hearing with the opportunity to be “heard upon all the evidence [a party] can adduce”);
State v. Beach, 46 N.E. 145, 146 (Ind. 1897) (explaining that due process right to a
hearing is denied when the law would “in effect, exclude the evidence of a party”).
new concept of the hearing-right as right to adversarial, and therefore (it was thought) more accurate, testing of evidence—combined to make the notion that due process guaranteed a right to present all the facts in civil cases increasingly plausible to mainstream American lawyers.276

First, at the heart of the Lochner era Court’s conception of fundamental rights were property rights, including not only rights to real property, but to intangible wealth.277 Damage judgments accordingly implicated the Court’s fundamental rights jurisprudence, since they deprived the defendant of a property interest, by imposing a pecuniary loss.278 Property, accordingly, could not be taken through an

276 The fusion is evident as early as an 1852 essay by Lysander Spooner—a then-marginal natural law theorist who was an archetypal man before his time. He believed, as the Lochner Court would hold, that a just legal process must protect fundamental common law liberties. See, e.g., SPOONER, supra note 133, at 23–26 (arguing that Magna Carta, with which he equated due process, was designed to nullify positive law that did not accord with natural “justice”). And, he argued, those fundamental rights are meaningless without reference to the evidentiary rules in which those rights are proven. As he said, “[i]f the government can dictate . . . the laws of evidence, it can not only shut out any evidence it pleases . . . but”—much as Chief Justice Marshall had done in Hamilton v. Russell, when he decreed certain evidence conclusive proof of a fraudulent conveyance—“it can require that any evidence whatever . . . be held as conclusive proof of any offence whatever.” Id. at 10. If the authority to determine the weight and sufficiency of evidence is vested in the government—by which he included not only legislatures, but the judge—“the government will determine its own powers over the people . . . and the people [will] have no liberties except such as the government sees fit to indulge them with.” Id. For Spooner, this dictated the proper allocation of institutional authority. “The authority to judge what are the powers of the government, and what the liberties of the people,” he said, “must necessarily be vested in one or the other of the parties themselves—the government, or the people.” Id. For Spooner, the choice was obvious. The jury must “judge of the existence of the law; of the true exposition of the law; of the justice of the law; . . . and of the admissibility and weight of all the evidence offered; otherwise the government will have everything its own way . . . and the trial will be, in reality, a trial by the government, and not a ‘trial by the country.’” Id.

Here, Spooner combined all three conceptual shifts: He insists on the protection of fundamental rights. And he views a right to a hearing as a right whose value lay in its truth-seeking value—its ability to accurately protect fundamental rights from erroneous deprivation—a value that could only be secured through the adversarial process. In order to protect the substance of fundamental rights, the integrity of that remedial process must be protected from the “government” (legislatures and judges), by preventing it from rationing the evidence that could be offered at trial. Spooner’s account is far cruder, and more radical, than any court has been willing to recognize. But it was nonetheless a harbinger of things to come.

277 See, e.g., Hovey v. Elliot, 167 U.S. 409, 418–19 (1897) (asserting that a default judgment in an equitable accounting dispenses with the “fundamental constitutional safeguards securing property”).

278 See, e.g., Brinkerhoff–Faris Trust & Savings Co. v. Hill, 281 U.S. 673, 680 (1930) (“The federal guaranty of due process extends to state action through its judicial as well as through its legislative, executive or administrative branch of government.”); Truax v.
arbitrary civil proceeding, any more than property interests could be infringed by arbitrary legislation.

Under the substantive due process wing of the *Lochner* era’s due process cases, liability rules must serve a legitimate public purpose rooted in the police power, satisfy means-ends scrutiny, and satisfy the Equal Protection Clause-derived bar on “class legislation.” But even when rules governing civil liability satisfied these substantive due process standards, procedures and remedial rules, the Court held, independently effected an arbitrary infringement, or “spoliation” of the defendant’s property if they created unacceptably high risk of an erroneous liability judgment.279

Rather than engage in difficult line drawing entailed by policing the exact risk of error that amounts to a deprivation of the right, courts settled for a blanket procedural prophylaxis: an adversarial hearing in which the defendant had an

Corrigan, 257 U.S. 312, 327–30 (1921) (characterizing “business” as including not only use of business premises, but also the pecuniary profits of the business, as a property right, and finding that remedial rules that deprive a business of meaningful protection from unlawful interference as an arbitrary infringement of that right); id. at 347 (Pitney, J., dissenting) (agreeing “[t]hat the right to conduct a lawful business, and thereby acquire pecuniary profits, is property, is indisputable,” but disagreeing that the remedial provision at issue amounted to an arbitrary infringement of that right); Dewey v. Des Moines, 173 U.S. 193, 202 (1899) (enforcing “personal liability” to pay for a property assessment would “so far as his personal liability is concerned, . . . amount to the taking of property without due process of law”); *Hovey*, 167 U.S. at 419 (“If the power to violate the fundamental constitutional safeguards securing property exists, and if they may be with impunity set aside by courts . . . [w]hy should not a court in a criminal proceeding deny to the accused all right to be heard . . . and sentence him to the full penalty of the law. No distinction between the two cases can be pointed out. The one would be as flagrant a violation of the rights of the citizen as the other, the one as pointedly as the other would convert the judicial department of the government into an engine of oppression . . .”).

279 Authorities recognizing that inadequate remedies for existing rights amount to an arbitrary infringement of the right include: *Brinkerhoff-Faris Trust*, 281 U.S. at 679, 682 (while state courts have discretion over adjective law and remedies, due process guarantees parties a “real opportunity” to establish a defense; denial of that opportunity is a “deprivation of . . . property” without due process of law); *Truax*, 257 U.S. at 329–30 (withdrawal of “real remedy” for unlawful invasion of property rights is “purely arbitrary or capricious”). Applied to one branch of remedial legislation—rules of evidence, particularly presumptive evidence—authorities interpreted this principle to require that rules must bear a “rational relationship” to the standard of liability. *See, e.g.*, Bandini Petroleum Co. v. Superior Court, Los Angeles Cnty., 284 U.S. 8, 18–19 (1931) (“The State, in the exercise of its general power to prescribe rules of evidence, may provide that proof of a particular fact, or of several facts taken collectively, shall be *prima facie* evidence of another fact when there is some rational connection between the fact proved and the ultimate fact presumed.”); Mobile, Jackson & Kansas City R.R. Co. v. Turnipseed, 219 U.S. 35, 43 (1910) (holding that an evidentiary presumption must have a “rational connection between the fact proved and the ultimate fact presumed, and that the inference of one fact from proof of another shall not be so unreasonable as to be a purely arbitrary mandate”).
opportunity to mount a meaningful defense based on the “all of the facts”—or, more specifically, all probative facts with a bearing on liability. The right to present all the facts in civil proceedings was the fruit of the Court’s early efforts to work out a notion familiar to modern lawyers: fundamental rights require heightened procedural protections against their erroneous deprivation. Lochner’s substantive and procedural due process cases were, in short, linked.

The Court’s approach to procedural due process, in turn, reflected the Lochner-era Court’s characteristic historicism: The Lochner Court, consistent with late nineteenth-century historicism, viewed historical evolution in legal practices Whiggishly—as a teleological revelation of the true principles of justice, fairness, and practical wisdom appropriate for the American people. History had given us the adversarial system, which had become entrenched by the late nineteenth century. Late nineteenth-century lawyers, looking back on its development, came to view particularized adversarial proceedings as the historical expression of an institution appropriate for a free people—an institution essential to accurately defining and enforcing individual rights. Courts, accordingly, did not need to fall back on their own fallible guesstimates about rules of evidence or procedures sufficient to ensure a reasonably accurate, and therefore nonarbitrary, judgment. History, and progress, had shown the way. Courts only needed to filter judicial

---

280 See Lindsley v. Natural Carbonic Gas Co., 220 U.S. 61, 82 (1911) (explaining due process forbids cutting off “the right to make a full defense”); Turnipseed, 219 U.S. at 43 (explaining that due process guarantees an opportunity to present a full defense based on “all of the facts”). For a discussion of these and other relevant Supreme Court decisions, see infra notes 293–313 and accompanying text.

281 See, e.g., Santosky v. Kramer, 455 U.S. 745, 753–54 (1982) (declaring that states must afford “fundamentally fair procedures” before depriving parents of a fundamental liberty interest in “personal choice in matters of family life”); Tijani v. Willis, 430 F.3d 1241, 1244 (9th Cir. 2005) (Tashima, J., concurring) (noting that “[w]hen . . . a fundamental right is at stake . . . the Supreme Court has insisted on heightened procedural protections to guard against the erroneous deprivation of that right” and discussing cases related to civil commitment proceedings).

282 Siegel, supra note 21, at 69–71 (explaining that nineteenth-century historicists believed history revealed either “laws that were true for all people” or the “permanent principles that . . . distinguish culturally correct law and legal change from culturally incorrect change” for a particular people); id. at 75 (explaining that for many historicists, “[c]ustomary law was a spontaneous and evolutionary expression of the nation. As with all natural and social phenomena . . . discoverable ordering principles governed its development”).

283 Kessler notes that late nineteenth-century lawyers, “seeking to promote a market economy and to affirm their national (and racial) superiority,” saw the adversarial system as the expression of the emerging values and institutions appropriate to a free people. Kessler, supra note 265, at 479–80. In particular, they saw the adversarial hearing as an essential element of a system dedicated to protecting individual rights. The American legal profession thought that “only formal, adversarial adjudication” could “ensure the development of clear, individual rights under the law—and especially those rights necessary to promote a society based on both freedom and free enterprise.” Id. at 480.
deprivation of property through a robust adversarial process in which parties presented all the facts bearing on the defendant’s liability.

The shift was gradual. Early cases in the 1870s and 1880s showed all the marks of transitional opinions—awkwardly mouthing the old separation-of-powers conception of due process while striking out in new directions that were difficult to pin on that old concept. An illustrative example, Chicago & Alton R.R. Co. v. People ex rel. Koerner,284 involved an enforcement action filed by Illinois’ railroad commission, alleging a railroad had engaged in rate discrimination. The Illinois constitution authorized the state assembly to “pass laws to . . . prevent unjust discrimination.”285 But a state statute provided that proof of price discrimination for similar freight transported equal distances, although from different places, “makes out a case against railway companies, which they are not allowed to meet by evidence showing the reason or propriety of the discrimination.”286

We might be inclined to see the case to present a straightforward interpretive question of substantive law—what did the framers of the Illinois constitution intend when they limited the legislature to prohibiting “unjust” discrimination? Did they think “unjust” discrimination included this sort of price discrimination? The Illinois Supreme Court, however, neatly dodged that question by conceptualizing the case as one out about due process restraints on the legislature’s power to restrict opportunities to present evidence.

The court’s reasoning was a mix of the old and the new. On the one hand, it raised questions about the fit between the statute and the purpose of the Illinois constitution’s “unjust discrimination” clause. A jury, it warned, might find that “a difference of price for the same distance of transportation[] is not necessarily an unjust discrimination,” since not all price discrimination necessarily injures the public.287 As such, the conduct on which liability hinged—simple price discrimination, an act that the court thought “may be innocent in itself”—has an insufficiently clear probative relationship to the conduct the legislature was constitutionally authorized to proscribe.288

Yet, rather than treat the risk of inaccuracy as one of fit between the “proof” made conclusive and the Illinois constitution’s “unjust discrimination” provision—an issue we, today, would say sounds in substantive due process—the court conceptualized it as a problem that implicated procedural due process, namely a right to a meaningful hearing. Echoing the reasoning of the old cases, the statute, said the court, was in effect a legislative judgment or decree: it “impose[d] . . . forfeiture of the franchise, which would often be equivalent to a fine millions of dollars” by treating the mere fact of discrimination as “conclusive evidence” that the discrimination was “unjust,” as the states constitution requires.289 As a result it

284 67 Ill. 11 (1873).
285 Id. at 23.
286 Id. at 24–25.
287 Id. at 25.
288 Id. at 24.
289 Id.
amounts to an “ex parte trial” that deprived the defendant of a “right . . . to offer what evidence it can by way of explanation.”

In doing so, though, the court also exhibited a new conception of the right to a hearing. Due process requires a hearing not simply because of the structural value of independent judicial judgment about the weight and sufficiency of adjudicative facts, implied the court, but because testing all the evidence through adversarial presentation reduces the risk that “innocent” defendants will be subjected to extraordinary penalties. The right to a hearing is no longer simply a right to a hearing before the right institution and therefore a restraint with most of its bite on legislature. It is also a right to a proceeding with safeguards against an “arbitrary” adjudication—that is, with safeguards that police against unacceptably large risks of inaccuracy. It was a right, the court’s logic implied, that restrained judges, too.

At the same time, the decision had overtones of the fundamental rights jurisprudence to come. While the court never said so explicitly, its pointed reference to the size and extent of the penalty underscored its concern that the defendant not be deprived of an important property interest without the procedural safeguards (specifically, opportunities for presenting evidence) necessary to ensure a suitably accurate, and therefore non-arbitrary, determination of guilt.

Jump forward nearly thirty years. In a series of decisions beginning in the first decade of the twentieth century, in years following *Lochner*, and extending until the very closing years of the *Lochner* era, the Supreme Court adopted and refined the logic of *Chicago & Alton Railroad*. In cases where fundamental private rights, including property interests threatened in a civil suit, were at stake, the Court held that due process permits deprivation of those interests only through a full adversarial, evidentiary hearing—one that gave the defendant an opportunity to present any evidence bearing on the merits of his defenses.

The first move came in *Hovey v. Elliot*. There, the Court overturned a default judgment entered against a defendant for contempt in a strongly worded

---

290 *Id.* at 26. The court also held that the due process clause, which guarantees a “right . . . to appear and defend in person,” is violated by the “operation of a law that substantially condemns without a trial.” *Id.* at 24. Furthermore, due process requires affording a company the opportunity “to offer what evidence it can by way of explanation.” *Id.* at 26. Similarly, the court explained that the statute envisions an “ex parte trial” by barring the defendant from “meet[ing] by evidence” the charges against them. *Id.* at 23–24.

291 *Id.* at 24, 27 (explaining that by “debarring the companies from all right of explanation,” and “taking . . . the privilege of showing the actual innocence or propriety” of the discrimination, the statute risks imposing liability on “an act that might be shown to be perfectly innocent”).

292 *Id.* at 27 (characterizing the imposition of liability as “arbitrary” because liability is conclusively presumed “from the proof of an act that might be shown to be perfectly innocent”).

293 *Id.* at 24 (emphasizing that the forfeiture of the franchise is equivalent to a fine of millions of dollars).

294 167 U.S. 409 (1897).
opinion. The default judgment, the Court said, “violate[d] the fundamental constitutional safeguards securing property.” Due process requires a hearing—in the form of a “trial”—before property of one person can be given to another. And “what plainer illustration could there be of taking property of one and giving it to another without hearing,” than a civil judgment “decree[ing]” the defendants liable to the plaintiff without “consider[ing] the merits of [defendants’] defence?” A decade later, in Mobile, Jackson & Kansas City R.R. Co. v. Turnipseed, and Lindsley v. Natural Carbonic Gas Co., the Court elaborated that a fair hearing requires an opportunity to raise a “full defense” in an evidentiary hearing that gives “the party affected a reasonable opportunity to submit . . . all of the facts bearing upon the issue” of his liability.

Due process required a “reasonable opportunity to “submit . . . all of the facts bearing” on a legal defense. But in Heiner v. Donnan the Court added that procedures rationing the defendant’s opportunity to present all probative evidence bearing on a defense were categorically unreasonable. Heiner arose out of Congress’s effort to streamline enforcement of the federal estate tax, which applied to gifts made in contemplation of death. When the Internal Revenue Service complained that it was difficult to determine “whether a particular transfer was or was not made in contemplation of death,” Congress created a conclusive presumption: gifts made within two years of the death of a decedent qualified.

---

295 Id. at 419.
296 Id. at 418 (stating that due process requires a court to “render judgment only after trial”).
297 Id. at 419, 446 (explaining that due process forbids “rendering a decree by refusing to permit a defendant to be heard in his defence or to consider the merits of a sufficient defence”).
298 219 U.S. 35 (1910).
299 220 U.S. 61 (1911).
300 Lindsley, 220 U.S. at 82 (emphasis added); Turnipseed, 219 U.S. at 43 (emphasis added).
301 Lindsley, 220 U.S. at 82 (explaining that due process requires an opportunity to present a “full defense” based on the “all of the facts bearing upon the issue”); Turnipseed, 219 U.S. at 43; see also Bandini Petroleum Co. v. Superior Court Los Angeles Cnty., Cal., 284 U.S. 8, 19 (1931) (explaining that due process requires “a reasonable opportunity to present the pertinent facts in his defense”); Jones v. Union Guano Co., 264 U.S. 171, 175 (1924) (stating that a party has a right “to try his action on the real facts”); Luria v. United States, 231 U.S. 9, 26 (1913) (due process protects a party’s right to “present his defense to the main fact thus presumed”); see also Phillips v. Ballinger, 37 App. D.C. 46, 51–52 (1911) (“Due process of law, in a case like the present, is had when full opportunity is presented to introduce all the evidence and arguments which the party interested deems important, and to be confronted with witnesses against him, he having had notice of the question at issue.”).
302 285 U.S. 312 (1932).
303 Id. at 312–14.
304 Id. at 313.
305 Revenue Act of 1926, ch. 27, § 302(c), 44 Stat. 9, 70 (1925).
The respondents, executors of an estate subject to IRS enforcement proceedings, claimed the presumption violated due process.

The Court agreed in a decision by Justice Sutherland that emphasized procedural, not substantive due process. Indeed, to connect the case with procedural due process cases like *Turnipseed*, the Court felt compelled to emphasize that “the conclusive presumption created by the statute” is a rule of evidence, not a rule of substantive law.306 “A rebuttable presumption,” said the Court, “clearly is a rule of evidence which has the effect of shifting the burden of proof . . . and it is hard to see how a statutory . . . presumption is turned from a rule of evidence into a rule of substantive law as the result of . . . making it conclusive. In both cases it is a substitute for proof.”307 And, citing *Turnipseed*,308 the Court concluded the presumption deprives a defendant of due process because it deprives him of a fair opportunity to “prove the facts of his case”309 by offering into evidence “every fact and circumstance tending” to support his legal defense—here, “to show the real motive of the donor.”310 Imposing liability on such terms amounts to an arbitrary taking, a “spoliation,” of the defendant’s property.311

Throughout this period, the Court maintained a high degree of continuity with the specific holdings of nineteenth-century cases, by recognizing that the legislature had a significant degree of flexibility to regulate the rules of evidence and the substantive defenses to which defendants are entitled. It might adjust burdens of proof, by legislating rebuttable presumptions, so long as the presumptions bore a rational connection to the ultimate fact proved.312 It might alter rules governing admissibility, based on suitably reasoned judgments about the evidence’s probative value.313 It might eliminate traditional legal defenses entirely, like the defense of contributory negligence or assumption of risk.314 These were the

307 Id. at 329.
308 Id.
309 Id. (explaining that “if a legislative body is without power to enact as a rule of evidence a statute denying a litigant the right to prove the facts of his case, certainly the power cannot be made to emerge by putting the enactment in the guise of a rule of law”).
310 Id. at 327 (stating that the presumption is arbitrary because it “excludes consideration of every fact and circumstance tending to show the real motive of the donor”); see also id. at 329 (stating that “a statute creating a presumption which operates to deny a fair opportunity to rebut it violates the *due process clause of the Fourteenth Amendment*” (emphasis added)).
311 Id. at 327 (asserting that an exaction under these rules “is not taxation but spoliation”).
313 See *West v. Louisiana*, 194 U.S. 258, 263–64 (1904) (explaining that states have wide latitude to adopt “reasonable” rules on admissibility, so long as they do not “substantially affect” the rights of the accused party).
314 See, e.g., Ariz. Employers’ Liability Cases, 250 U.S. 400, 421 (1919) (holding that common law defenses, like contributory negligence, “are not placed, by the *Fourteenth Amendment*, beyond the reach of the State’s power to alter them, as rules of future conduct
kind of “reasonable” regulations of opportunities to present evidence that *Turnipseed* greenlighted. But within the universe of probative evidence bearing on defenses that remain legally available, it simply could not bar, or ration, a defendant’s opportunity to present *all* the probative evidence—“every fact and circumstance”—bearing on those defenses.

**B. Lochner Overthrown**

1. **The Demise of the Lochnerian Right to a Hearing**

   The *Lochner* era came to an end in the late 1930s in *West Coast Hotel v. Parrish*[^315] and *United States v. Carolene Products Co.*[^316]—decided just five years after *Heiner*.[^317] While the post-New Deal Court did not, ultimately, repudiate the view that due process protects certain “fundamental” rights, *West Coast Hotel* repudiated the *Lochner* Court’s distinctive notion that economic interests are among those rights.[^318] Yet, it was a concern for the fundamental safeguards securing property that drove the Court’s reasoning in *Turnipseed* and *Heiner*.[^319] As a result, after *West Coast Hotel*, the future prospects of the *Lochner* era’s procedural due process cases (at least in civil cases) seemed gloomy. It would be strange, after all, to deny courts acting with the scope of their delegated authority over evidence and procedure the same flexibility the post-New Deal Court’s

[^315]: 300 U.S. 379 (1937).
[^316]: 304 U.S. 144 (1938).
[^317]: *West-Coast Hotel*, 300 U.S. at 379; *Carolene Products*, 304 U.S. at 152 & n.4.
[^318]: The post-*Carolene Products* Court has recognized a number of fundamental (noneconomic) liberty interests, including family and sexual privacy, but it has never recognized fundamental “property” interests. See, e.g., Ronald J. Krotoszynski, Jr., *Fundamental Property Rights*, 85 Geo. L.J. 555, 555 (1997) (“The Supreme Court’s modern substantive due process jurisprudence, which protects ‘fundamental’ liberty interests . . . has never been formally extended to encompass ‘fundamental’ property rights.”). Perhaps it may sweep too broadly to say that the modern category of fundamental rights categorically excludes *any* interests in “property.” See id. at 556 (“The most obvious explanation would be that property rights are simply less important than liberty interests and therefore are never ‘fundamental.’ However, the Supreme Court has never suggested that this is so, and, if it were, the Court’s most recent Takings Clause decisions would not make overall doctrinal sense.”). At a minimum, though, intangible corporate property interests in accumulated wealth (cash) implicated by civil liability judgments are clearly nonfundamental. See, e.g., E. Enters. v. Apfel, 524 U.S. 498, 550 (1998) (Kennedy, J., concurring in judgment and dissenting in part) (asserting that statutes imposing monetary liability implicate “the deferential standard of review applied in substantive due process challenges to economic legislation” and violate due process “only under the most egregious of circumstances”).
[^319]: *See supra* notes 298–313 and accompanying text.
substantive due process cases granted legislatures when property interests and other economic interests are at stake.

It’s hard to identify exactly when the Court first hinted at a break with cases like *Turnipseed* and *Heiner*, but a fair candidate is a foundational case in the post-New Deal Court’s procedural due process jurisprudence: *Hansberry v. Lee*, decided only eight years after *Heiner* but on the other side of the New Deal constitutional revolution. *Hansberry*, of course, stands for the proposition that due process requires adequate—meaning conflict-free—representation of class members. It says nothing about the right to present “all the facts” before one can be deprived of property. But that is the point. Based on the arguments presented, and the facts of the case, the Court might have done so, but it declined. *Hansberry*, as a result, might be seen as a silent, but significant, first step toward repudiating the *Lochner*-era understanding of procedural due process.

*Hansberry* involved a collateral attack on an Illinois class judgment upholding the validity of a racially restrictive covenant in Washington Park, a neighborhood adjacent to the University of Chicago’s Hyde Park enclave. In a prior lawsuit, *Burke v. Kleiman*, one Washington Park resident, purporting to represent the other residents, sued to enjoin a neighbor from renting a room to an African American family. The suit ended in a judgment upholding the covenant and enjoining the defendant from continuing the rental arrangement.

Washington Park residents sued years later to prevent Hansberry from buying a house in the neighborhood, relying on the restrictive covenant. Hansberry fought back, arguing that terms governing the covenant’s enforceability had not been satisfied. The plaintiffs countered that whether the agreement was effective was subject to *res judicata* as a result of the *Burke* judgment.

One of Hansberry’s counterarguments was a familiar one: that he was not adequately represented by the class representative, whose interest—in enforcing the covenant!—was diametrically opposed to his. But alongside this sensible argument, the NAACP sounded older, *Lochner*-era themes, which commentators have universally forgotten, or ignored.

The *Lochner*-esque argument rested on a singular fact about the *Burke* suit—it had not involved a presentation of all the facts bearing on the enforceability of the covenant. To be valid, the covenant must have obtained signatures of owners whose property accounted for 95 percent of the development. Rather than establish this fact in an adversarial proceeding after a fact-intensive presentation of

---

320 311 U.S. 32 (1940).
321 See *id.*
322 *Id.* at 42–44.
323 277 Ill. App. 519 (1934).
324 *Id.* at 521.
325 *Hansberry*, 311 U.S. at 37.
326 *Id.* at 38.
328 See *id.* at 35.
proof, the parties to the \textit{Burke} suit had simply stipulated that the requisite number of signatures had been obtained.\footnote{329 \textit{Burke}, 277 Ill. App. at 522 (noting the terms of the stipulation).} Instead, the \textit{Burke} suit was litigated entirely on a collateral question: whether conditions in the neighborhood (namely its racial mix) had changed to such a degree that enforcement of the covenant would work a hardship on residents.\footnote{330 \textit{Id.} at 521 (characterizing defendant’s argument).} We are inclined to see this problem either in the terms of issue, rather than claim preclusion—that is, the “judgment” respecting the number of signatures cannot be given \textit{issue} preclusive effect because the issue had not been actually litigated (a defense, incidentally, unhelpful in \textit{Hansberry}, a suit exposed to the separate defense of claim preclusion) or as further evidence of compromised representation. But \textit{Hansberry} presented the problem, in part, in terms reminiscent of the \textit{Lochner} Court’s procedural due process cases. At stake was Hansberry’s personal right to buy property, a fundamental interest.\footnote{331 \textit{See}, e.g., Brief of Petitioner, \textit{supra} note 327, at 53–54, 58–59 (arguing the right to own, “use,” and otherwise “deal” in property is a “fundamental right[] as a citizen of the United States’); Reply Brief of Petitioner at 15, \textit{Hansberry} v. \textit{Lee}, 311 U.S. 32 (1940) (No. 40-29) (noting that a “long line of decisions in this Court” extended due process to secure the “protection of other than real property rights,” including the right to “do business” in property).} He could not be deprived of that right through judicial decree without requisite procedural protections—namely an adversarial hearing that provided a meaningful opportunity to present all probative evidence in defense of his property rights.\footnote{332 Reply Brief of Petitioner, \textit{supra} note 331, at 12–15 (contending that Hansberry’s right to own and deal in property is a due-process protected property interest and that taking that interest without, among other protections, a meaningful opportunity to be heard in defense based on all the facts amounts to a spoliation, or arbitrary infringement, of his property rights).} Where, as here, the class suit was litigated based on a stipulated presumption that the covenant had obtained the requisite signatures, it could not, consistent with due process, be given preclusive effect when absentees present strong evidence that the stipulation was factually unfounded.\footnote{333 Brief of Petitioner, \textit{supra} note 327, at 36–37 (arguing that where the record shows “innumerable instances . . . where the individual purported parties signatory could have made substantial defenses,” application of the doctrine of \textit{res judicata} deprives them of a right to present a defense and, therefore, violates their due process rights); \textit{see} Reply Brief of Petitioner, \textit{supra} note 331, at 14–15 (giving the prior judgment preclusive effect amounts to a spoliation, or “arbitrary” deprivation, of their property rights).} Doing otherwise would deprive him, and other similarly situated class members, of a meaningful opportunity to make a full defense against the enforcement of the covenant based on all the facts, and amount to an “arbitrary” confiscation, or spoliation, of their property interests.\footnote{334 Reply Brief of Petitioner, \textit{supra} note 331, at 14–15.} In effect, the NAACP invited the Court to extend the logic of its procedural due process cases into the law of former adjudication—by recognizing that where a fundamental property right is at stake, concern for ensuring against inaccurate

\footnote{329 \textit{Burke}, 277 Ill. App. at 522 (noting the terms of the stipulation).} \footnote{330 \textit{Id.} at 521 (characterizing defendant’s argument).} \footnote{331 \textit{See}, e.g., Brief of Petitioner, \textit{supra} note 327, at 53–54, 58–59 (arguing the right to own, “use,” and otherwise “deal” in property is a “fundamental right[] as a citizen of the United States’); Reply Brief of Petitioner at 15, \textit{Hansberry} v. \textit{Lee}, 311 U.S. 32 (1940) (No. 40-29) (noting that a “long line of decisions in this Court” extended due process to secure the “protection of other than real property rights,” including the right to “do business” in property).} \footnote{332 Reply Brief of Petitioner, \textit{supra} note 331, at 12–15 (contending that Hansberry’s right to own and deal in property is a due-process protected property interest and that taking that interest without, among other protections, a meaningful opportunity to be heard in defense based on all the facts amounts to a spoliation, or arbitrary infringement, of his property rights).} \footnote{333 Brief of Petitioner, \textit{supra} note 327, at 36–37 (arguing that where the record shows “innumerable instances . . . where the individual purported parties signatory could have made substantial defenses,” application of the doctrine of \textit{res judicata} deprives them of a right to present a defense and, therefore, violates their due process rights); \textit{see} Reply Brief of Petitioner, \textit{supra} note 331, at 14–15 (giving the prior judgment preclusive effect amounts to a spoliation, or “arbitrary” deprivation, of their property rights).} \footnote{334 Reply Brief of Petitioner, \textit{supra} note 331, at 14–15.}
judgments trump considerations of finality when the property right hadn’t been litigated in a prior, fact-intensive proceeding. The argument, though, was thinly sketched by Hansberry’s lawyers, who struggled to connect the case to the old property-right cases without directly invoking the *Lochner* era’s economic liberties jurisprudence.335

The *Hansberry* Court ignored this argument without comment—a hint, perhaps, at a shift away from the *Lochner* era’s approach to procedural due

335 For example, Hansberry’s lawyers tried to link claim to older cases that limited representative actions to “joint” rather than “personal” rights: “personal” rights aren’t susceptible of representation, the argument went, and therefore the representatives can’t waive defenses relating to the personal right. Brief of Petitioner, supra note 327, at 38 (explaining that various state law defenses to the covenant’s enforcement are “personal to each signer” and cannot be represented in a class proceeding); see also Robert G. Bone, *Personal and Impersonal Litigative Forms: Reconceiving the History of Adjudicative Representation*, 70 B.U. L. REV. 213 (1990) (reviewing Stephen C. Yeazell, *From Medieval Group Litigation to the Modern Class Action* (1987)) (discussing the old distinction between representation of personal and joint or common rights). As the respondents noted, though, the Supreme Court had held that preclusive effect of a judgment was a matter for state law, and Illinois, in turn, had not recognized the personal rights limitation on representative actions, holding instead that representative actions were proper when claims were too numerous to efficiently litigate separately; it had, for good measure, extended broad preclusive effect to representative actions, including to stipulations and consent decrees entered into by class representatives. See Brief of Respondents at 23–26, 28–29, 31, *Hansberry v. Lee*, 311 U.S. 32 (1940) (No. 40-29) (collecting cases).

In their reply brief, the petitioners invoked precedent that anticipated the modern rule in *Cooper v. Fed. Reserve Bank of Richmond*, 467 U.S. 867 (1984), that class proceedings preclude only “common” claims, but not individual ones—an argument that, however, wasn’t applicable to the failure of the covenant’s condition precedent, which raised a question common to all property owners covered by the covenant. See Brief of Petitioner, supra note 327, at 38–39 (noting the defense-based failure of the condition precedent was shared by all class members).

And so the reply brief also sounded a more general theme reminiscent of *Lochner*-era cases like *Heiner* and *Turnipseed*: it (1) reiterated that the state decree implicated fundamental rights to deal in property and, therefore, implicated the due process clause’s core protections, see Reply Brief of Petitioner, supra note 331, at 15, and (2) emphasized that, before being deprived of that right, they must have a meaningful opportunity “to defend, and to be heard,” including a meaningful opportunity to present a full defense, based on all the facts. See id. at 12–13 (arguing that giving the *Burke* decree’s stipulation preclusive effect deprives them of that right’); id. at 14–15 (arguing that giving the state decree preclusive effect amounts to an arbitrary spoliation of their property interests and therefore violates the Fourteenth Amendment’s due process clause).

The closest the petitioner came to explicitly invoking the *Lochner* line of economic liberties cases, though, is in the reply brief’s pointed invocation (without citation) of the “long line of decisions in this Court” extending due process protections to rights “other than real property rights” and its cite to *Brinkerhoff-Faris* (the lone right to a meaningful defense decision of the *Lochner* era authored not by the “Four Horsemen,” or the old Fuller Court libertarians, but by a progressive—Justice Brandeis). Id. at 12, 15.
process. That demise of *Lochnerian* procedural due process would become explicit three decades later in a widely misunderstood case: *Weinberger v. Salfi*.336

*Salfi* followed hard upon two earlier (and, today, much maligned)1970s decisions, *Stanley v. Illinois*338 and *Cleveland Board of Education v. LaFleur*;339 in which a Brennan-lead Court had revived *Heiner* to strike down conclusive presumptions that restricted parties’ ability to prove their rights in state administrative proceedings affecting core privacy interests.340 In each, the state defended the presumption as a means of ensuring a tolerably accurate, yet “speed[y] and efficien[t],” resolution of most underlying claims.341 In each, the Court analyzed the constitutionality of the presumption in procedural due process terms that anticipated *Mathews v. Eldridge*’s weighing method of scrutiny.342 That is, the Court weighed the interests of the party threatened with deprivation, the error-risks of applying the presumption, and state interests in administrative efficiency.343 And in each case, the Court concluded, the need to safeguard the important interests of the party disadvantaged by the presumption from an erroneous deprivation outweighed the state’s concerns in administrative efficiency.344

In *Weinberger v. Salfi*, however, the Court drew a line in the sand. The respondent in *Salfi*, the surviving spouse of a deceased wage earner, claimed an entitlement to her husband’s social security benefits. A statutory provision, however, defined “wives” entitled to survivorship benefits to include only those spouses who had a marital relationship with the deceased wage earner for over nine months before the wage earner’s death.346 The government characterized this provision as a conclusive presumption that marriages entered into less than nine months before the spouse’s death were “sham[s]” designed to obtain social

---

337 *See, e.g.*, Easterbrook, *supra* note 12, at 113 (referring to the Court’s “late, un lamented flurry of [1970s-era] irrebuttable presumption cases”).
341 *Stanley*, 405 U.S. at 656 (noting the state defended the presumption based on administrative efficiency); *see also LaFleur*, 414 U.S. at 646–67.
343 424 U.S. 319, 334–45 (1976); *see also* Michael H. v. Gerald D., 491 U.S. 110, 153 & n.10 (1989) (Brennan, J., dissenting) (noting that cases like *Stanley* had analyzed irrebuttable presumptions through the lens of procedural, not substantive, due process).
344 *Stanley*, 405 U.S. at 656–57 (“The establishment of prompt efficacious procedures to achieve legitimate state ends is a proper state interest worthy of cognizance in constitutional adjudication. But the Constitution recognizes higher values than speed and efficiency . . . [W]hen, as here, the procedure forecloses the determinative issues of competence and care, when it explicitly disdains present realities in deference to past formalities, it needlessly risks running roughshod over the important interests of both parent and child.”); *LaFleur*, 414 U.S. at 646–67.
345 422 U.S. 749 (1975).
security benefits. The presumption, said the government, was a sensible prophylactic rule: it did a reasonable job of filtering sham relationships, avoided the enormous “expense and other difficulties of individual determinations” in a massive program the size of social security, and “protect[ed] large numbers of claimants who satisfy the rule from the uncertainties and delays of administrative inquiry into the circumstances of their marriages.”

Relying on Stanley and LaFleur, Salfi in turn claimed the “sham marriage” presumption deprived her of a fair hearing, because it denied her a right to rebut the presumption with particularized proof. The Court disagreed. In cases like Stanley, said the Court, the challenged presumption limited parties’ ability to resist a deprivation of fundamental liberty interests—to “conceive and to raise one’s children” or to exercise “personal choice in matters of marriage and family life.” The force of these decisions, emphasized the Court, was limited to cases implicating these fundamental interests (interests, it said, “far more precious . . . than property rights.”) By contrast, Salfi asserted an interest in welfare benefits that, said the Court, was not constitutionally protected. Therefore, Salfi had no basis for claiming any procedural due process protections.

Because Salfi treated a right to welfare benefits as a constitutionally unprotected interest, the Court didn’t need to engage in procedural due process balancing. But the fact the Court considered the nature of the interest in the first place implied rather strongly that it thought irrebuttable presumptions require a procedural due process evaluation when the presumption does affect constitutionally protected interests. Moreover, the Court’s emphasis that Stanley and LaFleur had involved deprivations of fundamental interests “far more precious . . . than property rights” had clear implications. In this case, even assuming welfare benefits could be considered constitutionally protected property, the government’s interest in administrative efficiency would easily outweigh the welfare beneficiary’s interest in the protection afforded by a maximally fact-intensive investigation.

---

347 See Salfi, 422 U.S. at 767–68.
348 Id. at 777, 781–82.
349 Id. at 771–72.
350 Id. at 771 (quoting Stanley v. Illinois, 405 U.S. 645, 651 (1972)).
351 Id. at 771 (quoting Cleveland Bd. of Educ. v. LaFleur, 414 U.S. 632, 639 (1974)).
352 Id. (quoting Stanley, 405 U.S. at 651).
353 After reviewing cases like Stanley and LaFleur, it emphasized the fundamental nature of the interests at stake in each and then concluded that “these cases are not controlling on the issue before us now,” because “[u]nlike the claims involved in Stanley and LaFleur, a noncontractual claim to receive funds from the public treasury enjoys no constitutionally protected status.” Id. at 771–72.
354 Id. at 771 (quoting Stanley, 405 U.S. at 651).
355 Cf. id. at 804–05 (Brennan, J., dissenting) (criticizing the majority’s view that efficiency outweighs “the individual’s interest in proving that the facts presumed are not true as to him”).
Having dispatched plaintiff’s procedural due process claim, the Court noted the plaintiff’s only remaining constitutional claim is that the presumption is not “rationally related to a legitimate legislative objective” and then proceeded to analyze the presumption as though it were a rule of substantive law, employing rational basis scrutiny. The Court found it did in fact satisfy rational basis scrutiny: “Congress can rationally conclude . . . that the difficulties of individual determinations” outweigh whatever gains in accuracy “they might be expected to produce.” Analyzed either from a procedural or a substantive due process standpoint, the presumption didn’t offend due process.

Salfi also (gratuitously) emphasized its reasoning applied not only to presumptions in cases involving claims for welfare benefits, but also to those that limit a defendant’s ability to prove defenses in civil suits—proceedings, in other words, affecting the kind of traditional property interests in intangible wealth that the Lochner Court had once viewed as fundamental, but that Salfi emphasized were not fundamental. Extending a Heiner-like protection against the use of presumptions that streamline civil litigation, by rationing opportunities to present particularized proof, throws a monkey wrench in the regulation of the economy, the Court warned. Therefore, it warred with the modern construction of the Fifth and Fourteenth Amendment, which permits wide regulatory latitude in just that area.

By suggesting its analysis applied to presumptions in ordinary civil litigation, Salfi sounded the death knell of the old Lochner-era conception of procedural due process. It strongly implied that where the interest at stake is a simple economic or property interest—as where a civil defendant is contesting civil liability—restricting opportunities for presenting particularized proof in defense of that lower-priority interest comports with due process so long as the evidence given legal effect has at least some rational probative relationship to standards of liability, and limiting the availability of particularized hearings serves important public interests, by reducing onerous litigation costs, promoting more consistent outcomes, or ensuring effective deterrence.

2. Scalia Sows Confusion

Unfortunately, Salfi’s significance is sometimes misunderstood, thanks in large part to Justice Scalia’s narrow reinterpretation of the conclusive presumption

356 Id. at 772.
357 Id. at 785.
358 See supra notes 350–55 and accompanying text.
359 Salfi, 422 U.S. at 772, 773 (explaining that “extend[ing] . . . Stanley . . . and LaFleur to the eligibility requirement in issue here would turn the doctrine of those cases into a virtual engine of destruction for countless legislative judgments which have heretofore been thought wholly consistent with the Fifth and Fourteenth Amendments,” including those “arising from legislative efforts to regulate private business enterprises”).
cases as pure substantive due process cases in his plurality opinion in *Michael H v. Gerald D*\(^{360}\) a decade later—an opinion that only the Chief Justice joined in full.

Irrebuttable presumptions, as noted earlier, fall within the nether zone of rules that reasonable people might classify as either substantive or procedural.\(^{361}\) Throughout the nineteenth century and into the *Lochner* era, courts conceptualized presumptions as procedural rules—that is, rules regulating the opportunity to present evidence. Today, however, courts view irrebuttable presumptions as rules defining parties’ substantive rights.\(^{362}\)

Justice Scalia shares this modern bias. Irrebuttable presumptions, he declared in *Michael H*, are simply general statutory classifications—rules that define substantive rights and reflect substantive policies.\(^{363}\) According to Scalia, irrebuttable presumption cases like *Salfi* “must ultimately be analyzed as calling into question not the adequacy of procedures but—like our cases involving classifications framed in other terms—the adequacy of the ‘fit’ between the classification and the policy that the classification serves.”\(^{364}\)

Scalia is right about many things, but he was wrong here. The *Stanley* Court had clearly treated the presumption there as a procedural due process problem, as the Court reaffirmed in the decade after it (and *Salfi*) were decided.\(^{365}\) Though it is true that part of the *Salfi* Court’s opinion analyzed the presumption at issue there in rational-basis terms (and found the presumption survived rational-basis scrutiny), Scalia ignored that it did so only after rejecting *Salfi*’s procedural due process argument on the merits, based on the nature of the interest at stake, implying, in turn, that the *Salfi* Court thought irrebutable presumptions do require procedural due process analysis when they affect constitutionally protected interests.\(^{366}\)

Scalia’s view of *Salfi* was, at least at the time, idiosyncratic. The majority of justices in *Michael H* agreed *Salfi* and its precursors like *Stanley* and *LaFleur* addressed a problem with procedural due process dimensions.\(^{367}\) As Justice

\(^{360}\) 491 U.S. 110 (1989).

\(^{361}\) *See supra* notes 131–47 and accompanying text.

\(^{362}\) *See supra* notes 131–47 and accompanying text.

\(^{363}\) 491 U.S. at 119–21.

\(^{364}\) *Id.* at 121.


\(^{366}\) *Michael H.*, 491 U.S. at 144 (Brennan, J., dissenting) (“[T]he Court fit the complaint in *Salfi* into the [substantive rather than procedural due process] category on the ground that the challenged law did not deprive anyone of a constitutionally protected interest. . . . Today’s plurality, in contrast, classifies this case as one invoking substantive due process before it considers the nature of the interest at stake.”).

\(^{367}\) *See RHONDA WASSERMAN, PROCEDURAL DUE PROCESS: A REFERENCE GUIDE TO THE UNITED STATES CONSTITUTION* 86–87 (2004) (the majority of justices in *Michael H*—Brennan, Marshall, Blackmun, White and Stevens—treated the problem of irrebutable presumptions as a procedural, not a substantive, due process problem). As Justice Brennan put it, in cases like *Salfi*, the State, by “declar[ing] a certain fact relevant, indeed controlling,” while “den[y]ing a particular class of litigants a hearing to establish that fact,” presents “precisely [the sort of issue that concerns] procedural due process.” *Michael H*,
Brennan put it, in cases like *Stanley* and *Salfi*, the State, by “declare[ing] a certain fact relevant, indeed controlling,” even as it “den[ies] a particular class of litigants a hearing to establish that fact” presents “precisely [the sort of issue that concerns] procedural due process.” Unfortunately, Scalia’s reading has won the day. Today, *Salfi* is remembered as a case that decided that legislatively mandated irrebuttable presumptions are substantive rules that should be analyzed under the substantive strain of the Court’s due process jurisprudence.

In some class actions that are the focus of this Article, the import of *Salfi* (even interpreted in the narrow and misleading way Scalia suggests) is clear: Some trial courts seem to conceive of the presumptions they adopt as interpretations of parties’ substantive rights. Consider, for example, a case in which a court rules that the underlying substantive law allows courts to irrebuttably presume reliance based on proof of uniform written misrepresentations of a product’s merits in cases involving mass injuries. In these cases, Scalia’s interpretation of *Salfi* tells us that the interpretation (assuming it is consistent with legislative intent) passes muster if it satisfies rational basis scrutiny—that is, if it’s rational to think the presumption’s over- or underinclusiveness is justified in order to conserve on the excessive cost and unpredictability of particularized evidentiary hearings.

---

491 U.S. at 153 (Brennan, J., dissenting). *Salfi*, Justice Brennan emphasized, did not hold “that a challenge to a conclusive presumption must rest on substantive rather than procedural due process”; it applied rational basis scrutiny to the presumption only after concluding that the presumption did not violate procedural due process. *Id.* As Rhonda Wasserman notes, Brennan, Marshall, Blackmun, White, and Stevens also agreed that at least in cases implicating important liberty interests, procedural due process entitled parties to the presentation of particularized probative evidence. See Wasserman, supra, at 85–87.

While the majority in *Michael H* agreed on these points, they disagreed about how to apply them. Four, led by Justice Brennan, dissented, asserting that the conclusive presumption at issue in the case had denied the petitioner a fair hearing. See 491 U.S. at 153 (Brennan, J., joined by Marshall and Blackmun, J.J., dissenting) (characterizing irrebuttable presumptions as procedural due process problem); id. at 157, 160–61, 163 (White, J., joined by Brennan, J., dissenting) (arguing the presumption denied the petitioner procedural protection due before depriving him of a fundamental liberty interest, by denying him an opportunity to rebut the presumption). One of the five, Justice Stevens, decided that the statute at issue, properly interpreted, didn’t preclude the petitioner from presenting evidence to establish the liberty interest he was seeking to enforce. See *id.* at 134–35 (Stevens, J., concurring) (“I recognize that my colleagues have interpreted [the statute] as creating an absolute bar that would prevent a California trial judge from regarding the natural father as either a ‘parent’ within the meaning of the first sentence of § 4601 or as ‘any other person’ within the meaning of the second sentence. That is ... an unnatural reading of the statute’s plain language ....”).

368 *Id.* at 153 & n.10 (Brennan, J., dissenting) (noting that “[t]he plurality’s bald statement that the holding in *Stanley* did not rely on procedure due process is therefore incorrect,” and noting, too, that *Salfi* dismissed the relevance of cases like *Stanley* only after concluding the statute in *Salfi* did not implicate a significant interest).

But in a variety of other class cases, courts are not clearly purporting to issue
a substantive ruling. Instead, they seem to assume they are exercising discretion,
delegated to them by Congress, to ration defendants’ procedural opportunities to
rebut proof of their liability. This is certainly the case in suits adopting sampling
or formulaic methods of proof. Courts do not seem to characterize sampling or
formulaic proof as changes in the parties’ legal rights. Instead, they seem to
characterize their decisions procedurally—as modification of the parties’
opportunities to present proof of those rights in order to accommodate plaintiff’s
use of the class device. True, sampling and formulaic proof both employ
irrebuttable presumptions. Sampling employs an irrebuttable presumption that
class members suffered an injury equivalent to the average claim in a test sample,
while formulaic proof employs an irrebuttable presumption that the proof identifies
real victims of discrimination. But it is not a presumption that applies generally,
as a new rule creating new substantive rights does. The irrebuttable presumption
arises because, for reasons of economy and equity, the court has exercised its
discretion to forbid defendants from rebutting that proof with the particularized
evidence they would be entitled to present in a nonaggregated proceeding.

If we take these decisions on their own terms—as procedural and not
substantive decisions—those who uncritically accept Scalia’s narrow interpretation
of Salfi risk missing Salfi’s import. When a defendant’s interest is in keeping its
property—in avoiding civil liability—Salfi implies that limitations on the
defendant’s opportunity for a particularized hearing also comport with procedural
due process so long as the insuperable difficulty of “individualized determinations”
would impose enormous costs of the legal system and hamstring the deterrence
objectives of the law, and so long as the evidence used to establish liability is
roughly accurate.

Salfi, in effect, anticipates the same conclusions a host of lower courts have
reached when considering restrictions on class defendants’ opportunities under
Mathews and Doehr, decided after Salfi. Namely, where defendants’ sole interest is
in avoiding or limiting damages, “adversarial resolution of each class member’s
claim would pose insurmountable practical hurdles” for plaintiffs and impose
enormous ancillary costs on courts, and evidentiary shortcuts are tolerably accurate
in the run of cases, limitations on defendants opportunities to present
individualized evidence “[do] not violate due process.”

---

370 See supra Part I.A.
371 See supra notes 42–58 and accompanying text.
372 See supra notes 42–58 and accompanying text.
373 See supra note 42 and accompanying text.
2004) (characterizing the decision to adopt formulaic proof as a pragmatic response to the
realities of class litigation—a form of “rough justice”); see also supra notes 38–58 and
accompanying text.
375 Hilao v. Estate of Marcos, 103 F.3d 767, 786–87 (9th Cir. 1996).
3. The Early Tradition Restored

Uncritical acceptance of Scalia’s myopic (if fashionably) view of irrebuttable presumptions, it bears (re)emphasizing, also risks blinding us to the significance of the historical record. Much of the evidence about the original understanding of due process developed here is drawn from a long skein of cases, dating from the earliest years after the framing, dealing with the due process implications of conclusive, or irrebuttable, presumptions—the most thoroughgoing restraint on opportunities to present proof possible. As we have seen, lawyers of the period conceived of these presumptions, in procedural and evidentiary, not substantive, terms, as rules denying defendants opportunities to present probative proof, rather than rules defining substantive rights. And, pre-Lochner, courts thought that sort of limit—when adopted as a matter of independent judicial discretion, in the service of equity and convenience—didn’t violate defendants’ procedural due process rights.

If we want to understand how the nineteenth-century legal community understood what constitutes due process, we have to look at these cases through their eyes, not Justice Scalia’s. And when we do, it is evident that the post-New Deal procedural due process cases and the older understanding converge.

At the beginning of the Court’s long effort to work out the content of “due process,” Chief Justice Marshall in Hamilton v. Russell had barred fact-intensive investigations of conveyances by debtors, by adopting a conclusive presumption that conveying title to property without relinquishing possession is conclusive evidence of fraud. And he had done so for many of the reasons the Burger Court would uphold the conclusive presumption challenged in Salfi. The presumption in Hamilton also reflected sensible effort to conserve on litigation costs, promote predictability, and in doing so, forestall efforts to manipulate legal proceedings to effectuate a “sham” or fraud on the legal system.

In line with the antebellum conception of due process, Marshall had not seemed to think that rationing parties’ opportunities for proving their claims or defenses based on these considerations violated parties’ constitutional rights. The Burger Court did not see any problem with similar restrictions on proof, either. Despite a gulf of two centuries, the pragmatic reasoning that led the old Marshall Court and the modern Burger Court to these parallel conclusions was fundamentally the same.

Salfi departed from antebellum cases in a nuanced way: its belief that judicial approval of a presumption of fact is guided by a (loose) constitutional requirement that the presumption reflect a sensible weighing of litigants’ interests,

376 See supra notes 360–74 and accompanying text.
378 See supra notes 129–47 and accompanying text.
379 5 U.S. (1 Cranch) 309, 318 (1803).
380 See supra notes 176–89 and accompanying text.
381 See supra notes 337–39 and accompanying text.
administrative efficiency, and public policy (requirements later formalized in Mathews and Doehr). The old antebellum cases, in contrast, had thought due process required nothing more than independent judicial sanction and confirmation of the weight and effect of specific evidence.382

Nuances aside, the approaches are fundamentally similar. Due process does not impose any firm restrictions on alterations in the type or quantity of evidence presented in ordinary civil proceedings. Everything, in practice, turns on a judicial judgment, guided by dictates of “good sense, equity, and convenience.” In Salfiti, the essentials of the old antebellum due process tradition had been, quietly, restored.

CONCLUSION:
APPRASING CLASS ACTION DEFENDANTS’ NEW LOCHNERISM

The Due Process Clause, says Wal-Mart, guarantees a right to defend against civil claims with “any relevant rebuttal evidence they choose,”383 But for most of our history, courts have thought otherwise.384 The common law imposed a variety of limits on parties’ ability to present probative evidence to economize on the cost and unpredictability of jury fact-finding.385 And, not surprisingly, after ratification, American courts—the inheritors of that tradition—did not construe the Due Process Clause to require a maximally fact-intensive evidentiary hearing. They treated it as a structural right to adjudication by the right institution—an independent judiciary.386 The idea that due process might bar courts from doing what they had always done—refine, and sometimes restrict, opportunities to present probative evidence in common law fashion in the service of equity and convenience—never seemed to have occurred to them.

Class action defendants bid to revive a different, brief-lived understanding from the Lochner era. Due process, the Lochner Court thought, guarantees a full-throated adversarial evidentiary hearing, in which each side has an opportunity to present “every fact and circumstance”—or “any relevant . . . evidence they choose,” as Wal-Mart says—bearing on liability.387 Class defendants agree.

Given its continuing status as an antiprecedent, Lochner is never a great recommendation for any construction of the Due Process Clause. And that’s doubly true here. The Lochner-era Court’s protection of a right to present “all the facts” was intertwined with the notion, deeply anathema to the post-New Deal due process framework, that property rights are fundamental and therefore deserve the type of heightened procedural safeguards that today we reserve for certain kinds of

383 See supra note 4 and accompanying text.
384 See supra Part III.
385 See supra Part III.A.
386 See supra Part III.B.
387 See supra note 4 and accompanying text.
important liberty interests.\textsuperscript{388} \textit{Lochner}-era cases like \textit{Turnipseed} or \textit{Heiner} are, to put it bluntly, relics of assumptions whose rejection lies at the foundation of post-New Deal constitutionalism.\textsuperscript{389} It’s not surprising that post-New Deal courts have turned their back on them.

But \textit{Lochner} is not just an inconvenient pedigree as a matter of recent precedent. The history developed here undermines the claims of class action defendants who want to invoke \textit{originalism} as a trump against recent precedent.

We’ve seen that historical practice can be useful to originalists in three ways. Some think that a long tradition of early historical construction of open-ended constitutional clauses, like the Due Process Clauses, is powerful evidence of their proper construction because the early construction reveals the kind of principle that the people who ratified the clause most likely intended to enact.\textsuperscript{390} Others think a long tradition of early historical practice liquidates ambiguity in a clause’s text by “fixing” the proper meaning of the clause.\textsuperscript{391} Still others think that while vague and abstract clauses delegate a degree of interpretive discretion to judges, long traditions are presumptive evidence of wise constructions.\textsuperscript{392}

Yet under any of these approaches to constitutional construction, the \textit{Lochner} era view that civil defendants have a right to present “all the facts” has a weak claim to guide us. Its approach is a departure from the historical understanding of the Due Process Clause that grew in the wake of the Fifth Amendment’s ratification and persisted for much of the first century of American due process jurisprudence.\textsuperscript{393} As a result, the \textit{Lochner}-era understanding is not very good evidence of the principles the Fifth Amendment’s framers thought they were enacting. Nor is it compelling for those who think, consistent with common theories of constitutional construction held by the Framers, that early practice fixes the construction of ambiguous clauses. Either way, the older, century-long tradition that \textit{Lochner} supplanted has a much stronger claim to guide construction.

The long chain of antebellum cases also raises some doubts about the \textit{Lochner}-era understanding as evidence of the meaning of the \textit{Fourteenth} Amendment’s Due Process Clause. After all, given the early tradition of interpretation—which persisted up until the ratification of the Civil Rights Amendments—there’s no compelling reason to think many Reconstruction era lawyers would have thought due process protects a right to present all the facts in civil cases. True, some of the conceptual shifts (\textit{e.g.}, changes in the way lawyers conceived of the relationship between rights and remedies and changing notions about the value of a hearing) that lead to \textit{Lochner}-era cases like \textit{Turnipseed} have their first glimmerings in the years around the Reconstruction era. But those shifts ripened and bore doctrinal fruit a generation later. Based on the state of the law at

\begin{itemize}
  \item \textsuperscript{388} See \textit{supra} Part IV.A.
  \item \textsuperscript{389} See \textit{supra} Part IV.A.
  \item \textsuperscript{391} See \textit{supra} Part II.B.
  \item \textsuperscript{392} See \textit{supra} Part II.B.
  \item \textsuperscript{393} See \textit{supra} Part III, IV.
\end{itemize}
the time the Fourteenth Amendment was ratified, many Reconstruction lawyers were instead more likely to have thought as we do: that there is no vested constitutional right to rules of evidence.\(^{394}\) For them, as for us, those rules were committed to reasoned judicial discretion in light of the spirit and guiding principles of the common law of evidence—accuracy, convenience, utility, and public policy.

Even if we concede the meaning of the Fourteenth Amendment’s Due Process Clause is, like the Fifth’s, ambiguous and calls for construction, the history presented in this Article matters. If, as many originalists think, long traditions have the best claim to guide construction of due process because the test of time is evidence of wise constructions, then the very long, robust and irrepressible practice of tolerating judicial alteration of opportunities to present all probative evidence in civil cases, in the service of convenience, utility, and public policy, has a powerful claim to guide construction of both the Fifth and Fourteenth Amendment’s Due Process Clauses.

The history of the Due Process Clause in effect confirms our common wisdom. Selecting fair procedures requires a hazardous and uncertain trade-off between “risk reduction and available resources, just as any substantive right against risk imposition does,”\(^{395}\) and it’s far from clear that that balance always favors a full-throated adversarial proceeding that lets in “all the facts.” Indeed, it’s not clear adversarial proceedings are best even if accuracy were the only measure of fair procedure: “Today,” as Robert Bone says, “we are acutely aware of the limitations of trial-like procedures [as a truth-seeking mechanism], especially in the strategic environment of litigation.”\(^{396}\) Given that selecting “fair” procedures is fraught with trade-offs and uncertainty, constitutionalizing the features of an ideal, full-throated adversarial hearing in civil cases without regard to experience and context, including cost, is unwise.\(^{397}\) Our traditions, which—the Lochner era aside—treat many choices in the realm of civil procedure and evidence as a subconstitutional matter left to judicial or legislative choice, bear that judgment out.

This is not to say that due process doesn’t include some, albeit uncontroversial, fixed content. The original understanding requires a deprivation of core private rights through a judicial hearing. Since 1791, the core hearing right has been understood to require (1) some opportunity to participate—personally or through an adequate representative—in the judicial proceeding that affects your rights, (2) fair notice, (3) an unbiased judge exercising independent judgment about adjudicative facts and unsettled parts of the law, and (4) a determination of

\(^{394}\) COOLEY, supra note 133, at 367.


\(^{396}\) Id. at 1016.

\(^{397}\) Id. at 1017 (“[F]ew people, if any, would think that reducing the risk of error is always important enough to justify substantial social investments . . . .”).
liability according to governing rules of law and procedures commanded by a legislature.

But to the extent that the legislature has given courts authority to alter and modify preexisting rules governing the weight and sufficiency of evidence, it doesn’t limit their ability on to ration opportunities for proof based on a reasoned trade-off between accuracy and efficiency.

This is not to say that class action defendants’ arguments might not have some merit on other grounds. Even if due process does not require federal courts to provide class defendants an opportunity to present individualized evidence, Congress might. But, what if it’s not clear what Congress wants? Should federal courts, take it on themselves to adjust parties’ opportunities to prove their claims or defenses in the service of efficiency or deterrence? Or should they wait for further, clearer legislative guidance? Lurking behind class defendants’ due process arguments is a difficult question about the exact responsibility that Congress ought to have for the choice about how class claims can be proven.

That’s a question for another article. But, as this Article should make clear, it’s an important question precisely because the Constitution doesn’t specify the exact level of opportunities parties should have to prove their civil claims and defenses. The Constitution leaves the answer to that question up to us.
A MINIMALIST APPROACH TO SAME-SEX DIVORCE:
RESPECTING STATES THAT PERMIT SAME-SEX MARRIAGES
AND STATES THAT REFUSE TO RECOGNIZE THEM

Robert E. Rains*

"Thus Grief still treads upon the Heels of Pleasure: Married in haste,
we may repent at leisure."
—William Congreve, The Old Batchelour, Act V, Scene 1 (1693)

I. INTRODUCTION

Unlike most modern countries, the United States has no general law of
domestic relations. The powers delegated in the Constitution to the Congress do
not include the governance of family law. Moreover, the Bill of Rights provides
that “[t]he powers not delegated to the United States by the Constitution, nor
prohibited by it to the States, are reserved to the States respectively, or to the
people.” Thus, in 1890, the US Supreme Court unequivocally stated, “[t]he whole
subject of the domestic relations of husband and wife, parent and child, belongs to
the laws of the States and not to the laws of the United States.” In the ensuing 120
years, Congress has, directly and indirectly, addressed multiple family law issues
utilizing its various delegated powers. But it remains true that there is no federal
law of marriage or divorce. Each of the fifty states has its own marriage and
divorce laws, and they are often in sharp conflict with each other. For example,
until the Supreme Court ruled such laws unconstitutional in Loving v. Virginia in
1967, sixteen states still prohibited interracial couples from getting married, while
thirty-four states authorized such unions.

* 2012 Robert E. Rains. Professor of Law and Co-Director of the Family Law Clinic,
The Pennsylvania State University Dickinson School of Law, Carlisle, PA USA. The
author wishes to thank his research assistant, Ujala Aftab, for her contributions to this
project.

1 See U.S. CONST. art. I, § 8, cl. 1.
2 U.S. CONST. amend. X.
4 See, e.g., infra text accompanying notes 15–16.
5 There have been efforts over the years to create uniform marriage and divorce laws
on the state level, but they have met with little success. The National Conference of
Commissioners on Uniform State Laws issued a proposed Uniform Marriage and Divorce
Act in 1970, but, to date, only eight states have adopted some form of that act. See UNIF.
6 388 U.S. 1 (1967)
7 Loving, 388 U.S. at 6. Had Barack Obama’s parents attempted to get married in
1961 in Virginia, or to have even lived in Virginia as a married couple, they would have
been subject to criminal prosecution as were the Lovings.

393
While the once heated debate over interracial marriage is today probably viewed in most circles as an embarrassing vestige of the era of “Jim Crow,” basic disagreements continue among the states as to who can marry whom. American states are fairly equally divided as to whether first cousins may marry.\(^8\)

But, of course, the current marriage issue that most animates vitriolic political dispute in the United States and elsewhere is the question of same-sex couples. This issue first came to the fore in the United States with the 1993 decision of the Hawaii Supreme Court in *Baehr v. Lewin*,\(^9\) in which that court ruled that several same-sex couples had stated a cause of action that Hawaii’s prohibition on same-sex marriage arguably violated the Hawaii State Constitution.\(^10\) This decision, which only called for a remand of the case, created a public firestorm. At the federal level, Congress enacted the “Defense of Marriage Act” (“DOMA”).\(^11\)

The federal DOMA has but two substantive provisions. One provision is that the United States government will not recognize a same-sex marriage for any federal purpose.\(^12\) The other provision addresses interstate concerns, specifically recognition by one state of a same-sex marriage legally performed in another state:

---


\(^9\) 852 P.2d 44 (Haw. 1993).

\(^10\) Id. at 52–54. *Baehr* was a sharp departure from prior state court decisions on the subject, all of which had rejected the concept of a right to same-sex marriage. See *Jones v. Hallahan*, 501 S.W.2d 588 (Ky. 1973) (asserting “marriage has always been considered as the union of a man and a woman”); *Baker v. Nelson*, 191 N.W.2d 185 (Minn. 1971) (holding Minnesota law “does not authorize marriage between persons of the same sex and that such marriages are . . . prohibited” and dismissing the appeal for “want of a substantial federal question”), 409 U.S. 810 (1972); *Anonymous v. Anonymous*, 325 N.Y.S.2d 499 (N.Y. Sup. Ct. 1971) (“The law makes no provision for ‘marriage’ between persons of the same sex.”); *DeSanto v. Barnsley*, 476 A.2d 952 (Pa. Super. Ct. 1984) (holding, “as a matter of law,” two persons of the same sex cannot contract a common law marriage).


\(^12\) 1 U.S.C. § 7 (2006). This provision has been and continues to be challenged. Most notably, on July 8, 2010, federal district Judge Joseph L. Tauro issued two companion decisions striking down this section. In *Massachusetts v. U.S. Department of Health and Human Services*, he ruled that this section violated both the Tenth Amendment and the Spending Clause of the U.S. Constitution. 698 F. Supp. 2d 234, 249, 253 (D. Mass. 2010). In *Gill v. Office of Personnel Management*, he further ruled that it violated “the equal protection principles embodied in the Fifth Amendment.” 699 F. Supp. 2d 374, 397 (D. Mass. 2010). The United States appealed both decisions to the First Circuit Court of Appeals, but the status of those appeals became murky when Attorney General Eric Holder announced on Feb. 23, 2011, that President Obama has concluded that this provision is unconstitutional and, “[g]iven that conclusion, the President has instructed the Department
No State, territory, or possession of the United States, or Indian tribe, shall be required to give effect to any public act, record, or judicial proceeding of any other State, territory possession, or tribe respecting a relationship between persons of the same sex that is treated as a marriage under the laws of such other State, territory, possession, or tribe, or a right or claim arising from such relationship.\textsuperscript{13}

This provision carves out an exception to the general American rule that a marriage validly entered into in one state will be recognized in all other states.\textsuperscript{14} There is a constitutional, as well as a common law, basis for this rule, as the Full Faith and Credit Clause provides: “Full Faith and Credit shall be given in each State to the public Acts, Records, and judicial Proceedings of every other State.”\textsuperscript{15} However, in enacting DOMA, Congress purported to rely on its enforcement power under the Full Faith and Credit Clause: “And the Congress may by general Laws prescribe the Manner in which such Acts, Records and Proceedings shall be proved, and the Effect thereof.”\textsuperscript{16} Whether this second provision of DOMA is a proper exercise of that enforcement power, or a violation of it, is a hotly debated question.\textsuperscript{17}

This Article will address the legal conundrum that arises when a person who validly entered a same-sex marriage in one state seeks a divorce in another state that refuses to recognize same-sex marriage. The Article will first discuss the interstate recognition of marriages and divorces in general, then the patchwork quilt of same-sex marriage laws in the United States, followed by a discussion on seeking a legal exit from a same-sex marriage in a state that does not recognize that marriage, and finally, suggest a path which will allow a court in the latter state not to defend the statute in such cases. I fully concur with the President’s determination.” See Statement of the Attorney General on Litigation Involving the Defense of Marriage Act, U.S. DEP’T JUST. (Feb. 23, 2011), http://www.justice.gov/opa/pr/2011/February/11-ag-222.html. A group of Republican leaders of the House of Representatives—the Bipartisan Legal Advisory Group—retained counsel and intervened to defend this section of DOMA. On May 31, 2012, a panel of the First Circuit Court of Appeals unanimously affirmed the district court in a consolidated opinion. Comm’n of Massachusetts v. U. S. Dep’t of Health and Human Servs., No. 10-2204, 2012 WL 1948017 (1st Cir. May 31, 2012).\textsuperscript{13} 28 U.S.C. § 1738C.\textsuperscript{14} See Peter Hay, Recognition of Same-Sex Legal Relationships in the United States, 54 AM. J. COMP. L. 257, 261 (2006).\textsuperscript{15} U.S. CONST. art. IV, § 1.\textsuperscript{16} Id.\textsuperscript{17} Compare Lynn D. Wardle, Non-recognition of Same-Sex Marriage Judgments under DOMA and the Constitution, 38 CREIGHTON L. REV. 365 (2005) (arguing Congress was acting within its authority under the Full Faith and Credit Clause when it enacted DOMA), with Hay, supra note 14, at 261 (arguing a “fine distinction” may be the key to DOMA’s constitutionality). A federal district court upheld this provision of DOMA against multiple constitutional challenges in Wilson v. Ake, 354 F. Supp. 2d 1298, 1303–04, 1309 (M.D. Fla. 2005).
to grant relief without violating the letter or spirit of state provisions barring recognition of same-sex marriage.

II. INTERSTATE RECOGNITION OF MARRIAGES (AND DIVORCES) IN GENERAL

The validity of a marriage entered into in another state is a matter that is litigated with some frequency, but not usually on a constitutional basis. Different courts have developed different frameworks for addressing this issue.

Perhaps the best-known case is the 1953 decision of the New York State Court of Appeals in In re May’s Estate.18 In that case, a Jewish uncle and niece were barred from marrying in New York.19 They traveled to Rhode Island, which generally prohibited such marriages but allowed them for persons of the Jewish faith.20 The marriage lasted thirty-two years until the wife’s death, and produced six children.21 An estate battle ensued between the widower and three of the children who claimed that their parents’ marriage was invalid under New York law, and therefore they were next of kin to their deceased mother.22 The surrogate court (that is, the trial court) agreed with the three children that their parents’ marriage was void because it was “opposed to natural law” and contrary to New York statutory law.23

The May’s Estate decision presents a number of interesting aspects. First, the court never addressed the constitutionality of allowing a marriage of persons of one faith where the same marriage would be declared void if the parties were of another faith. Could a couple convert from one religion to another to avoid a

---

19 See id. at 4–5.
20 Id. at 5.
21 Id.
22 Id.
23 Id.
24 Id. at 7.
25 Id. at 6.
26 Id.
27 Id. at 6–7.
28 Id. at 7.
marriage prohibition? If a couple such as the Mays lawfully married in one religion but later converted to another religion, would it affect their civil marriage? What if they were an interfaith couple? Could they choose which religion governed the validity of their marriage?

The second aspect of the May's Estate decision worth noting is the subjectivity involved in a civil court’s attempt to find and apply “natural law.” This was highlighted by the fact that a dissenting judge would have found that “[a]ll such misalliances are incestuous, and all, equally, are void.” 29 A court’s reliance on such an amorphous concept as natural law is akin to reliance on scripture, as often happens today in the battle over same-sex relationships. 30 Those resorting to such scriptural reliance would do well to recall that in 1959 the trial judge who sentenced the Lovings for their crime of inter-racial marriage found support from the Deity in doing so:

Almighty God created the races white, black, yellow, malay and red, and he placed them on separate continents. And but for the interference with his [sic] arrangement there would be no cause for such marriages. The fact that he [sic] separated the races shows that he [sic] did not intend for the races to mix. 31

It appears that none of the nine justices of the US Supreme Court who reversed the Lovings’ convictions shared the trial judge’s views of a mandate from the Lord nor did they fear divine retribution. 32

The third interesting lesson from May's Estate is that context is often critical in marriage recognition cases. The Mays’ marriage was of long duration, lasting over three decades, and happy enough to produce six children. 33 There is no indication that the couple ever separated, or that they doubted the validity of their union. 34 Their marriage was not attacked by either of them, but rather by three of their children who were apparently motivated by greed over their mother’s estate. 35

29 Id. at 9 (Desmond, J., dissenting).
32 Id. at 12 (holding that marriage restrictions based on race violate Equal Protection and Due Process); see also id. at 13 (Stewart, J., concurring) (“It is simply not possible for a state law to be valid . . . which makes the criminality of an act dependent upon the race of the actor.”).
33 In re May’s Estate, 114 N.E.2d at 5.
34 See id.
35 See id.
Might the result have been different if, shortly after their wedding in Rhode Island and return to New York, the bride had “come to her senses,” left her uncle/husband, and sued for an annulment? There is, of course, no way of knowing, but it seems far more likely that the New York courts would have declared the marriage void under those circumstances.

Often, courts invoke the notion of “comity” to validate an out of state marriage. Thus, in *Hesington v. Estate of Hesington*, the Missouri Court of Appeals opined: “However, as a matter of comity, Missouri will recognize a marriage valid where contracted unless to do so would violate the public policy of this state.” Note, however, that comity (giving deference to a foreign judgment, decree, etc.) is a lesser mandate than the constitutional mandate of full faith and credit—which itself is not absolute. In *Hesington*, a Missouri woman wished to establish that she was the widow of a deceased Missouri man by virtue of a common law marriage they had entered into in Oklahoma in 1978. At the time of the common law marriage ceremony, Oklahoma permitted common law marriages, but Missouri had abolished such marriages in 1921. The Missouri trial judge found that had the couple been Oklahoma residents, they would have met Oklahoma’s requirements for a common law marriage. Nevertheless, the trial judge ruled that the couple’s Oklahoma marriage was invalid in Missouri, and the appellate court affirmed. The appellate court noted with approval the Restatement (Second) of Conflict of Laws § 283(1) (1971): “The validity of a marriage will be determined by the local law of the state which, with respect to the particular issue, has the most significant relationship to the spouses and the marriage . . . .”

Nevertheless, applying the principle of comity, the court indicated that “Missouri will recognize a marriage valid where contracted unless to do so would violate the public policy” of Missouri. The court noted that while other states are split on the subject, the majority view is that a state that does not permit common law marriages will not recognize a common law marriage of its residents when the

---

36 640 S.W.2d 824, 826–27 (Mo. Ct. App. 1982).
37 *Id.* at 826.
38 Compare *id.* (“[A]sa matter of comity, Missouri will recognize a marriage valid where contracted unless to do so would violate the public policy of this state.”), with Williams v. North Carolina, 317 U.S. 287, 296 (1942) (“[E]ven though [a] cause of action could not be entertained in the state of the forum, either because it had been barred by the local statute of limitations or contravened local policy, the judgment thereon obtained in a sister state is entitled to full faith and credit . . . [and although ]ome exceptions have been grafted on the rule . . . [they] have been few and far between . . . .”).
39 *Hesington*, 640 S.W.2d at 824.
40 *Id.* at 825.
41 *Id.* at 824–25.
42 *Id.* at 824–25, 827.
43 *Id.* at 826.
44 *Id.*
common law marriage took place during a temporary sojourn to a state that permits such marriages.\footnote{Id. (collecting cases).}

The court found that when the Missouri legislature abolished common law marriage, it had as a purpose “to require some degree of solemnity and reliability in establishing a marriage of those domiciled in and residing in Missouri.”\footnote{See id. at 827.} Recognizing the Oklahoma common law marriage of the Missouri residents in this case would violate that public policy.\footnote{Id. (noting that the statute provided the “highest evidence” of the state’s public policy regarding common law marriage).}

On a strictly logical basis, it is hard to square the result in \textit{Hesington} with the result in \textit{May’s Estate}. Both involved marriages that were lawful where contracted. But the couple in \textit{May’s Estate} could not under any circumstances have married in their state of residence because of their consanguinity. There is no suggestion in \textit{Hesington} that there was any bar whatsoever to the Hesingtons’ marriage in their home state; they simply entered into their marriage in a less formal fashion than their home state allowed. In other words, their error only went to the “formalities” of marriage, not the essentials. Therefore, from a logical standpoint, Mr. Hesington’s widow had a stronger claim than Mrs. May’s widower.

Other than different courts addressing different cases at different times, the only reasonable explanation for the contradictory results is, again, context. The Mays were married for thirty-two years and produced six children. A ruling that their marriage was void would have almost certainly rendered those children “illegitimate” at a time when illegitimacy not only carried a great social stigma, but also far greater legal disadvantages than it does today.\footnote{Lili Mostofi, \textit{Legitimizing the Bastard: The Supreme Court’s Treatment of the Illegitimate Child}, 14 J. CONTEMP. LEGAL ISSUES 453 (2004) (describing the extension of rights to illegitimate children over the course of approximately forty years). But see Solangel Maldonado, \textit{Illegitimate Harm: Law, Stigma, and Discrimination Against Nonmarital Children}, 63 FLA. L. REV. 345 (2011) (arguing that social stigma and discrimination against illegitimate issue continue).} (It is indeed ironic that three of the Mays’ children were effectively arguing in court for their own illegitimacy.)

By contrast, the Hesingtons entered into their purported common law marriage less than two years before Mr. Hesington’s death.\footnote{\textit{Hesington}, 640 S.W.2d at 824.} There is no indication that their union was blessed with issue, hence there were apparently no children who would be deemed illegitimate by virtue of the Missouri court’s ruling.

A case applying yet another approach to marriage recognition is the 1984 Washington Court of Appeals decision, \textit{In re Estate of Shippy}.\footnote{678 P.2d 848 (Wash. Ct. App. 1984).} This case actually involved the laws of three states: Washington, California, and Alaska.\footnote{See id. at 849.} James Shippy executed a will in January 1972, leaving his estate to his then wife,
Marion. In January 1973, Marion obtained an interlocutory decree of divorce from James in California. James married Inge in Alaska in 1978 although his divorce from Marion was not final. James died in a plane crash in Alaska on July 15, 1981. On November 16, 1981, four months after James’ death, the California court entered the final decree *nunc pro tunc* divorcing James and Marion as of May 14, 1973. In the subsequent estate battle in Washington, the trial court found that Inge was not James’ surviving spouse because her marriage to James was void under Alaska law. The issue presented was which state’s law would control regarding the retroactive effect of a *nunc pro tunc* decree on an intervening second marriage. Under the majority view, including the law of Washington state, the later *nunc pro tunc* decree would validate the intervening marriage. Some states took the contrary position. Although Alaska courts had not addressed the issue, Alaska statutory law provided that, “[a] subsequent marriage contracted by a person during the life of a former husband or wife which marriage has not been annulled or dissolved is void.” Hence the Washington Court of Appeals concluded that if it applied Alaska law, James and Inge’s marriage would appear to be void. This would have defeated Inge’s claim because the counterpart of the general rule that a marriage validly entered into is valid everywhere is that a marriage invalidly entered into is invalid everywhere.

The court went on, however, to apply a choice of law approach that appears to be the polar opposite of those used in May’s *Estate* and Hesington. Relying on the Restatement (Second) of Conflict of Laws § 283 comment i (1971), it reasoned that the Alaska marriage would not be deemed invalid in Washington unless:

> [t]he intensity of the interest of the state where the marriage was contracted in having its invalidating rule applied outweighs the policy of protecting the expectations of the parties by upholding the marriage and

---

52 Id.
53 Id.
54 See id.
55 Id.
56 Id. The opinion does not explain the eight-year delay from the interlocutory decree to the final decree nor indicate who, if anyone, asked the California court to issue the final decree.
57 Id. at 849–50.
58 Id. at 850.
59 Id.
60 Id. (collecting cases).
61 Id. (citing ALASKA STAT. § 09.55.-080 (1982)).
62 Id. The court noted that Alaska did not have a provision for common law marriage, and that under Alaska statutory law, a “[m]arriage is prohibited and void if performed when (1) either party to the proposed marriage has a husband or wife living.” Id. (citing ALASKA STAT. § 25.05.021).
63 See Farah v. Farah, 429 S.E.2d 626, 629 (Va. Ct. App. 1993) (“A marriage that is void where it was celebrated is void everywhere.”).
the interest of the other state with the validating rule in having this rule applied.\textsuperscript{64}

Although the court was unable to determine James and Inge’s state of residence at the time of their Alaska marriage, it found that Washington had a substantial relationship to the parties because they resided in Washington when James died, property existed in Washington to be distributed, and probate proceedings were pending in Washington.\textsuperscript{65} Thus, Washington law would apply unless Alaska had a clearly contrary policy.\textsuperscript{66} Alaska law, by itself, did not establish such a policy.\textsuperscript{67} Indeed Washington had a similar statute, but its courts would still recognize such a marriage.\textsuperscript{68} Thus, “to protect the expectations of James and Inge,” the court applied Washington law and validated their Alaska marriage.\textsuperscript{69}

The \textit{Shippy} decision raises as many issues as it answers. The Shippys’ marriage was longer (five years)\textsuperscript{70} than the Hesingtons’ (two years), but considerably shorter than the Mays’ (thirty-two years). While the court explicitly concerned itself with James and Inge’s expectations, it did so at the expense of James’ children (who may or may not have been the product of his marriage to Marion).\textsuperscript{71} The most reasonable explanation is that James and Inge were unaware that his divorce from Marion had not been finalized. Although ignorance of the law is generally no excuse, the court simply chose to protect Inge if she was unfamiliar with the difference between a California interlocutory divorce decree and final divorce decree. Indeed, it is probable that James told her—and actually believed—that he was divorced from Marion. It appears that it was his intention to divorce Marion and, later, to marry Inge. Viewed this way, his error might, or might not, be deemed to have gone to the formalities—as opposed to the essentials—of marriage.

Some state legislatures have sought to proactively bar their residents who cannot marry in their state of residency from getting married in another jurisdiction. For example, Wisconsin enacted a law in 1971 to prevent Wisconsin “deadbeat dads” from marrying.\textsuperscript{72} The law generally barred parents who were in

\textsuperscript{64} \textit{Shippy}, 678 P.2d at 851 (quoting \textsc{Restatement (Second) of Conflict of Laws} § 283 cmt. i (1971)).

\textsuperscript{65} \textit{Id.}

\textsuperscript{66} \textit{Id.}

\textsuperscript{67} \textit{Id.}

\textsuperscript{68} \textit{Id.} at 851–52 (citing \textit{In re Estate of Storer}, 544 P.2d 95 (Wash. Ct. App. 1975)).

\textsuperscript{69} \textit{Id.} at 852.

\textsuperscript{70} \textit{Id.} at 849.

\textsuperscript{71} See \textit{id.} at 849, 852. James had two children, Dorothy Coe and Thomas Shippy. It is unclear whether they were the product of James’ marriage to Marion.

arrears in paying child support from marrying. 73 It specifically addressed out-of-state marriages:

   This section shall have extraterritorial effect outside the state; and s. 245.04(1) and (2) [providing that out-of-state marriages to circumvent Wisconsin law are void] are applicable hereto. Any marriage contracted without compliance with this section, where such compliance is required, shall be void, whether entered into in this state or elsewhere. 74

The United States Supreme Court struck down the law in its entirety, finding that it violated the Equal Protection Clause of the Constitution, and hence had no occasion to address the constitutionality of its extraterritorial provision. 75

Attacking this issue from the opposite perspective, in 1912 the National Conference of Commissioners on Uniform State Laws proposed the confusingly named “Uniform Marriage Evasion Act.” 76 Under that statute, a state would not permit a marriage to take place within its borders if it was between nonresidents who were forbidden to marry in their home state. (Hence the statute should have been called the Uniform Marriage Prohibition Evasion Act. It was intended to prevent certain people from evading marriage prohibitions in their home states, rather than evading marriage.) Most states already had some form of marriage evasion act. 77 The proposed uniform act was only adopted in five states, and the Uniform Law Commissioners withdrew it in 1943. 78 However, withdrawal by the commissioners of a uniform law does not repeal that law in any state that has already adopted it. 79 Only a state’s legislature can repeal a law (or that state’s courts may strike it down). Indeed, over six decades after the Uniform Marriage Evasion Act was withdrawn, the Massachusetts Supreme Judicial Court applied Massachusetts’ version of that act to bar same-sex couples from Connecticut, Maine, New Hampshire, and Vermont from getting married in Massachusetts. 80

73 See Wis. Stat. §§ 245.10(1)–(3).
74 Zablocki, 434 U.S. at 375 n.1 (alterations in original) (quoting Wis. Stat. § 245.10(5)).
75 Id. at 377.
78 See id.
79 See Cote-Whitacre, 844 N.E.2d 632 (Spina, J., concurring).
The US Supreme Court has had limited opportunity to address interstate marriage recognition. In 1888, in *Maynard v. Hill*, the Court upheld a ruling of the Supreme Court of the Territory of Washington that a decedent was married to his second wife at the time of his death. David Maynard had married Lydia Maynard in Vermont in 1828 and had two children by her. In 1852, allegedly with no notice to Lydia, David obtained a legislative divorce from her. Shortly thereafter, David married Catherine, with whom he lived until his death. In the ensuing estate battle, Lydia’s children asserted that Lydia was still legally married to David when he made a “donation claim” to certain land after the legislative divorce. Lydia’s children raised various due process objections to the legislative divorce, all of which were ultimately rejected. The Court did not directly address any interstate conflict of laws issues in *Maynard*.

In 1907, in *Travers v. Reinhardt*, the US Supreme Court reviewed a decision of the Court of Appeals of the District of Columbia addressing the marriage of a man from Washington, D.C. to a woman from West Virginia. The marriage took place in Virginia, but was defective there because of the lack of a proper minister; however, it was arguably ratified as a common law marriage in New Jersey during short stays there. The US Supreme Court affirmed the District of Columbia court’s finding that the parties had been validly common law married in New Jersey. As in *Maynard*, the Court did not directly address the standards for interstate marriage recognition.

In *Williams v. North Carolina*, a case that went to the US Supreme Court twice, the Court did address, under the Full Faith and Credit Clause, North Carolina’s refusal to recognize the marriage of two North Carolinians in Nevada. However, the validity of their marriage hinged on the recognition of the parties’ divorce decrees, which were issued by the State of Nevada and purported to dissolve the parties’ prior marriages to their respective spouses who remained in North Carolina.

---

81 125 U.S. 190 (1888).
82 See id. at 193, 215–16.
83 Id. at 191–92.
84 Id. at 192–93.
85 Id. at 193.
86 Id. at 193–95.
87 Id. at 193, 196, 216.
88 205 U.S. 423 (1907).
89 Id. at 432–33.
90 Id. at 433–34, 438–39. The parties also had lived in Maryland during the marriage.
91 Id. at 438.
93 *Williams I*, 317 U.S. at 289–92; *Williams II*, 325 U.S. at 234, 236.
Briefly, O.B. Williams married Carrie Wyke in 1916 in North Carolina and lived with her there until 1940. Lillie Shaver Hendrix married Thomas Hendrix in 1920 in North Carolina and lived with him there until 1940. In May 1940, O.B. and Lillie travelled to Las Vegas, Nevada (“Sin City,” then as now), where each filed for divorce in June 1940. Neither of their spouses was personally served in Nevada, although each apparently received notice of the proceedings. Neither entered an appearance or participated in any way in either divorce action. The Nevada court granted O.B. a divorce on August 26, 1940, and Lillie a divorce on October 4, 1940. Not letting the grass grow under their feet, O.B. and Lillie got married that same day in Nevada. Presumably, if they had remained in Nevada, they could have lived there together legally ever after.

But O.B. and Lillie returned to North Carolina, where they were tried, convicted, and sentenced to imprisonment for the crime of bigamous cohabitation. The North Carolina courts ruled that North Carolina was not required to recognize their Nevada divorce decrees under the Full Faith and Credit Clause.

The first time that the *Williams* case went to the US Supreme Court, in 1942, the Court presumed that the newlyweds had met Nevada’s domiciliary requirements for a divorce. Overturning past precedent, the Court ruled that a state is empowered to enter a divorce decree that is entitled to full faith and credit in all other states, as long as one of the spouses is domiciled in that state and provides “substituted service” on the other party that meets the requirements of due process. In other words, a state court—applying its own state divorce laws—can grant a divorce that is binding on both parties even when the marriage was entered into in another state, their entire married life took place in another state, and the defendant spouse has never set foot in the state issuing the divorce, was not served in that state and did not participate in the divorce action—as long as the defendant spouse has received “substituted service.” Finally, in *Williams I*, the Court remanded the case to the courts of North Carolina for further proceedings.

O.B. and Lillie were retried before a jury of their peers in North Carolina. The trial judge instructed the jury that O.B. and Lillie had the burden to

---

96 Id.
97 Id.
98 *Williams I*, 317 U.S. at 289–90.
99 Id.
100 Id. at 290.
101 Id.
102 Id. at 289–90.
103 Id. at 291.
104 Id. at 292–93, 302.
105 Id. at 299, 302–04.
106 Id. at 304.
107 *Williams II*, 325 U.S. at 233–36.
demonstrate that they were domiciled in Nevada at the time they obtained their divorces, and that the Nevada court’s recitation of *bona fide domicil* in their divorce decrees was “prima facie evidence,” but did not compel “such an inference.” If they had only gone to Nevada to get their divorces, intending to return to North Carolina on obtaining them, then they neither lost their North Carolina domicil nor acquiesced new domicils in Nevada. The jury duly convicted O.B. and Lillie again of bigamous cohabitation, and that conviction was upheld through the North Carolina courts.

On appeal to the US Supreme Court the second time, the critical issue was whether the North Carolina courts had failed to give full faith and credit to the Nevada divorce decrees, specifically insofar as those decrees found that O.B. and Lillie had bona fide domicil in Nevada. In *Williams II*, the Court ruled that, although the “fact that the Nevada court found that they were domiciled there is entitled to respect, and more,” the North Carolina courts were not bound by that finding. North Carolina was free to reexamine this issue and had done so, giving appropriate weight to the Nevada court’s findings. Concluding that North Carolina had not violated the full faith and credit clause, the Supreme Court affirmed the convictions for bigamous cohabitation.

The *Williams I* and *Williams II* decisions—made during the era of “migratory divorce,” when the unhappily married frequently left their spouse and home state to find a more conducive jurisdiction and congenial life partner—remain the law in the United States today. The result of those decisions for the individual litigants (O.B. and Lillie) was a truly anomalous situation. As far as Nevada was concerned, they were divorced from their original spouses and lawfully married to each other. As far as North Carolina was concerned, they were each married to their original spouses and it was criminal for them to hold themselves out as married to each other. The Supreme Court rather blithely acknowledged that if one state can review the validity of a divorce, and hence a remarriage, in another state, then “persons may, no doubt, place themselves in situations that create unhappy consequences for them.” And that is precisely the situation faced today by certain people who have entered into a same-sex marriage in one state that they have tried to lawfully exit in another state.

Two more Supreme Court full faith and credit cases in the domestic relations arena warrant brief discussion. The Court refined the *Williams I* doctrine in two subsequent decisions, both of which, not coincidentally, involved departing spouses who sought their legal freedom in Nevada.

---

108 *Id.* at 235–36 N.B. In describing *Williams*, this Article uses the Court’s spelling of “domicil” in the case. Modern usage is “domicile.”

109 *Id.*

110 *Id.* at 227.

111 *Id.*

112 *Id.* at 233–34.

113 *Id.*

114 *Id.* at 239.

115 *Id.* at 237.
In the 1948 case *Estin v. Estin*, the Court addressed the situation of Mr. Estin, who was married in New York in 1937 and lived there with his wife until they separated in 1942. In 1943, his wife filed an action against him in New York for a legal separation, which the court granted, along with $180 per month as legal alimony that, under New York law as it then existed, would continue until the parties were divorced. Mr. Estin, like other unhappy spouses before and since, headed out to Nevada in 1944 and brought a divorce action in 1945 (thereby clearly meeting the domiciliary requirement). His wife was notified of the action (thereby meeting the due process requirement), but entered no appearance and did not participate. Mr. Estin duly informed the Nevada court of the New York separation and alimony decree; nevertheless, the Nevada court entered a divorce decree with no provision for alimony. So, of course, Mr. Estin stopped paying his now ex-wife. She, naturally, sued him in New York to compel continued payments. He appeared in that action and moved to eliminate the New York alimony order on the basis of his Nevada divorce decree, but the New York courts ruled that the Nevada decree did not extinguish his ex-wife’s right to alimony under the earlier New York decree.

On appeal, the Supreme Court created the doctrine of “divisible” divorce, ruling that the Nevada decree was entitled to full faith and credit to the extent that it changed the marital status of the parties, but not insofar as it purported to change the “legal incidence of the marriage,” in other words, the alimony order. Because the alimony order was a property interest of the wife, Nevada could not affect that interest without personal jurisdiction over her, which it lacked.

A decade later, the Court refined the *Estin* divisible divorce doctrine in *Vanderbilt v. Vanderbilt*. The Vanderbilts were married in 1948 and lived in California, where they separated in 1952. She moved to New York, and he went to Nevada where he obtained a divorce decree in 1953, freeing both parties “from the bonds of matrimony and all the duties and obligations thereof.” Mrs. Vanderbilt received notice of the Nevada action but was not served in Nevada and did not participate. In 1954, the former Mrs. Vanderbilt filed suit in New York for a

---

116 334 U.S. 541, 541 (1948).
117 Id. at 542.
118 Id. at 542–43.
119 Id.
120 Id. at 543.
121 Id.
122 Id.
123 Id.
124 Id.
125 Id. at 549.
126 Id. at 549.
128 Id. at 416.
129 Id.
Mr. Vanderbilt appeared specially in that proceeding and argued that the Full Faith and Credit Clause compelled New York “to treat the Nevada divorce as having ended the marriage and as having destroyed any duty of support which he owed . . . .” The New York court recognized the Nevada decree as terminating the status of the parties’ marriage, but found that it did not preclude New York from directing Mr. Vanderbilt to pay support, which it duly ordered.

On appeal, the Supreme Court affirmed, reasoning that the fact that Mrs. Vanderbilt’s right to support had not yet been reduced to judgment did not materially distinguish the case from *Estin*. Since Mrs. Vanderbilt had not been subject to personal jurisdiction in the Nevada court, that court could not terminate her right to support in an *ex parte* proceeding.

The 1967 “miscegenation” case of *Loving v. Virginia* also presented a potential interstate marriage recognition issue. Two Virginia residents, Mildred Jeter, described as a Negro woman, and Richard Loving, a white man, had been married in Washington, D.C., pursuant to its laws. Shortly after their marriage, they returned to Virginia where they were indicted, pled guilty to, and were sentenced to jail for violating the Virginia anti-miscegenation statute. While the case might have been litigated and decided under the Full Faith and Credit Clause, that issue was not presented to the Court, which found that the statute violated both the Equal Protection and Due Process Clauses of the Fourteenth Amendment.

III. THE PATCHWORK QUILT OF SAME-SEX MARRIAGE LAWS IN THE UNITED STATES

With regard to lawful recognition of same-sex couples, states generally fall into three main categories: 1) those that permit such couples to enter into marriage or a quasi-marriage relationship such as civil union or registered partnership; 2) those that do not permit same-sex couples to enter into legal marriage or marriage-type relationships but recognize such relationships if entered into elsewhere, at least for some purposes; and 3) those that prohibit and do not recognize same-sex marriages or quasi-marriage relationships.

At the time of this writing, six states permit same-sex couples to marry: Massachusetts (as of 2004), Connecticut (2008), Iowa (2009), Vermont

---

130 *Id.*
131 *Id.*
132 *Id.* at 417–18.
133 *Id.* at 418.
134 *Id.* at 418–19.
136 *Id.* at 2.
137 *Id.* at 2–3 N.B. The trial judge suspended the sentence for 25 years on condition that the Lovings leave Virginia and not return together during that period of time.
138 *Id.* at 10–12.
139 June 2012.
Three additional states have passed laws permitting same-sex marriage, which had not yet taken effect as of this writing. On February 8, 2012, the Washington State Legislature enacted a bill to allow same-sex couples to marry, which Governor Christine Gregoire signed on February 13, 2012. Opponents have stated that they will seek to block implementation through a referendum measure. New Jersey’s legislature passed a bill allowing same-sex marriage on February 16, 2012, which was quickly vetoed by New Jersey Governor Chris Christie. Soon afterward, Maryland enacted a law permitting same-sex marriages, which was signed by Maryland Governor Martin O’Malley and will go into effect (unless blocked) on January 1, 2013.

On Dec. 18, 2009, Washington, D.C. Mayor Adrian Fenty signed bill 18-482, which legalized same-sex marriage in the District of Columbia. It became effective March 2010, after Chief Justice John Roberts, acting as circuit justice for the District, refused to issue a stay. California permitted same-sex couples to enter into marriage for approximately six months in 2008, during which time it is reported that

140 In re Opinions of the Justices to the Senate, 802 N.E.2d 565 (Mass. 2004).
147 Yardley, supra note 146.
149 Kate Zernickie, Christie Keeps Promise to Veto Gay Marriage Bill, N.Y. TIMES, (Feb. 17, 2012), http://www.nytimes.com/2012/02/18/nyregion/christie-vetoes-gaymarriage-bill.html. Under state law, the New Jersey Assembly has two years, until January 2014, to override the veto. Id.
150 See Maryland Legislature Enacts Same-Sex Marriage Law, 38 FAM. L.REP. (BNA), no. 18, 2012, at 1227.
approximately 18,000 couples entered these unions. California voters approved Proposition 8 in November 2008, banning such marriages. The California Supreme Court subsequently upheld Proposition 8, but it also ruled that those same-sex marriages that had been lawfully entered into remained valid. A federal district court subsequently struck down California Proposition 8 as unconstitutional. On February 7, 2012, a divided panel of the Ninth Circuit Court of Appeals affirmed the district court. The Ninth Circuit denied rehearing en banc by order of June 5, 2012.

Several other states permit same-sex couples to enter into variously named forms of legally recognized quasi marriages. In the midst of the *Baehr v. Lewin* litigation, the Hawaii legislature enacted a law in 1997 allowing same-sex couples to become “reciprocal beneficiaries” with many of the “rights and benefits available only to married couples.” Similarly, Vermont created “civil unions” for same-sex couples in 1999 after its supreme court ruled that denying such couples the benefits of marriage violated the state constitution. Parties to a Vermont civil union were to have “all the same benefits, protections and responsibilities under law, whether they derive from statute, administrative or court rule, policy, common law or any other source of civil law, as are granted to spouses in marriage.” When Vermont amended its marriage law to permit same-sex couples to marry as of September 1, 2009, it also repealed the procedure for such couples to enter civil unions, while allowing existing civil unions to continue and allowing parties in civil unions to marry their civil union partners if they so choose. In 2004, New Jersey enacted its “Domestic Partnership Act,” permitting same-sex and opposite-sex couples to register as domestic partners and obtain some of the rights of married couples. In late 2006, New Jersey enacted a Civil Union Act, amending the 2004 Domestic Partnership Act. Under the Civil Union Act, two eligible individuals of the same sex can enter a civil union and “receive

---

156 *Strauss*, 207 P.3d at 48.
158 *Perry v. Brown*, 671 F.3d 1052 (9th Cir. 2012).
164 2003 N.J. Laws 246.
the same benefits and protections and be subject to the same responsibilities as spouses in a marriage.”

The latest state to create a statutory framework for same-sex (and opposite-sex) couples to enter into a civil union is Illinois. On January 31, 2011, Illinois Governor Pat Quinn signed legislation creating civil unions in that state, effective June 1, 2011.167 The Governor’s Office noted that California, Nevada, New Jersey, Oregon, Washington State, and Washington, D.C. all have civil union or similar laws on the books.168

Such state quasi-marriage laws have not been consistent as to the means to dissolve a civil union, domestic partnership, etc., but the trend has been to apply the same rules that apply to married couples. For example, under Washington State’s 2007 registered domestic partnership law, a member of a registered domestic partnership could exit that legal status by the simple expedient of filing a notice of termination and paying a filing fee.169 However, in 2009, the Washington State legislature amended the law to make those in registered domestic partnerships subject to the same rules as married people:

It is the intent of the legislature that for all purposes under state law, state registered domestic partners shall be treated the same as married spouses. Any privilege, immunity, right, benefit, or responsibility granted or imposed by statute, administrative or court rule, policy, common law or any other law to an individual because the individual is or was a spouse, or because the individual is or was an in-law in a specified way to another individual, is granted on equivalent terms, substantive and procedural, to an individual because the individual is or was in a state registered domestic partnership, or because the individual is or was, based on a state registered domestic partnership, related in a specified way to another individual. The provisions of [this act] shall be liberally construed to achieve equal treatment, to the extent not in conflict with federal law, of state registered domestic partners and married spouses.170

Oregon law places the same burden upon a party to a domestic partnership; that partnership will be treated like a marriage for purposes of dissolution:

An individual who has filed a Declaration of Domestic Partnership may not file a new Declaration of Domestic Partnership or enter a marriage with someone other than the individual’s registered partner unless a

---

168 Id.
judgment of dissolution or annulment of the most recent domestic partnership has been entered. This prohibition does not apply if the previous domestic partnership ended because one of the partners died.\textsuperscript{171}

New Jersey follows the same pattern with its civil unions:

The dissolution of civil unions shall follow the same procedures and be subject to the same substantive rights and obligations that are involved in the dissolution of marriage.\textsuperscript{172}

The point is that most states that permit same-sex couples to enter legally recognized, quasi-marital relationships, will normally require a disillusioned member of that couple to obtain a divorce in order to become legally free, just as if she were in a state that permits same-sex marriage by name.

There is another, smaller group of states that will not allow same-sex couples to marry or enter into quasi-marital relationships, but will recognize same-sex marriages validly entered into elsewhere, at least for certain purposes. New York State was a notable example before it authorized same-sex marriage in 2011.\textsuperscript{173} Similarly in May 2009, prior to allowing same-sex marriages to be performed there, the Washington D.C. Council had voted to recognize same-sex marriages from other jurisdictions.\textsuperscript{174} In February 2010, the Attorney General of Maryland issued a formal opinion that Maryland may recognize such marriages.\textsuperscript{175} In May 2010, the Maryland Department of Budget and Management announced that it was extending health benefits to the same-sex spouses of active and retired state employees who were validly married in another state.\textsuperscript{176} In January 2011, the Attorney General of New Mexico issued a formal opinion, not binding on New Mexico courts, that “a same-sex marriage that is valid under the laws of the

\textsuperscript{171} OR. REV. STAT. § 106.325(3) (2009).
\textsuperscript{173} See Godfrey v. Spano, 920 N.E.2d 328 (N.Y. 2009); C.M. v. C.C., 867 N.Y.S.2d 884 (2008) (holding that principles of comity permitted New York to recognize, and thus exercise jurisdiction over, a couple’s same-sex marriage in Massachusetts). Thus, a New York trial court was able in 2010 to grant a divorce to a same-sex couple who had entered into a civil union in Vermont without addressing the difficult issues that a state law prohibiting recognition of same-sex marriages would have presented. Parker v. Waronker, 918 N.Y.S.2d 822, 822 (N.Y. Supp. Ct. 2010).
\textsuperscript{175} Whether Out-of-State Same-Sex Marriage That is Valid in the State of Celebration May be Recognized in Maryland, 95 Md. Op. Att’y Gen. 3 (2010).
\textsuperscript{176} Maryland Offers Health Benefits to Workers in Same-Sex Marriages from Other States, 36 FAMILY LAW REP. (BNA) 1335, 1335 (2010).
The largest group of states are those that not only do not allow same-sex marriage (or quasi marriage), but also explicitly provide by a state statute or constitutional provision that they will not recognize a same-sex marriage validly entered into elsewhere. Following the federal DOMA, many of these state provisions are known as “mini-DOMAs” or “state DOMAs.” A survey published in the BNA Family Law Reporter in June 2010 concluded:

As of June 2, 2010, 45 states prohibit same-sex marriage. Ten do so through statute only, four through state constitution amendments only, 27 through both statute and state constitution amendments, two through case law (New York and New Jersey), and two through the state attorney general’s office (New Mexico and Rhode Island). Depending on one’s statutory construction, approximately 40 of those expressly refuse to recognize same-sex marriages of other jurisdictions, and some of those more broadly refer to other same-sex relationships.178

Pennsylvania enacted a typical mini-DOMA in 1996, containing two new statutory provisions. The first defines marriage as “[a] civil contract by which one man and one woman take each other for husband and wife.”179 The second addresses interstate recognition:

It is hereby declared to be the strong and longstanding public policy of this Commonwealth that marriage shall be between one man and one woman. A marriage between persons of the same sex which was entered into in another state or foreign jurisdiction, even if valid where entered into, shall be void in this Commonwealth.180

IV. BETWEEN A ROCK AND A HARD PLACE: SEEKING A LEGAL EXIT FROM A SAME-SEX MARRIAGE IN A STATE THAT DOES NOT RECOGNIZE THAT MARRIAGE

Americans are a famously restless people. Two centuries ago, Alexis de Tocqueville observed that, “[i]n the United States a man builds a house in which to spend his old age, and he sells it before the roof is on; . . . he settles in a place, which he soon afterwards leaves to carry his changeable longings elsewhere.”181 Those words are even truer in today’s world of high-speed transportation and the Internet than when they were written in the 1830s.

179 23 PA. CONS. STAT. § 1102 (2012).
180 Id. § 1704.
181 2 ALEXIS DE TOCQUEVILLE, DEMOCRACY IN AMERICA 136 (Alfred A. Knopf 1945).
Moreover, the break up of a serious relationship often triggers, or is triggered by, a move of one or both of the parties to that relationship. A party may leave and put distance between herself and her spouse or partner to escape abuse, to take a new job, to be near or live with family members or friends who can provide a support system (especially when she has minor children), to follow or join a new significant other, or simply to get a “fresh start.” Normally the physical departure from the relationship and the situs of the relationship precedes any serious thought about legally ending the relationship. Indeed, physical separation is often deemed by one or both of the parties to be part of a “trial separation.”

Additionally, for a variety of reasons, couples often get married in a jurisdiction where they do not reside. They may marry where one or both have family. They may choose to have a “destination wedding” in some romantic or vacation location. As was the case in May’s Estate, they may temporarily leave a jurisdiction where they cannot marry, travel to a jurisdiction where they can and do marry, and then return home.

For all these reasons, it is not surprising that an individual may well reside in a different jurisdiction from the one in which she married at the time that she decides to initiate divorce proceedings. If she has left a same-sex marriage (or quasi marriage) and is domiciled in a jurisdiction that refuses to recognize that marriage, she is likely to find herself in a form of legal limbo. A recent Pennsylvania case, Kern v. Taney, illustrates her dilemma.

Two women, Carole Kern and Robin Taney, were married in Massachusetts. Subsequently, Carole moved to Pennsylvania and filed for a divorce, utilizing the Pennsylvania no-fault divorce ground of irretrievable breakdown of the marriage. Robin did not appear to defend the action. However, the Attorney General of Pennsylvania intervened in order to defend the constitutionality of Pennsylvania’s mini-DOMA.

The trial judge reasoned that, “relief under the Divorce Code can only be obtained by parties who are recognized to be married.” Under the second section of Pennsylvania’s mini-DOMA, Section 1704 of the Domestic Relations Code, quoted above, the parties could not be recognized as married. Therefore, Carole attacked the constitutionality of the act, asserting that it violated her substantive due process and equal protection rights to marry under both the Pennsylvania and United States Constitutions.

The trial court dismissed all of Carole’s constitutional challenges, finding that homosexuals have no fundamental right to be married to each other. The court

---

183 Id. at 559.
184 Id. at 559–60.
185 See id.
186 Id.
187 Id. at 562.
188 Id. at 562–63.
189 Id. at 564.
190 Id. at 574.
applied the “rational basis” test. In arguably a circular piece of reasoning, the court concluded, “The amendment did not expand, limit, alter or otherwise change the law of the Commonwealth of Pennsylvania. As such, the legislation did not impose an inequality on homosexuals.” Accordingly, the court could not grant her a divorce. The court did, however, offer Carole an alternative legal solution:

Plaintiff has a concern that she has no available remedy in Pennsylvania, and since she does not qualify under the residency requirements of Massachusetts, she is unable to obtain a divorce. While it is true that Pennsylvania cannot grant her a divorce, there is no reason why she cannot seek relief under section 1704, requesting the court to have her marriage declared void.

However, a declaration in Pennsylvania that Carole and Robin’s Massachusetts marriage was void as against Pennsylvania public policy would hardly be the equivalent of a Pennsylvania divorce decree. Under Williams I, a divorce decree should be entitled to full faith and credit in all states. It is difficult to believe that a decree of annulment based on Pennsylvania’s public policy against same-sex marriage would be accorded full faith and credit in those states that permit such marriages, especially Massachusetts. So, with a Pennsylvania annulment, Carole might well find herself in the “unhappy” circumstance that befell O.B. and Lillie in the Williams litigation. She would be married in one state and not in another. As was the case with O.B. and Lillie, it would remain questionable whether she could legally remarry. If, after obtaining an annulment in Pennsylvania, she were to marry a man in Pennsylvania, could they honeymoon on “Old Cape Cod” per Patti Page’s old chartbuster? If they did, could not Massachusetts arrest, try and punish her for bigamy under Massachusetts law, just as happened to O.B. and Lillie seven decades ago in North Carolina? Indeed, could not that fate befall her if she were to go to any of the states that either permit or recognize same-sex marriage?

Presumably the only effective remedy theoretically available to Carole would be to file for divorce in Massachusetts. But, in Massachusetts, as elsewhere in the United States, it is significantly more time-consuming to get divorced than to get married. As noted by the trial court, there is no residency requirement to be married in Massachusetts, but to get a divorce generally the parties have to have resided in Massachusetts together for a year preceding the commencement of the action. Since Carole had to have resided in Pennsylvania for six months before

191 Id.
192 Id. at 575.
193 Id. at 576.
194 Id.
195 See supra text accompanying notes 93–105.
filing her Pennsylvania divorce action, she would have to move to Massachusetts—and presumably find housing and employment—for a year just to commence a divorce action there.

The result in *Kern v. Taney* is consistent with that reached in other mini-DOMA jurisdictions in similar situations (with three recent notable exceptions that will be discussed infra). Thus, in 2007, in *Chambers v. Ormiston*, the Rhode Island Supreme Court was presented with this certified question:

> May the Family Court properly recognize, for the purpose of entertaining a divorce petition, the marriage of two persons of the same sex who were purportedly married in another state?

In *Chambers*, two Rhode Island women, Margaret Chambers and Cassandra Ormiston, had married each other in Massachusetts in 2004, and then returned to reside together in Rhode Island. In October 2006, Ms. Chambers filed for divorce in Rhode Island. The Family Court was concerned that it lacked jurisdiction and asked for guidance from the state’s highest court as to whether the parties were married under Rhode Island law. The Rhode Island courts assumed that the parties’ marriage was valid under Massachusetts law. But, the Rhode Island Supreme Court ruled that “marriage” under its state statute is “the state of being united to a person of the opposite sex.” Since the parties, therefore, were not married under Rhode Island law, the Rhode Island courts lacked jurisdiction to entertain a divorce action.

Like the Pennsylvania trial court in *Kern*, the Rhode Island Supreme Court expressed some sympathy for the thwarted plaintiff:

> We know that sometimes our decisions result in palpable hardship to the persons affected by them. It is, however, a fundamental principle of jurisprudence that a court has no power to grant relief in the absence of jurisdiction, as is true in the instant case. Ours is not a policy-making branch of the government. We are cognizant of the fact that this observation may be cold comfort to the parties before us. But, if there is

---

198 23 PA. CONS. STAT. § 3104(b) (2012).
199 The U.S. Supreme Court has upheld the constitutionality of a one-year residency requirement to commence a divorce action in *Sosna v. Iowa*, 419 U.S. 393, 410 (1975).
200 See infra text accompanying notes 228–245.
201 935 A.2d 956 (R.I. 2007).
202 Id. at 958.
203 Id.
204 Id. at 958–59.
205 Id. at 959.
206 Id. at 958–59.
207 Id. at 962.
208 Id. at 967.
to be a remedy to this predicament, fashioning such a remedy would fall within the province of the General Assembly.\(^{209}\)

In 2008, in *O’Darling v. O’Darling*,\(^{210}\) the Oklahoma Supreme Court ruled that a trial court judge had properly vacated a divorce decree of a couple that had been married in Canada, where the trial judge learned after entering the decree that both parties were women.\(^{211}\) The state supreme court admonished counsel for the plaintiff for having failed to disclose the fact that the marriage was between two women, hence invalid under Oklahoma law.\(^{212}\)

In 2010, a Texas court of appeals likewise ordered dismissal of a divorce action filed between two men who had been married in Massachusetts in *In the Matter of the Marriage of J.B. and H.B.*\(^{213}\) In *J.B. and H.B.*, the trial court had granted the divorce, ruling that the state’s constitutional and statutory provisions barring recognition of same-sex marriage violated the Equal Protection Clause of the Fourteenth Amendment to the United States Constitution.\(^{214}\) On appeal by the state, the Texas Court of Appeals reversed and ordered dismissal of the divorce action for lack of subject matter jurisdiction.\(^{215}\) The Texas Constitution had been amended in 2005 to provide:

(a) Marriage in this state shall consist only of the union of one man and one woman.
(b) This state or a political subdivision of this state may not create or recognize any legal status identical or similar to marriage.\(^{216}\)

Further, the Texas Family Code had been amended to provide in Section 6.204:

(b) A marriage between persons of the same sex or a civil union is contrary to the public policy of this state and is void in this state.
(c) The state or an agency or political subdivision of the state may not give effect to a:

(1) public act, record, or judicial proceeding that creates, recognizes, or validates a marriage between persons of the same sex or a civil union in this state or in any other jurisdiction; or

---

\(^{209}\) *Id.* at 966–67.
\(^{210}\) 188 P.3d 137 (Okla. 2008).
\(^{211}\) *Id.* at 138.
\(^{212}\) *Id.* at 139. The state supreme court remanded the case for procedural reasons. *Id.* at 140.
\(^{213}\) 326 S.W.3d 654 (Tex. App. 2010).
\(^{214}\) *Id.* at 659.
\(^{215}\) *Id.* at 681.
\(^{216}\) *Id.* at 663 (citing *TEX. CONST.* art I, § 32).
right or claim to any legal protection, benefit, or responsibility asserted as a result of a marriage between persons of the same sex or a civil union in this state or in any other jurisdiction.217

The appellate court readily concluded that these constitutional and statutory provisions barred the same-sex divorce action.218 Thus the court was compelled to address whether these provisions violated the United States Constitution.219 Applying the rational basis test, the court concluded: “Texas’s marriage laws are rationally related to the goal of promoting the raising of children in households headed by opposite-sex couples.”220

Finally, the appellate court noted that the plaintiff could file a “voidance action” seeking to have his marriage annulled, even though he would not have all the ancillary remedies available in that action that he would have had in a divorce action, such as spousal maintenance and community property rights.221 The court quite unconvincingly disagreed with his contention that such a declaration of voidance might not be recognized in other jurisdictions.222 But the court failed to provide any cogent reason why Massachusetts, for example, would give full faith and credit to a declaration that a Massachusetts same-sex marriage is void as against public policy.223

In a similar case, Rosengarten v. Downes,224 decided by the Appellate Court of Connecticut six years before Connecticut authorized same-sex marriage, the court ruled that Connecticut courts lacked jurisdiction to entertain an action by one of its residents to dissolve a same-sex civil union he had entered in Vermont.225 The court reasoned that, “[i]f Connecticut does not recognize the validity of such a union, then there is no res to address and dissolve.”226

217 Id.
218 Id. at 669–70.
219 Id. at 670.
220 Id. at 677.
221 Id. at 678–79.
222 Id. at 679.
223 There is one Texas trial court case in which the court granted a divorce by agreement to two women who had been married in Massachusetts. The state tried to intervene unsuccessfully. The state appealed, but the Court of Appeals, Austin, ruled that the state lacked standing and dismissed the appeal. State v. Naylor, 330 S.W.3d 434 (Tex. App. 2011), petition for review filed Mar. 21, 2011. Thus, while the divorce decree remains valid, the appellate decision cannot be construed as an affirmance on the merits, nor is it inconsistent with In the Matter of the Marriage of J.B. and H.B.224
225 Id. at 172, 184.
226 Id. at 175. The refusal of most American states to recognize valid same-sex marriages from other jurisdictions does not always disadvantage one or both parties to such a marriage. In the anomalous case of In re Marriage of Bureta, a former husband sought to end his pension payments to his ex-wife on the grounds that she had remarried. 164 P.3d 534, 534 (Wash. Ct. App. 2007). She had traveled to Oregon with her female partner,
It was not until June 2011 that a state appellate court in a mini-DOMA jurisdiction found a way to grant relief to an individual seeking legal escape from a foreign same-sex marriage. In Christiansen v. Christiansen, two women, Paula and Victoria, had been legally married in Canada in 2008. Paula filed an apparently uncontested divorce action against Victoria in Wyoming in 2010. The district court dismissed the case for lack of subject matter jurisdiction, applying the now familiar reasoning that since the forum state does not recognize same-sex marriage, the state’s divorce law did not apply. In a brief and unanimous opinion, the Wyoming Supreme Court reversed the district court and remanded the case.

The Wyoming Supreme Court expressly limited its analysis to recognition of a foreign same-sex marriage for the sole purpose of granting a divorce. “The question of recognition of such same-sex marriages for any other reason, being not properly before us, is left for another day.”

The Court viewed the matter as one of statutory construction, attempting to resolve statutory provisions in apparent conflict with each other. Wyoming Statute Annotated §20-1-111 provides, “all marriage contracts which are valid by the laws of the country in which contracted are valid in this state.” But, Wyoming’s mini-DOMA defines a marriage as “a civil contract between a male and a female person . . . .” Significantly, however, Wyoming’s mini-DOMA “does not speak to recognition of a same-sex marriage validly entered into [elsewhere].”

The Court acknowledged long-standing case law that there are exceptions to Wyoming’s recognition of validly entered-into foreign marriages: “namely, marriages which are deemed contrary to the law of nature as generally recognized in Christian countries, such as polygamous and incestuous marriages, and those which the legislature of the state has declared shall not be allowed any validity, because contrary to the policy of its laws.”

---

227 I use “foreign” in the sense of extra-territorial. This could mean another state, although in this case the parties were married in a foreign country.
228 253 P.3d 153 (Wyo. 2011).
229 Id. at 154.
230 Id. Victoria did not file a brief in the subsequent appeal. Id.
231 Id. at 154–55.
232 Id. at 157.
233 Id. at 154 n.1.
234 Id. at 155 (quoting Wyo. Stat. Ann. § 20-1-111 (2009)).
237 Id. (citing Hoagland v. Hoagland, 193 P. 843, 843–44 (Wyo. 1920)).
However, the exceptions are meant to be narrow, lest they “swallow the rule.”238 Thus, for example, although Wyoming will not permit a common law marriage to be created within the state, it will consider valid a common law marriage legally entered into in another state.239 Accordingly, the Court concluded that “recognizing a valid foreign same-sex marriage for the limited purpose of entertaining a divorce proceeding does not lessen the law or policy in Wyoming against allowing the creation of same-sex marriages.”240

The Court noted that all that was being sought was a divorce, that the parties were “not seeking to live in Wyoming as a married couple . . .” and, importantly, that they “are not seeking to enforce any right incident to the status of being married.”241

In July 2011, between the date that the New York legislature enacted its Marriage Equality Act and that Act’s effective date, an appellate court in New York reached a similar conclusion in Dickerson v. Thompson.242 Two women, Audrey and Sonya, had entered into a civil union in Vermont. Unable to meet Vermont’s residency requirements for a dissolution action, Audrey brought an action in New York to dissolve the civil union, and Sonya did not defend that action. The trial court dismissed the action for lack of subject matter jurisdiction, and the appellate court reversed and remanded.243 On remand, the trial court entered “a declaration relieving the parties from all rights and obligations arising from the civil union, but denied that portion of the motion seeking a dissolution of the union.”244 The appellate division again reversed. “We disagree with the [trial court’s] conclusion that, in the absence of any legislatively created mechanism in New York by which a court could grant the dissolution of a civil union entered into in another state, it was powerless to grant the requested relief.”245

Most recently, in May 2012, the Maryland Court of Appeals reached a similar conclusion in Port v. Cowan.246 Two women, Jessica and Virginia, had been legally married in California in 2008 when such marriages could be legally performed there. They separated two years later by mutual agreement. Subsequently Jessica filed for divorce in Maryland on the ground of voluntary separation, and Virginia answered the complaint in a “no contest” manner. The couple had no children, and neither raised a financial claim against the other. Nevertheless the trial court denied the divorce on the ground that the marriage was not valid under Maryland law.247

238 Id.
239 Id.
240 Id.
241 Id.
243 Id. at 97.
244 Id. at 99, 123.
245 Id.
247 Id. at *1.
The Maryland Court of Appeals unanimously reversed.\textsuperscript{248} Although at the
time of this case Maryland Family Law provided that “only a marriage between a
man and a woman is valid in this State,”\textsuperscript{249} it did not specifically address the
recognition of out-of-state same-sex marriages legally performed in another
jurisdiction. The Court found that “for purposes of the application of its domestic
divorce laws,” the doctrine of comity compels recognition of the marriage, and that
such recognition is not repugnant to Maryland public policy.\textsuperscript{250}

Unfortunately, the \textit{Christiansen, Dickerson} and \textit{Port} decisions will be of little
or no value to unhappy spouses locked in same-sex marriages in most of the
United States. The approach of the Wyoming Supreme Court in \textit{Christiansen},
whatever its merits under Wyoming law, cannot be utilized in the vast majority of
mini-DOMA states. That court was not confronted with a state statute explicitly
barring recognition of a foreign same-sex marriage, nor was the \textit{Dickerson} court
confronted with such a statute in New York.\textsuperscript{251} Indeed, the Maryland Court of
Appeals in \textit{Port} noted that whereas other states, such as Pennsylvania and Virginia,
have enacted specific statutory provisions preventing recognition of foreign same-
sex marriages, Maryland's statute is silent on the subject.\textsuperscript{252} But, as noted above,
approximately forty of forty-five mini-DOMA states do have statutory or
constitutional provisions explicitly barring such recognition.\textsuperscript{253}

\section*{V. THREADING THE NEEDLE: A PATH FORWARD}

Legal scholars who have examined this issue have proposed various ingenious
solutions to address it, none of which, as the cases cited above show, have
commanded judicial respect.

Professor Barbara J. Cox, herself in a same-sex marriage entered into in
Ontario, Canada, has argued that courts in mini-DOMA states:

\begin{quote}
should consider whether an ‘incidents of marriage’ approach to the
issue in the case may lead them to recognize the civil union, domestic
partnership, or marriage based on the policy reasons behind that
disputed issue. They should work as hard to honor the relationships of
same-sex couples as they have worked to honor the relationships of
opposite-sex couples.\textsuperscript{254}
\end{quote}

\begin{footnotes}
\item[248] \textit{Id.}
\item[249] \textit{Id.} (quoting MD. CODE ANN., FAM. LAW § 2-201 (2009)).
\item[250] \textit{Id.} at *6.
\item[251] \textit{See supra} text accompanying notes 234–237, 245.
\item[253] \textit{See supra} text accompanying note 178.
\item[254] Barbara J. Cox, \textit{Using an “Incidents of Marriage” Analysis When Considering
Interstate Recognition of Same-Sex Couples’ Marriages, Civil Unions, and Domestic
\end{footnotes}
Under this approach a court could address the benefits, rights, and responsibilities flowing to a couple, without necessarily recognizing the marriage itself. There are two major problems with this approach. First, unless those benefits, rights or responsibilities flow out of a valid antenuptial agreement, they don’t exist absent a valid marriage. Second, a finding of a valid marriage will be not only contrary to the state’s mini-DOMA, but also be politically untenable in a state that has enacted such a statute or constitutional amendment. Indeed, the very plea that courts in such states should work “hard to honor the relationships of same-sex couples” is doomed to failure (absent, of course, repeal of the state mini-DOMA).

Professor Linda Silberman has taken a more cautious approach. She has proposed “balanced choice-of-law rules,” along the line of the old “marriage evasion” laws whereby the problem is avoided by having:

states . . . limit the application of their same-sex marriage or civil union laws to members of their own community—either through a residency requirement or by restricting application of the law to persons who do not face an impediment to such a marriage under the laws of the jurisdictions where they reside or intend to reside.

There are two main problems with this approach. First, it provides no avenue of legal redress to the person who entered a same-sex marriage while residing or intending to reside in a same-sex marriage jurisdiction, who later—for any of myriad reasons—relocates to a mini-DOMA state. Second, as a practical matter, the genie is already out of the bottle. The first same-sex marriage state, Massachusetts, repealed its “marriage evasion” act in 2008, after its courts used that act to bar same-sex couples from mini-DOMA states from getting married in Massachusetts. Proponents of repeal explicitly noted that Massachusetts had an economic interest in becoming a same-sex marriage destination:

State officials said they expected a multimillion-dollar benefit in weddings and tourism, especially from people who live in New York. A just-released study commissioned by the State of Massachusetts concludes that in the next three years about 32,200 couples would travel here to get married, creating 330 permanent jobs and adding $111 million to the economy, not including spending by wedding guests and tourist activities the weddings might generate.

255 Id. at 718–19.
256 Id.
258 Id. at 2204, 2213.
“We now have this added pressure, given what’s happened in California, that we really think that it is a good thing that we be prepared to receive the economic benefit,” State Senator Dianne Wilkerson, a Democrat who sponsored the repeal bill, said Tuesday after the vote.260

Several law student notes and comments have struggled heroically to resolve the issue of same-sex divorce in mini-DOMA jurisdictions. Writing in the Hastings Law Journal in 2003, Jessica A. Hoogs proposed that states create a “uniform dissolution proceeding,” presumably through legislative enactment.261 Given the failure of the states to generally adopt the Uniform Marriage and Divorce Act262 and the political divide over same-sex unions, this clever idea appears to be infeasible.

Writing in the Marquette Law Review in 2009, Louis Thorson suggested three methods that Wisconsin courts could use in same-sex divorce cases: 1) bar access to the courts for relief, 2) apply Wisconsin divorce law, or 3) have Wisconsin courts apply the laws of the state where the relationship was founded.263 He acknowledged that while all three approaches have their justifications, they also have their own difficulties.264 He admitted that the second approach, applying Wisconsin divorce law, “likely would violate both the Wisconsin Statutes and the Wisconsin Constitution.”265

Writing in the Boston University Law Review, also in 2009, John M. Yarwood argued that mini-DOMA states should create property distribution mechanisms for same-sex couples seeking to terminate an out-of-state same-sex marriage.266 While this might be a “consummation devoutly to be wished,”267 unfortunately it probably falls within the category of wishful thinking, given current political realities.

Writing in the Santa Clara Law Review in 2010, Danielle Johnson proposed, “courts should use an incidental approach to marriage recognition when

262 See supra note 5.
264 Id. at 619.
265 Id. at 642.
266 John M. Yarwood, Breaking Up is Hard to Do: Mini-DOMA States, Migratory Same-Sex Marriage, Divorce, and a Practical Solution to Property Division, 89 B.U. L. REV. 1355, 1388 (2009).
267 WILLIAM SHAKESPEARE, HAMLET, PRINCE OF DENMARK 51 (W.G. Clark & W.A. Wright eds., 2d ed. 1874).
considering a divorce petition in order to avoid unreasonably burdensome, illogical results.\footnote{268} She argued cogently that:

When the law of the forum state conflicts with, or is silent on, the legality of the underlying marriage, the court can use an incidental approach to marriage recognition and consider the divorce as an incident of that marriage. By recognizing the marriage for the limited purpose of the divorce, the court can confine its consideration of the relationship so as to avoid addressing the validity of the underlying marriage. The ability to legally end a marriage validly performed in another state is an incident of that marriage that should be available uniformly across the states, regardless of whether that state disagrees with the underlying marriage. Parties seeking an uncontested dissolution of their union are not asking the court to validate the union; they are simply asking the court to dissolve it. By refusing to perform a divorce in a same-sex couple’s home state, some states have made it incredibly burdensome for that couple to legally end their relationship.\footnote{269}

She concluded:

Using the incidental approach, the court can view divorce as an incident of marriage, analyze the policies behind the incident at issue, and then decide whether the marriage should be recognized for the sole purpose of performing the divorce.\footnote{270}

While this approach has the merit of being practical and is similar to what I will suggest, it has one fatal flaw. It would require a court in a mini-DOMA state to do something it is prohibited from doing: recognize a same-sex marriage. Any effort to bridge the enormous divide between those states that permit same-sex marriage and those that consider it an anathema is obviously fraught with peril. Bearing in mind Justice Holmes’ aphorism that, “[t]he life of the law has not been logic: it has been experience,”\footnote{271} surely an incremental approach which respects the position of anti-same-sex marriage jurisdictions while providing relief to their unhappily wed citizens is the most likely of success.

The author’s proposal for same-sex divorce is minimalist: Where a party to a same-sex marriage seeks a simple, uncontested, no-fault divorce in a mini-DOMA jurisdiction, the court can and should grant the divorce without inquiring into or addressing the validity of the marriage.

\footnote{269} Id. at 245–46.
\footnote{270} Id. at 253–54.
\footnote{271} OLIVER WENDELL HOLMES, JR., THE COMMON LAW I (Harvard Univ. Press 1881).
It must be acknowledged that this proposal will not aid the happily married (or quasi-married) same-sex couple now residing in a mini-DOMA jurisdiction. Under current law, they have no benefits flowing out of their marital relationship other than those that might be secured by contract. This proposal will not circumvent the incidents of marriage rules articulated by the Court in *Estin*, *Vanderbilt*, and their progeny. Even a court that might be persuaded to grant a divorce would probably be barred from addressing financial issues that it would normally resolve in the dissolution of an opposite-sex marriage. The proposal would also provide no relief, for example, to a member of a same-sex couple whose spouse is negligently killed in a mini-DOMA state, who wishes to bring a wrongful death claim.272

The proposed solution has several important benefits. First, it is completely consistent with dominant legal practice in the United States today. Since the advent of no-fault divorce in California in 1970,273 all states have made efforts to simplify the divorce process and make it less adversarial.274 Based on the author’s three decades of family law practice, it would be truly extraordinary for a court to spend its time in an uncontested no-fault divorce questioning the validity of the marriage.

Second, and in the same vein, judicial resources are scarce and judicial time precious. How does it benefit the court or the parties to waste limited judicial resources inquiring into the validity of a marriage when the only action before the court is an uncontested one to terminate the marriage?

Third, as noted, courts in some of the cited cases have recognized the hardship imposed on their own residents by refusing to grant a divorce in this situation.275 Hence, one may be able to appeal to the judge’s sense of equity in seeking such a result.

Fourth, this proposal is neither fanciful nor radical. The author has served as codirector of his law school’s Family Law Clinic for almost three decades. During this time, the clinic has filed divorce complaints where it was far from clear that the client was legally married. For example, in one case, the client and her husband had separated years before, and she had no way to contact him.276 She recalled receiving some papers from a lawyer long ago about a divorce but had long since lost them and didn’t even know what state they were from. She asked the clinic if she were already divorced, and, of course, no one could tell her.277 The only practical option to clarify her legal situation was to file a divorce and serve her

---

272 See, e.g., Littleton v. Prange, 9 S.W.3d 223, 223, 231 (Tex. Ct. App. 1998) (holding that the person who had married a male decedent could not maintain a wrongful death action as a surviving spouse because she was a male-to-female transsexual and hence in a non-recognized same-sex marriage).


274 See id.

275 See, e.g., supra text accompanying note 241.

276 Since this case did not result in a reported decision, the client’s name is not cited here for privacy reasons.

277 There is no national register of divorces in the United States.
husband by publication. He did not enter an appearance, and the court granted her a no-fault divorce without further ado. In another case, where the parties had had a marriage ceremony in another state but it appeared that they had failed to obtain a marriage license, the clinic filed a divorce for the wife, and the husband appeared and defended on the grounds that there was no valid marriage. Once the defendant spouse raised the issue, the court quite properly held a hearing on the subject (and ruled that there was a valid marriage).\textsuperscript{278} The point is that it is perfectly appropriate—and commonplace—to file a divorce even where a party’s marital status might be questioned, and a court will not ordinarily waste its time conducting an inquiry into marital status when a simple, no-fault divorce is uncontested.\textsuperscript{279}

Fifth, while a purist might question the logic of granting a divorce from a void marriage, there is nothing that inherently prevents a court from granting a divorce where an annulment might also be available. Pennsylvania statutory law contains an explicit example. Section 3304(a)(1) of the Pennsylvania Domestic Relations Code, “[g]rounds for annulment of void marriages,” provides:

(a) \textbf{General rule}—Where there has been no confirmation by cohabitation following the removal of an impediment, the supposed or alleged marriage of a person shall be deemed void in the following cases:

(1) Where either party at the time of such marriage had an existing spouse and the former marriage had not been annulled nor had there been a divorce except where that party had obtained a decree of presumed death of the former spouse.\textsuperscript{280}

Thus, a woman (or man) who discovers, as one of the clinic’s clients did, that her spouse was married all along to someone else, may seek and obtain an annulment of her void marriage. But, she also has a second legal option: divorce. Section 3301(a)(4) of the Domestic Relations Code, “[g]rounds for divorce,” provides:

\textsuperscript{278} Jagdeo v. Dookharan, 58 Cumb. 195 (2009).
\textsuperscript{279} This approach is also completely consistent with the only appellate case in Pennsylvania addressing same-sex marriage, De Santo v. Barnsley, 476 A.2d 952, 956 (Pa. Super. Ct. 1984). In De Santo, one man sued another man in divorce claiming that they had entered into a common law marriage. \textit{id.} at 952. The defendant filed an answer denying that the defendant and the plaintiff were ever married or were capable of being married. \textit{id.} Since the defendant put the existence of the marriage at issue, it was entirely appropriate for the trial court to address that matter, and it did so, finding that there was no valid marriage. \textit{id.} That finding was affirmed on appeal, with the superior court ruling as a matter of law that two persons of the same sex could not contract a common law marriage in Pennsylvania. \textit{id.}
\textsuperscript{280} 23 PA. CONS. STAT. § 3304(a)(1) (1990).
(a) Fault—The court may grant a divorce to the innocent and injured spouse whenever it is judged that the other spouse has: . . .

(4) Knowingly entered into a bigamous marriage while a former marriage is still subsisting.  

The fact that such a marriage is void and subject to annulment does not prevent a court from granting a divorce.

Finally, it can be readily and honestly argued that this approach is fully consistent with the mini-DOMA states’ anti-same-sex marriage position. The cases where courts have denied a divorce have had the counter-productive result of preserving a same-sex marriage rather than terminating it. By refusing to grant the divorce, the court is assuring that its resident remains in the very same-sex marriage that is antithetical to the state’s public policy. For reasons stated above, even an annulment in the mini-DOMA state is unlikely to free its resident from her same-sex marriage in states that recognize such marriages. On the other hand, a divorce granted in compliance with the dictates of Williams I would be entitled to full faith and credit in all states.

Indeed, in striking down a mandatory filing fee for poor people seeking divorces, the Supreme Court recognized the inextricable connection between the right to divorce and the right to marry:

Our conclusion is that, given the basic position of the marriage relationship in this society’s hierarchy of values and the concomitant state monopolization of the means for legally dissolving this relationship, due process does prohibit a State from denying, solely because of inability to pay, access to its courts to individuals who seek judicial dissolution of their marriages.

The short-term result of a universally recognized divorce is one fewer same-sex marriage. For the plaintiff spouse, the long-term result may be either: 1) remaining single, or 2) entering an opposite-sex marriage, or 3) entering another same-sex marriage. The first two long-term outcomes carry forward the state’s antisame-sex marriage position. The third outcome is actually neutral: the individual is still in a same-sex marriage, albeit a new one, and the sum total of same-sex marriages is not affected. The second outcome is not at all fanciful. Individuals have been known to leave same-sex relationships and, then or later, form opposite sex relationships. It would be the height of irony for a court’s

281 Id. at § 3301(a)(4).
283 Of course, it is possible that the other party to the initial same-sex marriage might also remarry another person of the same sex, which would create an additional same-sex marriage.
284 See, for example, L.S.K. v. H.A.N., a Pennsylvania child support case between two formerly lesbian partners, in which the court noted that both women are now married. 813
refusal to grant an uncontested divorce to someone in a same-sex marriage to result in that person’s not being truly legally free to enter into an opposite-sex marriage, the very institution the mini-DOMA states are supposedly trying to preserve and support.

A.2d 872, 875 n.2 (Pa. Super. Ct. 2002). Similarly, in the long-running interstate custody battles between former Vermont civil union partners, Lisa Miller and Janet Miller-Jenkins, see Miller-Jenkins v. Miller-Jenkins, 12 A.3d 768 (Vt. 2010), Lisa Miller has purportedly “renounced her homosexuality,” rediscovered her Baptist faith, and “is often flanked by others who’ve renounced their homosexuality and joined the faith.” See Lorraine Ali, Mrs. Kramer vs. Mrs. Kramer, NEWSWEEK (Dec. 6, 2008), http://www.newsweek.com/2008/12/05/mrs-kramer-vs-mrs-kramer.print.html.
THE COMMON LAW OF DISABILITY DISCRIMINATION

Mark C. Weber*

Abstract

In many cases alleging race and sex discrimination, plaintiffs append common law claims to cases asserting federal or state statutory causes of action. In other race and sex cases, plaintiffs put forward these common law claims without making any federal or state statutory claims. Less frequent, and much less frequently discussed by scholars, are common law claims for conduct constituting disability discrimination. Nevertheless, there are sound theoretical and practical reasons to develop a common law of disability discrimination.

On the theoretical side of the discussion, federal statutory disability discrimination claims are not exclusive, and the common law can both draw from and influence statutory developments. The evolution of the common law can be part of the adaptation of the social and legal environment that is needed to achieve equality for people with disabilities. Practically speaking, there are numerous obstacles to statutory disability discrimination claims; the common law may provide redress when statutory remedies are blocked. Common law claims may face difficulties of their own, however, and the law may need to be reformed to facilitate just results in common law cases.

Existing scholarship includes several prominent discussions of disability and the law of torts, but there has been little development of the most important tort and contract remedies for disability discrimination. This Article seeks to contribute to the scholarly discussion by considering common law remedies for disability discrimination in a systematic way and discussing how to align the remedies more closely with the goal of protecting civil rights of individuals with disabilities.

* © 2011 Mark C. Weber. Vincent DePaul Professor of Law, DePaul University. Thanks to participants in workshops at Marquette University Law School and the Southeastern Association of Law Schools Annual Conference for comments on earlier versions of this paper. Special thanks to Cheryl Anderson, Carrie Basas, Ravi Malhotra, Bruce Ottley, Laura Rothstein, and Paul Secunda for their comments on the manuscript. Many thanks to my research assistants: Kim Brown, Jennifer Coco, Christopher Cook, and Seth Condon.
INTRODUCTION

More than twenty-five years ago Professor Marc Greenbaum wrote, “If courts permit discrimination plaintiffs to pursue their claims under a common law theory, a revolutionary change in the enforcement of the employment discrimination laws will occur.”¹ Common law discrimination claims, he said, would enhance plaintiffs’ chances of success and expand the remedies they may obtain.² In the past quarter-century, race and sex discrimination plaintiffs have regularly filed claims asserting violations of common law duties, often attaching these claims to lawsuits charging violations of federal or state civil rights statutes.³ Scholars have considered the applicability of these common law claims and analyzed how they might supplement or stand apart from statutory discrimination remedies.⁴ Less frequent, and far less frequently studied, is the practice of disability discrimination plaintiffs bringing common law claims.

The idea of common law remedies for disability discrimination may be counterintuitive. Unlike misconduct based on race or sex, failure to employ or accommodate a person who has a disability does not necessarily suggest tortious misconduct or violations of contract rights. Even outright harassment on the basis of disability has proven a chancy basis for common law relief.⁵ Nevertheless, it remains theoretically and practically sound to assert common law actions in cases of discrimination on the basis of disability.

As a theoretical matter, a breach of contract rights or a traditional tort cause of action should be available if the activity is intentionally or by effect discriminatory, whenever the elements of the common law claim are met.⁶ Statutory claims under the Americans with Disabilities Act (ADA) are not meant to have any preemptive or supplanting effect.⁷ Common law remedies may draw analogies from statutory claims, and construction of statutory remedies may draw reciprocally from analogies to tort and contract.⁸ Moreover, common law reflects

² Id. at 68–69.
⁴ See infra note 31 and accompanying text (collecting sources on common law remedies for race and sex discrimination).
⁵ See MARK C. WEBER, DISABILITY HARASSMENT 109–12 (2007) (“Like statutory disability harassment claims, many, many employment, education, and other kinds of intentional-infliction cases fail, and those that fail greatly outnumber those that succeed on motions to dismiss or motions for summary judgment or at trial.”) (collecting cases).
⁶ See infra Part I.
⁸ See infra text accompanying notes 32–35.
the community’s sense of morality. As disability discrimination becomes more widely recognized as a social wrong, a common law remedy is a natural development. The application of the common law also fits in with a civil rights approach to disability, one that recognizes the role of the environment of physical spaces and attitudes in keeping people with disabilities from achieving social equality and that looks to legal tools to alter physical and attitudinal barriers.9

From the perspective of claimants, there are also practical benefits to asserting common law claims. These are varied, but they include the coverage of a greater range of defendants in employment and public accommodations cases, the prospect of higher damages awards, protection of a larger class of claimants, possible relief from limitations and exhaustion requirements, greater ease in establishing causation, and even a higher likelihood of avoiding summary judgment and reaching the jury.10 Much disability discrimination remains unremedied.11 The unemployment rate among people who have disabilities is almost 60 percent higher than that of people without disabilities.12 Unsurprisingly, the incidence of poverty among working-age adults with a disability is about one and one-half times that of comparable individuals without a disability.13 Even though the recent Americans with Disabilities Act Amendments Act (ADAAA) clarified that the federal disability discrimination law has broad coverage of claimants with

9 See infra text accompanying notes 51–64.
10 See infra Parts II.B & IV.A.
disabling conditions,\textsuperscript{14} it did nothing to alter the ADA’s limits on coverage of employers, its exhaustion provisions, and its restrictions on remedies, so barriers to remediation persist. Common law claims that could provide a remedy for discriminatory conduct against people with disabilities include contract actions such as breach of promise to accommodate and discharge in violation of the implied covenant of good faith and fair dealing.\textsuperscript{15} Tort claims include negligence, wrongful discharge in violation of public policy, tortious interference with contract, invasion of privacy and defamation, assault and battery, and intentional infliction of emotional distress, perhaps even a tort for disability discrimination itself.\textsuperscript{16} These claims may prevail when statutory relief under the ADA fails because of the barriers just described. Nevertheless, there are drawbacks and obstacles to asserting common law claims. These include potential preemption by workers’ compensation and state antidiscrimination statutes, difficulties with asserting vicarious liability, and an unreconstructed, minimalist view of common law duties among some judges.\textsuperscript{17} It is necessary to examine how to surmount these barriers to liability.

Several prominent scholars have been intrigued by the intersection of disability rights and the common law. With characteristic prescience, Jacobus tenBroek wrote forty-five years ago of tort law’s impact on people with disabilities’ “right to live in the world.”\textsuperscript{18} His goal was applying law—and reforming law—to promote integration of people with disabilities as equals in society at large.\textsuperscript{19} His torts article paid special attention to issues such as the right to have service animals in places of public accommodations,\textsuperscript{20} removal and

\textsuperscript{14} Pub. L. No. 110-325, 122 Stat. 3553 (2008) (codified in scattered sections of 42 U.S.C.). See generally Alex B. Long, Introducing the New and Improved Americans with Disabilities Act: Assessing the ADA Amendments Act of 2008, 103 NW. U. L. REV. COLLOQUIY 217 (2008), \url{http://www.law.northwestern.edu/lawreview/colloquy/2008/44} (describing expansion of coverage and other provisions in ADAAA). In the ADAAA, Congress disapproved Sutton v. United Air Lines, Inc., 527 U.S. 471, 482 (1999) (holding that impairments must be evaluated in their mitigated state in determining if an individual has a disability), and Toyota Motor Manufacturing v. Williams, 534 U.S. 184, 197 (2002) (holding that terms of disability definition must be strictly construed), § 2(b)(2)–(5); adopted the rule that impairments must be evaluated in a state not mitigated by medication, appliances, or bodily systems (except, in general, ordinary eyeglasses), § 3(4)(E); provided that major life activities whose substantial impairment entails coverage include major internal bodily systems and functions, § 3(2); and established that persons covered by virtue of being regarded as having an impairment need not be perceived to have an impairment that limits a major life activity, § 6(a)(1).

\textsuperscript{15} See infra Part IV.B.

\textsuperscript{16} See infra Part III.B.

\textsuperscript{17} See infra Part IV.B.


\textsuperscript{19} See id. at 843–52.

\textsuperscript{20} See id. at 852–58 (“The Rights of Dogs and the Rights of Men”).
prevention of architectural barriers, application of the reasonable person standard in connection with negligence and contributory negligence, common carrier liability, and enactment and interpretation of white cane laws. Further development of common law and disability discrimination issues came from Adam Milani, who in 1999 tied the evolution of tort law to the effort to guarantee civil rights for people with disabilities—an effort that led to landmark antidiscrimination legislation such as the ADA. Professor Milani sought to evaluate whether the social movement and statutory development had led to changes in tort law that promoted integration and equality. He found continuing deficiencies in areas such as the application of the reasonable person standard and accommodations from common carriers.

These pathbreaking works on torts and disability did not light on any number of the more plausible common law remedies for discriminatory conduct, however. They did not consider liability for breaches of promises to accommodate, or good-faith and fair-dealing claims for failure to provide equal treatment in the workplace, nor suits over invasion of privacy, assault and battery, and intentional infliction of emotional distress, when employers, operators of public accommodations, or others treat people with disabilities unfairly. Notably, Professor Frank Ravitch pointed out the availability of common law relief for discrimination in his early article on disability harassment, and some other work, including my own, takes up common law remedies for harassment. But these analyses center on intentional infliction claims seeking relief for the creation and maintenance of hostile environments; they do little to expand the discussion to

---

21 See id. at 859–63.
22 See id. at 863–83, 896–902.
23 See id. at 883–96.
24 See id. at 902–11. Typically, these laws free a person who is blind or partially sighted from being charged with negligence if the person is leading a service dog or carrying a white cane, and they make a driver who runs into the person negligent as a matter of law. See id. at 902.
26 See id. at 328, 416–17.
27 Some additional sources that address particular common law remedies will be discussed in the context of the respective causes of action.
29 See WEBER, supra note 5, at 98–121.
contract and other tort ideas, or to other forms of discrimination. Many prominent sources discuss common law remedies for race and sex discrimination in considerable detail, yet there has been surprisingly little extension of those ideas to disability discrimination. This Article seeks to fill the gap, and to set out the need for and potential of common law claims addressing disability discrimination.

Part I of this Article takes up theoretical reasons to develop common law causes of action for disability discrimination. Part II focuses on practical reasons plaintiffs might be interested in pursuing those claims. Part III looks at particular contract and tort remedies for conduct that unfairly discriminates against people with disabilities. Part IV discusses how these remedies might succeed where statutory ones fail, and further considers obstacles to common law liability and how they might be overcome.

I. THEORETICAL REASONS FOR LOOKING TO COMMON LAW REMEDIES

Looking at the topic in a theoretical light, there are five reasons to develop common law claims that stand apart from the ADA and other statutory remedies. First, the common law exists independently of the statutory regime, and if common law claims apply in situations where Congress or a state legislature has also seen fit to add statutory provisions, the courts should enforce the common law whenever the claim is justified, unless the legislative body affirmatively acted to

preempt it. Stating the point in this fashion may exaggerate the independence of common law and statutory law, when in fact the two exist in a reciprocal relationship. Statutes sometimes codify the common law, and in the modern era, courts interpreting the common law often draw analogies from statutory provisions. In particular, statutory race and sex harassment claims may be said to grow out of tort claims for intentional infliction of emotional distress as well as assault and battery, although early advocates of statutory harassment claims viewed these causes of action as inadequate in application. After the judicial recognition of harassment liability under Title VII, many courts have revitalized intentional-infliction law insofar as it relates to racially and sexually motivated misconduct, applying ideas drawn from the Title VII harassment cases. If courts come to appreciate the applicability of common law claims in disability discrimination cases, then continuing statutory development may influence common law interpretation. For example, one of the earlier cases recognizing an ADA claim for hostile environment discrimination involved a public school student who was repeatedly humiliated by her teacher; the appeals court took much of the reasoning it used to support the ADA cause of action and applied it to overturn the dismissal of an intentional infliction of emotional distress claim under Virginia law.

Second, although it is correct that some of the restrictions on the ADA are due to political compromises, it is also true that the ADA was never intended to be an

---


33 See, e.g., CATHARINE A. MACKINNON, SEXUAL HARASSMENT OF WORKING WOMEN 173 (1979) (“To treat [sexual harassment] as a tort is less simply incorrect than inadequate.”).


36 See, e.g., 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, 518–19 (1991) (“One discordant note in the ADA statutory scheme is the presence of several provisions that expressly exempt certain types of enterprises from the obligations imposed by the Act or that exclude certain classes of individuals with disabilities from the protection of the ADA . . . . None of these exemptions and exceptions was part of the bill as originally introduced; each was added to the ADA with the agreement of the bill’s sponsors in the name of compromise.”); Timothy M. Cook, The Americans with Disabilities Act: The Move to Integration, 64 TEMPLE L. REV. 393, 462 (1991) (describing provisions applicable to state and local government, stating, “As a statutory compromise, Congress
exclusive remedy or to preempt state law, either statutory or common. The statute says so explicitly. 37 By creating a one-way antidiscrimination ratchet—the ADA provides a lower limit on what covered entities must do, but not an upper one—the federal law effectively invites states to impose greater obligations on employers and others if legislatures or state courts see fit to do so.

Third, just as occurred with the recent passage of the ADAAA, the statutes of some states and foreign jurisdictions may provide ideas for making the ADA more effective. 38 So too the development of common law in the states may provide ideas for further improving the federal statute. If positive effects stem from expanded coverage and enhanced relief under common law provisions, those developments may inspire federal lawmakers when they consider the next generation of disability discrimination statutes. 39 The “laboratories of democracy” 40 point has been made so often as to become a cliché, but the fact remains that well received legal developments at the state level can serve as a model for federal legislation. 41

Fourth, common law remedies reflect and advance the social judgment that discrimination is a moral wrong. Common law causes of action for disability

 adopted the weaker standard permitting waivers for undue burdens and fundamental alterations of programs, but only for those portions of the ADA rules governing ‘program accessibility, existing facilities,’ and ‘communications,’ (i.e., architectural and communications barriers).”).

37 42 U.S.C. § 12201(b) (2006) (“Nothing in this chapter shall be construed to invalidate or limit the remedies, rights, and procedures of any Federal law or law of any State or political subdivision of any State or jurisdiction that provides greater or equal protection for the rights of individuals with disabilities than are afforded by this chapter.”).

38 For example, the 2008 ADAAA includes a provision similar to that found in the British Disability Discrimination Act of 1995, which provides broad coverage of people with disabilities, but has an exception for people whose sole disabling condition is the need for ordinary eyeglasses. Compare Disability Discrimination Act, 1995, c. 50, § 1, sch. 1, ¶ 6(3)(a) (Eng.) (eyeglasses exception), with Pub. L. No. 110-325, § 3(4)(E)(i)(I), 122 Stat. 3553, 3556 (2008) (ordinary eyeglasses and contact lenses exception). It may also be argued that interpretations of the reasonable accommodation provision of the ADA would benefit if courts looked to interpretation of parallel provisions in disability discrimination statutes of various states and Canada. See Mark C. Weber, Unreasonable Accommodation and Due Hardship, 62 FLA. L. REV. 1119, 1170–71, 1177 (2010).

39 Of course, a similar point may be made regarding state lawmakers.

40 New State Ice Co. v. Liebmann, 285 U.S. 262, 311 (1932) (Brandeis, J., dissenting) (“It is one of the happy incidents of the federal system that a single courageous State may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country.”).

41 In the most celebrated recent example, the 2006 Massachusetts health care reform led to the national one in 2010. See, e.g., Beth Waldman, Perspective, Massachusetts Health Care Reform, 11 HEALTH & HUMAN RTS., no. 2, 2010, at 1, 5–6, available at http://hhrjournal.org/blog/perspectives/ massachusetts-health-care-reform/ (comparing Massachusetts provisions to proposals then in Congress). However, the significance of examples of state legislation inspiring federal legislation should not obscure the possibility of state common law developments influencing federal legislation or interpretations of federal statutes.
discrimination are no different in this respect than common law causes of action for other wrongful conduct, and should be just as freely available. Social morality is the basis of much of judge-made law, and the law in turn reflects prevailing moral norms. Of course, limits exist on what the common law can or should accomplish in enforcing morality, but that does not undermine the essential connection. Tort law, in particular, protects human dignity. Disability discrimination, like other unjustified discrimination based on status, is an affront to human dignity. Liability for intentional infliction of emotional distress is especially closely tied to social judgments about the moral offensiveness of the defendant’s conduct, so it is more than apt to look at that tort for liability for discriminatory conduct. The status-based vulnerability of an individual as the member of a class or group militates in favor of such a common law remedy.

See Robert C. Post, The Social Foundations of Privacy: Community and Self in the Common Law Tort, 77 CALIF. L. REV. 957, 970 (1989) (“[T]he common law attempts not to search out and articulate first ethical principles . . . but instead to discover and refresh the social norms by which we live.”).


Ehrenreich, supra note 31, at 22 (“The common law of tort poses a promising way to understand and address the many forms of workplace harassment. Modern tort law embraces the concept of ‘dignitary harm,’ a harm that injures ‘personality interests’ rather than one’s physical well-being.” (footnote omitted)). Elsewhere in the same article, she states, “When we think about sexual harassment cases, the relevance of . . . common-law notions of human dignity—and therefore tort causes of action like assault, battery, invasion of privacy, false imprisonment, defamation, and intentional infliction of emotional distress—should be obvious at once.” Id. at 27.

See Chamallas, supra note 31, at 2119 (“The antidiscrimination principle . . . is a widely shared cultural norm that we would expect to see reinforced in private law, particularly in the articulation of duties owed by persons who are in a position to inflict serious harm and to restrict the opportunities and potential of others.”). Professor Chamallas is by no means a lone voice in support of intentional-infliction liability, though a few writers are more guarded in their advocacy. See, e.g., Gergen, supra note 3, at 1694 (“I think it is self-evident why we might want to compensate employees and punish firms or their agents through a tort remedy in extreme cases of cruelty or slander, if the torts could be limited to those extreme cases.”).

The Second Restatement expresses this point with respect to intentional infliction of emotional distress, albeit in condescending language: “The extreme and outrageous character of the conduct may arise from the actor’s knowledge that the other is peculiarly susceptible to emotional distress, by reason of some physical or mental condition or peculiarity.” RESTATEMENT (SECOND) OF TORTS § 46 cmt. f (1965).
Moreover, causes of action such as intentional infliction of emotional distress apply in cases of race and sex harassment.\textsuperscript{48}

Some authorities are skeptical of the use of tort law to address what they view as group- or class-based violations of civil rights, but this skepticism derives more from the historical inadequacy of the common law in addressing matters such as harassment than from the present-day potential of the common law.\textsuperscript{49} Pursuing common law claims does not entail abandoning statutory civil rights claims that are explicitly based on group- or class-related discrimination, such as ADA or state human rights act causes of action.\textsuperscript{50} The remedies can and should coexist, and either or both should be available when the proper interpretation of common law principles or statutory terms applies to the case at hand.

Fifth, the assertion of tort and contract rights in the courts fits in with leading ideas about civil rights in relation to disability. The central idea of what is often termed the social model of disability is that physical or mental conditions do not by themselves disable; instead, disability arises from a dynamic between those conditions and the environment.\textsuperscript{51} In the most commonly used example, a person who uses a wheelchair for mobility is not disabled but for curbs, steps, and, frequently, attitudes that prevent full participation in society.\textsuperscript{52} This model has been criticized,\textsuperscript{53} but it remains the fundamental insight behind the ADA’s use of

\textsuperscript{48} See, e.g., Alcorn v. Anbro Eng’t, 468 P.2d 216, 217–19 (Cal. 1970) (upholding claim for intentional infliction of emotional distress on basis of supervisor’s racial insults and firing of claimant); Maksimovic v. Tsogalis, 687 N.E.2d 21, 22 (Ill. 1997) (permitting action for assault, battery and false imprisonment in case alleging sex harassment in workplace).

\textsuperscript{49} See MacKinnon, supra note 33, at 173.

\textsuperscript{50} This point applies as well to tort actions based on sexual harassment. Ehrenreich, supra note 31, at 54 (noting continuing availability of Title VII to address group-based harm inflicted by harassment and further noting that context of sex discrimination could be raised in tort actions when no statutory claim is asserted).

\textsuperscript{51} See Paula E. Berg, Ill/Legal: Interrogating the Meaning and Function of the Category of Disability in Antidiscrimination Law, 18 YALE L. & POL’Y REV. 1, 9 (1999) (“This social-political model rejects the premise of the moral and biomedical perspectives that disability is inherent within the individual. . . . [I]t understands disability as contextual and relational. . . . [The model] understands as a broader social construct reflecting society’s dominant ideology and cultural assumptions. While it acknowledges the existence of biologically based differences, the social-political model locates the meaning of these differences—and the individual’s experience of them as burdensome—in society’s stigmatizing attitudes and biased structures rather than in the individual.” (footnotes omitted)).

\textsuperscript{52} See, e.g., Mary Crossley, The Disability Kaleidoscope, 74 NOTRE DAME L. REV. 621, 658–59 (1999).

civil rights law to compel change in the physical surroundings, workplace routines, and pervasive attitudes that limit the opportunities of people with disabilities.\textsuperscript{54} Consistent with civil rights ideas of using legal tools to force society to change to honor the rights of outsiders and treat them equally,\textsuperscript{55} so too common law remedies may provide incentives for employers and others to make expenditures and adjust standard practices to remove environmental and attitudinal barriers. In fact, the whole universe of judiciably enforceable rights and duties is part of the social environment, something that should itself adapt to facilitate full participation on a plane of equality for people who have disabling conditions.

Common law approaches may present some risks to the promotion of disability rights ideas. One corollary of the civil rights model is that people with disabilities need to take the role of agents rather than passive recipients of social interventions.\textsuperscript{56} It does not help to cast them as helpless victims of others’ conduct.\textsuperscript{57} Nevertheless, pursuing claims through the legal system is one means of taking control of one’s environment, and so the assertion of judicial remedies empowers individuals to take a role in creating the environment in which they live.\textsuperscript{58} Moreover, using the judicial system helps place individual accounts of discriminatory conduct into the broader context of how the society affects people with disabilities, opening possibilities for social change.\textsuperscript{59}

\textsuperscript{54} See, e.g., Wendy Hensel, \textit{The Disabling Impact of Wrongful Birth and Wrongful Life Actions}, 40 Harv. C.R.-C.L. L. Rev. 141, 150 & n.48 (2005) (“[S]ome scholars have credited the political awareness engendered by the minority model for the passage of the Americans with Disabilities Act and comparable civil rights legislation.”) (collecting authorities).

\textsuperscript{55} \textsc{Ruth O’brien, Crippled Justice: The History of Modern Disability Policy in the Workplace} 207–21 (2001) (contrasting rehabilitation emphasis of traditional approaches to disability with legal emphasis of civil rights approach).

\textsuperscript{56} The title of James Charlton’s book on disability discrimination throughout the world is particularly revealing in this regard: \textsc{James I. Charlton, Nothing About Us Without Us} (1998). So too is the title of the leading journalistic account of the social and political dynamics behind the passage of the ADA: \textsc{Joseph P. Shapiro, No Pity: People with Disabilities Forging a New Civil Rights Movement} (1993).

\textsuperscript{57} Martha Minow, \textit{Surviving Victim Talk}, 40 UCLA L. Rev. 1411, 1428–29 (1993) (noting that victimhood generates sympathy and attention but may diminish sense of possibility and power and capacity for choice and action).

\textsuperscript{58} There is a double edge to this reality, however. Lawyers may not always be fully responsive to the directions and interests of their clients with disabilities, see Stanley S. Herr, \textit{Representation of Clients with Disabilities: Issues of Ethics and Control}, 17 N.Y.U. Rev. L. & Soc. Change 609, 633–34 (1989–90), and the lawyers who handle disability rights cases, even cases that reach the highest levels, may not be the most strategically minded or experienced of advocates, see Stein et al., \textit{supra} note 11, at 1662–63.

\textsuperscript{59} \textit{Cf.} Minow, \textit{supra} note 57, at 1437 (noting need to connect victims’ individual stories with larger understandings of social structures and histories).
Professor Laura Rovner has written eloquently about how the process of litigating a disability discrimination case may undermine important disability rights ideas about equality and independent agency.\textsuperscript{60} As she notes, first to prevail on the merits the plaintiff must embrace the identity of having a disability, with all the stereotypes that entails, and second, to obtain significant damages the plaintiff may have to adopt an identity as a victim for whom the jury will feel pity.\textsuperscript{61} Claims litigated under the common law may be better than statutory actions with regard to Rovner’s concerns. For example, in many tort and some contract cases there will be no need to make a specific showing of disability because disability is in itself not an element of the cause of action.\textsuperscript{62} Nevertheless, as with statutory claims, individuals who have been discriminated against will need to present themselves as wronged and deserving compensation without presenting themselves as abject figures. In keeping with the ideas behind the social model of disability, crucial to this task will be an emphasis on the defendant’s conduct in creating the harm to plaintiff, rather than the role of the disability itself.\textsuperscript{63} The mission of the lawyer will be to reinforce the principles behind disability rights by inspiring outrage against the discriminator based on the discriminatory conduct rather than pity for the plaintiff based on the disability.\textsuperscript{64}


\textsuperscript{61} Id. at 251–52.

\textsuperscript{62} See infra Part III.B.6 (discussing intentional-infliction cause of action). It might also be argued that the taking on of an identity as a person with a disability may itself be empowering, and may clothe an individual with the combined strength of a movement of people asserting rights. See, e.g., Rod Michalko, \textit{The Difference That Disability Makes} 70 (2002) (“Coming out [as disabled] is a political matter and not a psychological one . . . . Coming out as disabled implies the necessity of reconnecting disability and identity.”).

\textsuperscript{63} Professor Vidmar has commented that in medical malpractice cases, some unusually high verdicts appear to stem from jury anger at the outrageousness of the defendant’s negligence rather than from any other source. Neil Vidmar, \textit{The American Civil Jury for Ausländer (Foreigners)}, 13 Duke J. Comp. & Int’l L. 95, 119 (2003) (“Some highly suspicious verdicts involving multi-million dollar compensatory awards in medical malpractice cases are documented in another article. Although the alleged behavior of the medical professionals in these cases presents grounds for jurors . . . to feel anger towards the defendants, the huge compensatory damage awards are difficult to defend.” (footnote omitted)).

\textsuperscript{64} As Professor Rovner notes, for at least some claimants, vindication of legal rights and personal integrity is far more important than the size of the monetary award. See Rovner, \textit{supra} note 60, at 315. The same, of course, is true for many a plaintiff whose privacy has been invaded or who has been defamed. Nevertheless, obtaining money to pay the lawyer in a common law case remains a problem in the absence of a cash judgment and contingency fee.
II. PRACTICAL DEFICIENCIES OF THE FEDERAL STATUTORIAL REGIME FOR CLAIMANTS

Mapping the practical deficiencies of the ADA’s protections against disability discrimination entails comparing the very real and broad guarantees against discrimination in the statute with the limits that the legislation embodies. Those limits suggest areas in which common law causes of action could be particularly useful for disability discrimination claimants.

A. What the ADA Provides

The Americans with Disabilities Act contains general provisions and five titles. The general provisions include findings and statutory purposes, as well as definitions.\(^{65}\) Title I of the statute forbids employment discrimination;\(^{66}\) Title II bans discrimination in state and local government services;\(^{67}\) Title III bars discrimination by privately-operated places of public accommodation; Title IV requires telecommunications providers to enhance services for people who are deaf, hearing impaired, or speech impaired;\(^{68}\) and Title V sets out miscellaneous matters such as abrogation of state immunities, prohibitions on retaliation and coercion, authorization of attorneys’ fees, and other provisions.\(^{69}\)

The ADA as written has broad coverage, defining disability to include “a physical or mental impairment that substantially limits one or more major life activities of [an] individual; a record of such an impairment; or being regarded as having such an impairment.”\(^{70}\) The ADA is a comprehensive ban on discrimination. It requires reasonable accommodations for employees with disabilities,\(^{71}\) forbids employment practices that impose disparate impact unless they are job-related and consistent with business necessity,\(^{72}\) demands that merchants make reasonable modifications to operations,\(^{73}\) and imposes a wide range of guarantees of accessible construction.\(^{74}\) The law further authorizes regulations that, among other things, force state and local governments to make reasonable modifications in their policies and practices.\(^{75}\) Other federal laws ban

\(^{66}\) Id. §§ 12111–12117.
\(^{68}\) 47 U.S.C. § 225.
\(^{69}\) 42 U.S.C. §§ 12201–12213.
\(^{70}\) 42 U.S.C.A. § 12102(1).
\(^{71}\) 42 U.S.C. § 12112(b)(5)(A).
\(^{72}\) Id. § 12112(b)(6).
\(^{73}\) Id. § 12182(b)(2)(A)(ii).
\(^{74}\) Id. § 12183; see also 28 C.F.R. § 35.151 (2011) (governing new construction and alteration in state and local public entities’ facilities).
\(^{75}\) See 42 U.S.C. § 12134 (authorizing regulation found at 28 C.F.R. § 35.130(b)(7)).
disability discrimination by federal grantees, in public and private housing, in air travel, and by federal government agencies.

B. What the ADA Lacks

The strength and breadth of the ADA’s guarantees do not obscure the limits of the legislative regime. These limits include ADA coverage restrictions based on employer size in Title I cases; damages limits in Title I; unclear coverage of independent contractors and volunteers under Title I; restrictions on liability of individuals and enterprises that are not employers under Title I or covered entities under the other titles; limitations and exhaustion requirements; potential difficulties with coverage of plaintiffs under the statutory definition of disability; potential causation difficulties; lack of coverage of entities operated by religious organizations under Title III; lack of damages under Title III; and difficulties with getting to trial in actions in federal court.

80 State statutory provisions may help fill some of these gaps and compensate for some of these weaknesses, for those statutes need not be construed identically with the ADA. See, e.g., Victor v. State, 4 A.3d 126, 148 (N.J. 2010) (permitting claim for denial of reasonable accommodation or failure to engage in interactive process of determining accommodations under state law without showing of adverse employment consequences despite contrary federal ADA interpretations). However, many state disability discrimination statutes simply track the ADA and receive similar judicial and administrative constructions. See, e.g., Kelly v. Drexel Univ., 94 F.3d 102, 105 (3d Cir. 1996) (“While the Pennsylvania courts are not bound in their interpretations of Pennsylvania law by federal interpretations of parallel provisions in Title VII, the ADA, or the ADEA, its courts nevertheless generally interpret the PHRA [Pennsylvania Human Right Act] in accord with its federal counterparts.” (citations omitted)); Tourville v. Securex, Inc., 769 So. 2d 491, 492 n.1 (Fla. Dist. Ct. App. 2000) (“The Florida Civil Rights Act should be construed in conformity with the Rehabilitation Act and the Americans with Disabilities Act.” (citations omitted) (internal quotation marks omitted)); City of Columbus Civil Serv. Comm’n v. McGlone, 697 N.E.2d 204, 206–07 (Ohio 1998) (“We can look to regulations and cases interpreting the federal Act for guidance in our interpretation of Ohio law.”); Thomann v. Lakes Reg’l MHMR Ctr., 162 S.W.3d 788, 796 (Tex. Ct. App. 2005) (“[W]e are guided by federal law in construing [Labor Code] chapter 21.”). Although the California Fair Employment and Housing Act (FEHA) is in some respects more protective of claimants than the ADA in covering individuals whose impairments limit major life activities, rather than only those whose impairments substantially limit major life activities, CAL. GOV’T CODE § 12926.1(c) (West 2011), when “the particular provision in question in the FEHA is similar to the one in the ADA, the courts have looked to decisions and regulations interpreting the ADA to guide construction and application of the FEHA,” Hastings v. Dep’t of Corr., 2 Cal. Rptr. 3d 329, 335 n.12 (Cal. Ct. App. 2003).
1. Coverage Restrictions Based on Employer Size

Perhaps because of concerns over the reach of Article I Commerce Clause power, Congress restricted coverage of the ADA Title I to employers with fifteen or more employees, which is the same threshold that applies to claims under Title VII of the Civil Rights Act.81 Plaintiffs in race discrimination cases, however, have an additional federal statutory remedy under the Civil Rights Act of 1866, and that provision has no restrictions on the size of the employer.82 Thus, whereas race discrimination claimants can proceed against smaller businesses, at least up to the actual limit of congressional commerce authority, ADA claimants are barred from doing so. Well more than ten million Americans are employed by small businesses that are exempt from the ADA.83

2. Damages Limits in Employment Cases

Plaintiffs in ADA Title I cases face restrictions on damages for pain and suffering and punitive damages, based on the size of the employer. Thus an employer with 100 employees is liable for a maximum $50,000 in damages, an employer with 200 has a maximum exposure of $100,000, the employer with 500 pays no more than $200,000, and even the largest employers (501 employees and more) are liable for no more than $300,000.84 All employers enjoy a safe harbor in cases over the denial of reasonable accommodations. If the employer satisfies the obligation of engaging in good faith efforts, in consultation with the employee, to identify and make a reasonable accommodation, the employer is not liable for damages.85 If the case is based on disparate impact, there are no damages at all.

81 42 U.S.C. § 12111(5)(A); cf. id. § 2000e(b) (defining covered employers as those with fifteen or more employees each working day in each of twenty or more calendar weeks, definition identical to ADA Title I provision). The Supreme Court sustained the constitutionality of the Civil Rights Act of 1964 as an exercise of congressional power over interstate commerce. Katzenbach v. McClung, 379 U.S. 294 (1964) (addressing Title II of the Civil Rights Act of 1964). The ADA expresses the purpose “to invoke the sweep of congressional authority, including the power to enforce the fourteenth amendment and to regulate commerce . . . .” 42 U.S.C. § 12101(b)(4).


83 The Census Bureau maintains statistics on people employed by firms with given numbers of employees, but the categories are not fully congruent with the fifteen-employee threshold of the ADA. As of 2008, 6,086,291 people worked at firms of one to four employees; 6,878,051 at firms of five to nine employees, and 8,497,391 at firms of 10 to 19 employees. Firms of all sizes had 120,903,551 employees in total. Statistics About Business Size (Including Small Business), U.S. CENSUS BUREAU, http://www.census.gov/epcd/www/smallbus.html (last visited Feb. 6, 2012).


85 Id. § 1981a(a)(3).
only equitable relief such as back pay and orders to not discriminate in the future.\(^{86}\) In addition, a few courts have ruled that in cases concerning retaliation, coercion, and harassment in employment they may award nothing more than equitable relief.\(^{87}\)

3. Unclear Coverage of Independent Contractors and Volunteers

The ADA’s employment title does not mention independent contractors, and several courts have held that those in an independent contractor relationship at their workplace lack the protections of the statute.\(^{88}\) Neither Title I’s employment provisions nor Title III’s public accommodations provisions give clear coverage to volunteers, leaving volunteers without any guarantee of accommodation or fair treatment.\(^{89}\)

4. Restrictions on Individual Liability

Moreover, individuals and other potential defendants who are not among the covered entities listed in the ADA are, generally speaking, not liable to ADA

\(^{86}\) Id. § 1981a(a)(2).

\(^{87}\) See, e.g., Alvarado v. Cajun Operating Co., 588 F.3d 1261, 1265 (9th Cir. 2009); Kramer v. Banc of Am. Secs., 355 F.3d 961, 968 (7th Cir. 2004). This position has been challenged. See Mark C. Weber, Workplace Harassment Claims Under the Americans with Disabilities Act: A New Interpretation, 14 STAN. L. & POL’Y REV. 241, 262 (2003).


claimants. This may create difficulties if the covered entity has shallow pockets or is out of business altogether. It is also apt to produce a situation where there is nobody to sue when some individual—say a coemployee or supervisor—acts contrary to employer policy but nevertheless harms the plaintiff using power derived from the employer. By contrast, under the common law, relief may exist against individuals for coworker or other peer misconduct even when there is no knowledge or other involvement by the school, employer, or other organization subject to the ADA.

5. Statute of Limitations and Exhaustion Requirements

In ADA employment cases, claimants frequently are barred by a short statute of limitations, and may also find their claims undone by the failure to have filed them with the EEOC or appropriate state agency before bringing them to court. These problems are more serious than might appear, because the limitation provisions are not well known and are difficult to interpret, and because courts have applied the exhaustion provision with great severity, requiring, for example, that the charge before the EEOC provide notice not just of the fact that the employee has a claim, but what the specific claim is. Thus an employee who asserted discrimination in the EEOC filing will not be able to bring a claim for retaliation as well if it was not identified in the charge, even when the employee drafted the charge without assistance of counsel.

---

90 This is not true for all provisions of the ADA, for example, the antiretaliation and harassment provision, § 12203. See LaManque v. Mass. Dep’t of Emp’t & Training, 3 F. Supp. 2d 83, 90 (D. Mass. 1998) (upholding individual liability).
92 In § 12117(a), Title I of the ADA incorporates the remedies of Title VII of the Civil Rights Act, including the limitations provision, § 2000e-5(f)(1).
93 See Maynard v. Pneumatic Prods. Corp., 256 F.3d 1259, 1264 (11th Cir. 2001) (applying requirement of timely filing of charge with EEOC or designated state administrative agency).
94 A leading treatise on disability discrimination law devotes an entire chapter to statutes of limitations. Peter Blanck et al., Disability Civil Rights Law and Policy ch. 18 (2004).
96 See Brown v. City of Cleveland, 294 F. App’x 226, 233–34 (6th Cir. 2008) (dismissing retaliation and hostile environment claims for failure to exhaust by inclusion in EEOC charge).
6. Lingering Problems with Protected-Class Coverage?

Unlike some other civil rights laws, with regard to the bulk of its provisions the ADA does not protect everyone, but instead protects only a class of covered individuals. To be covered as a person with a disability, the individual must meet at least one of three criteria: have a physical or mental impairment that substantially limits one or more major life activities of the individual, a record of such an impairment, or be regarded as having such an impairment. Although the ADAAA corrected several restrictive interpretations of the coverage of plaintiffs in ADA cases, problems still linger. The ADAAA represents progress from the claimant’s point of view. It affirms that impairments are to be evaluated in a state unmitigated by medical intervention (excluding ordinary eyeglasses or contact lenses) or “learned behavioral or adaptive neurological modifications.” It further provides an expanded list of major life activities, and says that the definition of disability is to be construed in favor of broad coverage and the phrase “substantially limits” should be interpreted consistently with findings and purposes of the law that disapprove a restrictive Supreme Court holding on the subject. Episodic impairments and those in remission are covered, and to be regarded as having a disability, the individual need not be perceived as having a limit on a major life activity.

But the ADAAA did not eliminate the “substantially limits” language altogether, as had been proposed and as California’s disability discrimination

---

97 See, e.g., 42 U.S.C.A. § 12112(b)(5)(A) (West 2011) (forbidding “not making reasonable accommodations to the known physical or mental limitations of an otherwise qualified individual with a disability who is an applicant or employee . . . .”). Some provisions of the ADA, such as the medical examinations and retaliation sections, apply more broadly. See, e.g., 42 U.S.C. § 12112(d) (banning various classes of medical inquiries and examinations); id. § 12203 (banning retaliation and coercion). The ADAAA changed the general prohibition on discrimination and effectively altered some specific prohibitions to no longer explicitly require the claimant to have a disability. See, e.g., 42 U.S.C.A. § 12112(a), (b). How courts will treat this modification remains unknown.


101 Id. at 122 Stat. 3555.

102 Id. (appearing to reference id. at § (2)(b)(2)–(3), 122 Stat. 3554 (referring to Toyota Motor Mfg., Ky. v. Williams, 534 U.S. 184 (2002))).

103 Id. at 122 Stat. 3556.

104 Id. at 122 Stat. 3555.

This requirement forces courts and juries to make difficult and ultimately arbitrary judgments about individuals, and it reinforces the misguided idea of a deserving class of people with disabilities as opposed to others with lesser impairments who supposedly reap windfall benefits from the statute. Moreover, courts have found the statute not to be retroactive to conduct occurring before its January 1, 2009 effective date.

7. Problems with Proving Causation Based on Disability

Even if coverage problems eventually disappear in the wake of the ADAAA, in some cases plaintiffs will still have a challenge convincing the judge or jury that the discrimination is because of disability. A number of courts have made this showing particularly difficult by requiring causal links between a requested accommodation and the major life activity that the claimant’s disability affects, or between the accommodation and the impairment as narrowly viewed. These courts defy the Supreme Court, which held in *US Airways, Inc. v. Barnett* that these causal links were not to be demanded, over a dissenting opinion that insisted they should be required. Nevertheless, the resistance of judges to

---

106 CAL. GOV’T CODE § 12926.1(c) (West 2011).
107 Anderson, supra note 105, at 99–102. Moreover, although some of these individuals will be covered under the regarded-as prong of the law, particularly now that the regarded-as standard has been relaxed, they will not be entitled to reasonable accommodations if that is the sole basis of their coverage. Id. at 126–28. The ADAAA actually reinforces the latter point. Pub. L. 110-325, § 6(a), 122 Stat. 3558 (providing that entities subject to ADA need not provide reasonable accommodations or reasonable modifications to persons who meet definition of disability solely under regarded-as provision); see Jeannette Cox, *Crossroads and Signposts: The ADA Amendments Act of 2008*, 85 Ind. L.J. 187, 189, 210–14 (2010) (“The amendments also strengthen the connection between the severity of endogenous limitations and the right to ADA accommodations. Accordingly, the amendments may unwittingly underwrite the assumption that the ADA is not a traditional civil rights statute but is instead a welfare benefits statute that confers special benefits to compensate for endogenous biological limitations.”). In recent writing, Professor Anderson has noted that adoption of an impairment-only definition of disability, one that lacks any inquiry into limits on activities, could trigger a backlash that would lead to restrictive readings of other provisions of the law. Cheryl L. Anderson, *Ideological Dissonance, Disability Backlash, and the ADA Amendments Act*, 55 Wayne L. Rev. 1267, 1272–73 (2009).
108 E.g., Kemp v. Holder, 610 F.3d 231, 236 (5th Cir. 2010).
111 Id. at 397.
112 Id. at 412 (Scalia, J., dissenting).
applying the flexible causal requirement embodied in the statute makes reliance on
the ADA riskier than it ought to be.

Moreover, a number of courts have required that disability be the but-for
cause of the adverse action that the claimant challenges, despite the efforts of
Congress in the Civil Rights Act of 1991 to overturn such an interpretation of
Title VII. These courts follow an interpretation that the Supreme Court imposed
on the Age Discrimination in Employment Act (“ADEA”) in Gross v. FBL
Financial Services, Inc. Like ADEA, with respect to employment the ADA
incorporates remedial language originally found in Title VII, but it does not
explicitly reference the amended provision of Title VII that was adopted to correct
the but-for interpretation in multiple causation cases. The application of this
reasoning to the ADA, however, flies in the face of the decision by Congress in the
ADA to refrain from copying the language of section 504 of the Rehabilitation Act
that bars discrimination against a person with a disability “solely by reason of her
or his disability.” Congress took this step specifically to avoid eliminating
liability when another efficient reason supported the adverse action. For
example, if a person is terminated both because he or she needs an accommodation
and because of race, the individual would not be entitled to relief because either
ground, standing alone, would support the decision. The House Committee

---

113 Serwatka v. Rockwell Automation, Inc., 591 F.3d 957, 962 (7th Cir. 2010); Ross
v. Indep. Living Res. of Contra Costa Cnty., No. C08-00854 TEH, 2010 WL 2898773, at
*6 (N.D. Cal. July 21, 2010). Similarly, when the plaintiff in an ADA case alleges
discrimination on the basis of his or her association with someone who has a disability,
some courts have required a showing that “the disability of the relative or associate was a
determining factor in the employer’s decision.” Den Hartog v. Wasatch Acad., 129 F.3d
1076, 1085 (10th Cir. 1997); see Dewitt v. Proctor Hosp., 517 F.3d 944, 952 (7th Cir.
Cir. 2004). The Serwatka decision held open the possibility that the ADA Amendments
Act’s adoption of language forbidding discrimination on the basis of disability might alter
the outcome in cases after the effective date. 591 F.3d at 961 n.1.


Mixed Motive Discrimination, 68 N.C. L. REV. 495, 507–23 (1990) (criticizing decision on
basis of multiple-causation principles applicable to tort and contract cases and
philosophical ideas about causation).


303, 368.

120 See id. (“A literal reliance on the phrase ‘solely by reason of his or her handicap’
leads to absurd results. For example, assume that an employee is black and has a disability
and that he needs a reasonable accommodation that, if provided, will enable him to perform
the job for which he is applying. He is a qualified applicant. Nevertheless, the employer
rejects the applicant because he is black and because he has a disability. In this case, the
specifically rejected that interpretation in the ADA.\(^{121}\) Despite the weakness of the courts’ position, once again, reliance on the ADA may leave claimants without remedy. Although some common law claims may entail difficult causation problems, others, such as intentional infliction of emotional distress, do not require plaintiffs to prove causation based on disability.\(^{122}\)

8. Exemption in Title III for Religious Entities

Another limit on the ADA is that public accommodations operated by religious entities are not covered.\(^{123}\) Thus a school or day care center with a religious affiliation need not be accessible, and the discriminatory conduct of its officials will not create liability under the Act except with respect to employment. Even universities that are religiously affiliated are bound by the ADA only with regard to employment; their obligations toward students and other non-employees derive from the fact they receive federal funding, which triggers coverage under section 504 of the Rehabilitation Act.\(^{124}\)

employer did not refuse to hire the individual solely on the basis of his disability—the employer refused to hire him because of his disability and because he was black. Although the applicant might have a claim of race discrimination under title VII of the Civil Rights Act, it could be argued that he would not have a claim under section 504 because the failure to hire was not based solely on his disability and as a result he would not be entitled to a reasonable accommodation.”).\(^{121}\) Id. (“The Committee, by adopting the language used in regulations issued by the executive agencies [which omits the ‘solely by reason of’ language], rejects the result described above.”). See generally Mark C. Weber, Disability Discrimination by State and Local Government: The Relationship Between Section 504 of the Rehabilitation Act and Title II of the Americans with Disabilities Act, 36 WM. & MARY L. REV. 1089, 1111 (1995) (discussing congressional decision to adopt interpretations of section 504 in ADA but to reject sole-causation language).

\(^{123}\) 42 U.S.C. § 12187.

\(^{124}\) After the Civil Rights Act of 1991, Pub. L. No 102-166, 105 Stat. 1071, (codified as amended in scattered sections of 42 U.S.C.), overruled Grove City College v. Bell, 465 U.S. 555 (1984), a university that receives federal financial aid or research money would be hard pressed to deny that it is covered by section 504 in all its operations. 29 U.S.C. § 794 (a)–(b). Section 504’s provisions governing entities such as private universities are similar to, but not exactly the same as, those of the ADA governing comparable public institutions. See Weber, supra note 121, at 1097–1116 (discussing the similarities and differences between Section 504 and Title II of the ADA).
9. Absence of Monetary Remedies in Public Accommodations Claims

Under ADA Title III, monetary damages are never available for private causes of action regarding discrimination in places of public accommodation. The ADA took this approach from the public accommodations title of the Civil Rights Act of 1964, but unlike race discrimination claimants covered by that law, people with disabilities who are discriminated against in public accommodations lack the additional remedy in damages provided by 42 U.S.C. § 1981.

Professor Michael Waterstone has amassed extensive empirical evidence showing “that people with disabilities are still at the margins of society in areas covered by [ADA] Titles II and III, [and] these low numbers demonstrate underenforcement of these Titles, rather than lack of a need or desire by potential plaintiffs to bring cases.” He points out that obstacles such as the absence of a damages remedy and the difficulty with obtaining and keeping standing for injunctive relief in the absence of such a remedy deter potential plaintiffs from litigation, leading in turn to noncompliance by entities bound by the law. Since compensatory and punitive damages are the default remedy for actions at common law, these claims may help fill the remedial gap and enhance compliance with nondiscrimination obligations in public accommodations. Of course, it remains to be seen whether awards will be large enough to correct the deficiency in cases not involving severe emotional distress or actual physical injury.

10. Federal Forum-Related Problems

In 1986, the Supreme Court loosened the standards under which federal district courts could grant summary judgment to defendants in civil litigation. See 42 U.S.C. § 12188(a). Damages and civil penalties are available in actions brought by the Attorney General. Id. § 12188(b)(2)(B)–(C).

See id. at § 12188 (provision of ADA Title III incorporating remedies and procedures provision of Civil Rights Act Title II, 42 U.S.C. § 2000a–3(a)).


Waterstone, supra note 11, at 1854 (collecting studies).

Id. at 1870 (“Considering the convergence of no damage remedy and the increasingly doubtful prospects for attorneys’ fees, the low number of Title III cases at the appellate level . . . makes sense. Individual plaintiffs have very few incentives to bring these cases.”). Common law remedies typically do not include attorneys’ fees, of course, but there is the prospect of taking a contingent fee from a damages award.

Id. at 1872.

Id. at 1854. As Professor Waterstone notes, Title III enforcement efforts by the Justice Department have been modest at best. See id. at 1873–74.

See Celotex Corp. v. Catrett, 477 U.S. 317, 322 (1986) (“We think that the position taken by the majority of the Court of Appeals is inconsistent with the standard for summary judgment set forth in Rule 56(c) of the Federal Rules of Civil Procedure. . . . In our view,
Some federal judges seem especially eager to use this license in ADA cases, displaying a remarkable inclination to make up their own minds about the facts and to keep the case away from the jury. The language they use can be revealing. In *Filar v. Board of Education*, a substitute schoolteacher with arthritis who was unable to drive or walk long distances and needed to be placed at a school close to public transportation, asked for an exception to the school board’s roving assignment system for substitutes. The court of appeals affirmed a grant of summary judgment for the defendant, declaring: “[T]he question is whether her requested accommodation was reasonable, and we don’t think it was.” “[A]spects of the request convince us that it was just not reasonable.” Except at the furthest extremes, reasonable accommodation is an issue of fact, not of law. The court failed to consider whether a reasonable jury might have come to a different conclusion about reasonableness of the accommodation.

The plain language of Rule 56(c) mandates the entry of summary judgment, after adequate time for discovery and upon motion, against a party who fails to make a showing sufficient to establish the existence of an element essential to that party’s case, and on which that party will bear the burden of proof at trial.); see also Anderson v. Liberty Lobby, Inc., 477 U.S. 242, 252–56 (1986) (suggesting greater availability of summary judgment in public-figure defamation case); Matsushita Elec. Indus. Co. v. Zenith Radio Corp., 475 U.S. 574, 595–98 (1986) (reinstating entry of summary judgment in factual context of antitrust case).

See Ruth Colker, *The Americans with Disabilities Act: A Windfall for Defendants*, 34 HARV. C.R.-C.L. L. REV. 99, 101 (1999) (“My review of the litigation outcome data—combined with my individualized review of every appellate decision and many of the district court cases decided since the ADA became effective in 1992—leads to the conclusion that district and appellate courts are deploying two strategies that result in markedly pro-defendant outcomes under the ADA. Courts are abusing the summary judgment device and failing to defer to agency guidance in interpreting the ADA.”); see also Sharona Hoffman, *Settling the Matter: Does Title I of the ADA Work?*, 59 ALA. L. REV. 305, 326–29 (2008) (discussing the role of judicial attitudes in accounting for low win rates in ADA employment litigation); Jeffrey A. Van Detta & Dan R. Gallipeau, *Judges and Juries: Why Are So Many ADA Plaintiffs Losing Summary Judgment Motions, and Would They Fare Better Before a Jury? A Response to Professor Colker*, 19 REV. LITIG. 505, 510 (2000) (collecting and analyzing data from “jury research in connection with actual litigation, [that] reveals a general public that is much more enlightened on issues of disability and workplace accommodation than are many employers—and is thus much less likely to produce pro-defense outcomes than current dispositive motion practice,” but attributing failure to reach juries to poor advocacy in litigated cases).

*Filar v. Board of Education*; 526 F.3d 1054 (7th Cir. 2008).

*Id.* at 1059.

*Id.* at 1067–68.

*Id.* at 1067.

*Id.* at 1067–68.


This point applies beyond ADA cases. See Patrick E. Higginbotham, *The Present Plight of the United States District Courts*, 60 DUKE L.J. 745, 746 (2010) (“[T]he motion for summary judgment . . . has displaced the trial as the destination point for litigation. Today it is unlikely that a trial date will ever be set, and rarer still that a trial date will have any meaning to the court and hence to the parties.”). Additional problems with reaching a
State courts sometimes grant summary judgment, but in general state courts display greater willingness to trust juries to decide issues of fact. Accordingly, it may be wise for claimants to base their complaints solely on state common law and statutory grounds in the absence of diversity removal jurisdiction, so that defendants cannot successfully remove the cases to the United States District Court on the basis of federal question jurisdiction.

Jury in cases filed in federal court are also to be expected due to the plausibility test that those courts now apply to pleadings under Bell Atlantic Corp. v. Twombly, 550 U.S. 544 (2007), and Ashcroft v. Iqbal, 129 S. Ct. 1937, 1950 (2009). See Paul D. Carrington, Politics and Civil Procedure Rulemaking: Reflections on Experience, 60 Duke L.J. 597, 648–55 (2010); cf. id. at 645–48 (also criticizing Supreme Court’s expansion of summary judgment in absence of rulemaking).

The plaintiff, of course, remains master of the complaint, free to rely on nonfederal grounds for seeking relief and thus defeat federal question jurisdiction. Holmes Grp., Inc. v. Vornado Air Circulation Sys., Inc., 535 U.S. 826, 831 (2002) (“[S]ince the plaintiff is the master of the complaint, the well-pleaded-complaint rule enables him, by eschewing claims based on federal law, to have the cause heard in state court.” (internal quotation marks and ellipsis omitted)); see, e.g., Reuther v. Shiloh Sch. Dist. No. 85, No. 07-CV-689-JPG, 2008 WL 191195, at *1 (S.D. Ill. Jan. 18, 2008). In Reuther, the court remanded a negligence and willful and wanton conduct action to state court in a case concerning a child with cerebral palsy who received injuries (1) when he was required to practice washing hands but was placed at a non-ADA compliant sink and not strapped in his wheelchair, fell forward and suffered a concussion when striking his head on the sink; (2) broke out in hives after being given a latex ball by an aide, even though the school district was aware of the child being highly allergic to latex; and (3) was forced to remain in a painful position in a medical device even though crying and complaining of pain. The court held that the allegation of breach of duties under the Individuals with Disabilities Education Act and the ADA did not create a federal claim when the parents did not bring a federal claim for compensation but solely alleged state tort law claims. Id.

Diversity removal is more restrictive than federal question removal. See 28 U.S.C. § 1441(b) (2006) (providing for defeat of diversity removal jurisdiction if any properly joined defendant is citizen of state where action is brought).
III. POTENTIAL COMMON LAW CAUSES OF ACTION

Common law claims that relate to tort, particularly intentional infliction of emotional distress, are the most thoroughly developed in the case law. But other tort causes of action, as well as some actions in contract, may apply when employers, merchants, and others discriminate on the basis of disability. Consider first contract suits and then tort liability.

A. Contract

Two causes of action that sound in contract are particularly relevant when discussing how to remedy disability discrimination: breach of promises to accommodate and breach of the implied covenant of good faith and fair dealing in employment contracts.

1. Promises to Accommodate

Defendants may be liable at common law if they make promises to accommodate but fail to follow through on them. Perhaps the best-known example of this type of claim is *Guckenberger v. Boston University*, 144 in which students at the university sued over the failure to provide adequate services and accommodations for students with learning disabilities. 145 The students did not prevail on their most noteworthy claim—that high officials of the university created a hostile environment in violation of the ADA and Section 504 by declaring students with learning disabilities to be a “plague” and making other disparaging public comments about them. 146 But the court nonetheless entered judgment against the university on a pendent claim under Massachusetts law for breach of contract in failing to provide the course substitutions, learning disability support, and acceptance of documentation of disability that the student disability services office had promised the students when they were deciding to attend the school. 147 Similarly, in a case involving employment, a Massachusetts court upheld

145 *Id.* at 311.
146 *Id.* at 312, 315. The court said that although various “comments, viewed objectively, may certainly be offensive to learning-disabled students . . . they are not physically threatening or humiliating.” *Id.* at 315. Further, imposing liability on the basis of university officials’ speeches would “have serious First Amendment implications.” *Id.* at 316.
a contract claim based on the failure to provide the promised accommodation of a delayed employment start date to permit the plaintiff’s recovery from a disabling condition. In Wisconsin, a federal court upheld a state law claim for breach of a partnership agreement when the plaintiff’s law partner withheld contractually due compensation after the plaintiff was forced to work at home because of his sickle-cell disease.

Employment situations also provide the opportunity for claims for breach of contract based on failure to provide accommodations pursuant to promises in an employee handbook. In one case, a customer service agent with a psychological impairment alleged that his employer denied him the reasonable accommodation of transfer to a different supervisor. The court denied summary judgment on the employee’s action for breach of contract based on a representation in the employee manual that it was the employer’s policy to provide reasonable accommodations. Many cases of this type fail, however, due to disclaimers or other barriers to establishing justifiable reliance on promises to accommodate found in student or employee handbooks or manuals.

individual not covered by ADA). The Supreme Court reversed Russell because the court of appeals improperly deferred to the district court’s interpretation of the state law when it should have applied de novo review, Salve Regina Coll. v. Russell, 499 U.S. 225, 233 (1991), but there is no ground to believe that the state courts would have interpreted the state law differently.


Muwonge v. Eisenberg, No. 07-C-0733, 2008 WL 753898, at *10 (E.D. Wis. Mar. 19, 2008). The court also allowed the claim for intentional infliction of emotional distress to proceed. Id. at *6.


Id. at 1065–66; see also Jamison v. Campbell Chain Copper Tool, No. 1:07-CV-0324, 2008 WL 857526, at *4 (M.D. Pa. Mar. 27, 2008) (denying motion to dismiss an action alleging that employee handbooks assured continued employment contingent on satisfactory job performance).

See, e.g., Pacella v. Tufts Univ. Sch. of Dental Med., 66 F. Supp. 2d 234, 241–42 (D. Mass. 1999) (holding in the case of a dental student alleging disability on the basis of an eye condition that the terms of the student handbook were not contractually binding, following state holdings that employee manuals would not be considered part of implied contract with employee when the employer retained the right to unilaterally modify manual’s terms; there was no negotiation over the terms of manual; no special attention was called to the manual by the defendant; and there was no indication plaintiff signed the manual or otherwise manifested assent or acknowledged the terms); Guckenberger v. Boston Univ., 974 F. Supp. 106, 151 & n.38 (D. Mass. 1997) (discussing disclaimer found in Boston University catalogs and brochures and dismissing claims for breach of contract for want of proof of reliance on representations made in materials). When the elements of contract are not met but the worker justifiably relies to her detriment, there may be a remedy in promissory estoppel. See Falk v. U.H.H. Home Servs. Corp., 835 F. Supp. 1078, 1081 (N.D. Ill. 1993) (denying motion to dismiss claim under Illinois law of promissory
2. Implied Covenant of Good Faith and Fair Dealing

The dominant rule in the United States is that, unless otherwise specified, contracts of employment are at will, and either party may terminate the arrangement for any cause or no cause at all, provided that the cause is not an unlawful one. Nevertheless, there remains a nonwaivable implied covenant of good faith and fair dealing in employment contracts, “which includes an agreement by each not to hinder the other's performance under, or to deprive the other of the benefit of, the contract.” In Ward v. Sorrento Lactalis, Inc., a federal district court ruled that a state law claim for breach of the covenant of good faith and fair dealing survived summary judgment on the basis of the same facts that supported claims under the ADA and the Idaho Human Rights Act. The facts were not fully spelled out in the opinion but included that the plaintiff had a back impairment that prevented him from lifting and bending and required multiple surgeries; he was fired immediately after returning from time off to recover from a third back operation without being given a chance to demonstrate he could perform the job.

---


154 RESTATEMENT (THIRD) OF EMP’T § 2.07(a) (Tentative Draft No. 2 (revised), 2009). At present, only a minority of states follow this doctrine and provide a common law remedy in employment cases. In California, for example, there is no independent claim for breach of implied covenant of good faith and fair dealing for at-will employees. Guz v. Bechtel Nat’l Inc., 8 P.3d 1089, 1112 (Cal. 2000) (“To the extent Guz’s implied covenant cause of action seeks to impose limits on Bechtel’s termination rights beyond those to which the parties actually agreed, the claim is invalid. To the extent the implied covenant claim seeks simply to invoke terms to which the parties did agree, it is superfluous. Guz’s remedy, if any, for Bechtel’s alleged violation of its personnel policies depends on proof that they were contract terms to which the parties actually agreed.”); see also Diaz v. Fed. Express Corp., 373 F. Supp. 2d 1034, 1066 (C.D. Cal. 2005) (entering summary judgment against plaintiff on good faith-fair dealing claim). Recognition of a limited cause of action for violation of the duty in the Restatement may promote the acceptance of the tort.


156 See id. at 1195 (citing IDAHO CODE § 67-5901).

B. Tort Claims

Only a fraction of the instances in which people with disabilities are discriminated against will furnish the basis for an action in contract. Causes of action for tort are more likely to apply. Moreover, damages awards for tort are more apt to include items such as pain and suffering or emotional distress in general than are actions in contract; the same is true for punitive awards. As the court said in one contract case involving allegations of disability discrimination, “Generally, tort damages such as punitives or, as here, damages for emotional distress, are recoverable for breach of contract only if the breach involved conduct that was independently tortious and outside the risks contemplated by the contract.”158 Tort claims include negligence, wrongful discharge in violation of public policy, tortious interference with contract, invasion of privacy, defamation, assault, battery, and intentional infliction of emotional distress. There is even the glimmer on the horizon of a separate tort of disability discrimination itself.159

1. Negligence

Common carriers’ duties towards passengers with disabilities could be addressed under any number of contract or tort theories, but it is customary to consider them to be specialized obligations on a particular class of actors to avoid behaving negligently. As Professor Milani has shown, due at least partly to the influence of access statutes, common carriers have been found to have an obligation to serve people with disabilities as well as obligations to make necessary accommodations and provide special assistance, at least when placed on notice of the need to do so.160 If the individual with a disability is injured as a result of the failure to provide an accommodation, the carrier is responsible in damages.161

Common carriers aside, negligence actions may arise when employers or others cause harm by careless treatment of individuals with disabling conditions. A Connecticut court denied the defendant’s motion for summary judgment on a negligent infliction of emotional distress claim relating to the manner in which the

---


159 See infra Part III.B.7.

160 Milani, supra note 25, at 376–81 (collecting cases). The refusal of common carriers to serve passengers with disabilities who travel without attendants was a particular concern of Professor tenBroek. See tenBroek, supra note 18, at 891–96 (collecting cases).

161 See Cunningham v. Vincent, 650 N.Y.S.2d 850, 853–54 (N.Y. App. Div. 1996) (upholding damages liability for person with paraplegia injured in fall from wheelchair while being lowered from van). Professor Ehrenreich has suggested that heightened obligations common carriers have to protect passengers from sexual harassment should be imposed on others, such as employers. Ehrenreich, supra note 31, at 47–52.
plaintiff was fired after an exacerbation of the plaintiff’s back injury from working without the accommodation he had requested.\textsuperscript{162} Evidence supported his allegations that supervisors spoke rudely to the plaintiff’s wife, delayed the plaintiff when he sought to leave the workplace to obtain medical treatment, insisted that he take a drug test, then terminated him when he responded angrily.\textsuperscript{163} The court applied state precedent establishing that an employer is liable for conduct in the termination process that involves an unreasonable risk of causing emotional distress that could result in illness or bodily harm.\textsuperscript{164} The court did not require outrageous conduct to be shown, but rather relied on the imposition of risk caused by the company’s agents.\textsuperscript{165}

2. Discharge from Employment in Violation of Public Policy

An employer commits a tort when it fires or takes other adverse actions against employees because they have engaged in activity protected by public policy.\textsuperscript{166} The new Employment Restatement places its emphasis on retaliatory conduct as the basis for the tort, although it acknowledges that the principle may cover more than retaliation for whistle-blowing, obeying professional conduct codes, discharging public duties or obligations reasonably believed to be imposed by law, and filing charges or claiming benefits under an employment statute.\textsuperscript{167}

Accommodation-denial claims may fit within the retaliation framework. In \textit{Fotheringham v. Avery Dennison Corp.},\textsuperscript{168} a California appellate court overturned a grant of summary judgment against plaintiff with respect to common law public-policy and statutory claims for discharge in retaliation for requesting an accommodation through an attorney.\textsuperscript{169} The presence of a common law remedy for

\begin{footnotesize}
\begin{enumerate}
\item[162] Tomick v. United Parcel Servs., Inc., No. CV064008944, 2010 WL 2196576, at *3 (Conn. Super. Ct. Apr. 23, 2010). Tomick requested a driver’s helper on his delivery route and was assured he would be provided one, but never was. \textit{Id.} at *1.
\item[163] \textit{Id.} at *1–2.
\item[164] \textit{Id.} at *3 (citing Perodeau v. Hartford, 792 A.2d 757 (Conn. 2002) and Parsons v. United Techs. Co., 700 A.2d 655 (Conn. 1997)).
\item[165] In fact, the court granted summary judgment against the plaintiff on the claim for intentional infliction of emotional distress. \textit{Id.} at *4. An intriguing possibility is that a state court could deem the violation of a federal statute relating to disability, which does not have its own cause of action for compensatory damages relief, to constitute negligence, on analogy to common law negligence actions based on violations of federal statutes such as the Federal Food, Drug, and Cosmetic Act, 21 U.S.C. §§ 301-399d (2006). \textit{See} Merrell-Dow Pharms., Inc. v. Thompson, 478 U.S. 804, 812 (1986) (finding no federal question jurisdiction for state-law negligence action premised on violation of Food, Drug, and Cosmetic Act). \textit{See generally} Rothstein, supra note 28, at 373–74 (discussing prospect of tort action based on violation of federal special education law).
\item[166] \textit{RESTATEMENT (THIRD) OF EMP’T § 4.01} (Tentative Draft No. 2 (revised), 2009).
\item[167] \textit{Id.} § 4.02.
\item[169] \textit{Id.} at *6–9.
\end{enumerate}
\end{footnotesize}
Retaliatory conduct would solve the remedial gap that a number of courts have created for retaliation claims under the ADA. Some courts have a more expansive idea of the public policy tort than merely a claim for retaliation, and the Restatement allows for such an interpretation. One example is the Ward case discussed above, in which the court upheld a claim under Idaho law for wrongful discharge of the worker who had just returned from a third back surgery. The court determined that Idaho law provides a tort claim when legislation protects a given class of individuals but does not provide a statutory civil remedy. In other words, whenever the state disability discrimination statute appears designed to protect the plaintiff in a case but does not give any specific relief, the public policy tort will supply one.

3. Tortious Interference with Contract

Tort liability exists when an individual intentionally, and without legal justification, interferes with a contract between two other people by inducing or otherwise causing one of them not to perform the contract, to the damage of the other. This cause of action would apply in the not-unusual situation where someone, say a human resources department manager or an immediate supervisor, promises to provide accommodations such as relaxation of rules or light duty for an employee with a disability, but someone else interferes with the arrangement.

---

170 See supra text accompanying note 87 (describing some courts’ refusal to award damages for retaliation in ADA cases).
171 Restatement (Third) of EMP’T § 4.02 (Tentative Draft No. 2 (revised), 2009) (“An employer is subject to liability in tort . . . for disciplining an employee who acting in a reasonable manner . . . (f) engages in other activity directly furthering a substantial public policy.”).
173 Id. at 1194.
174 The court left open the possibility that even when the state statute provides a civil remedy the tort may still apply. Id. In a prominent case, the California Supreme Court affirmed as modified a judgment on a jury verdict for a plaintiff who claimed discharge in violation of public policy and violation of the state fair employment statute in connection with disability harassment leading to termination. Roby v. McKesson Corp., 219 P.3d 749, 799 (Cal. 2009).
175 Restatement (Second) of TORTS § 766 (1979).
176 As a general matter, the person interfering with the agreement to accommodate would need to be acting outside the scope of his or her agency on behalf of the employer to be liable, but this would be the case if the defendant were acting contrary to the employer’s instructions or the scope of the responsibilities the employer gave the defendant. See Cedar Hill Props. Corp. v. E. Fed. Corp., 575 So. 2d 673, 676 (Fla. Dist. Ct. App. 1991) (“An agent of a corporate party to a contract, acting within his capacity and scope as an agent, cannot be considered to be a separate entity outside of the contractual relationship which can tortiously interfere with that relationship.”); cf. Restatement (Third) of AGENCY § 3.01 (2005) (“Actual authority . . . is created by a principal’s manifestation to an agent that, as reasonably understood by the agent, expresses the principal’s assent that the agent take..."
This situation was present, though this claim not raised, in what is perhaps the best known instance in which a federal court of appeals upheld a claim for disability harassment under the ADA, *Fox v. General Motors Corp.*\(^{177}\) In that case, the direct supervisor of an employee with a severe back impairment arranged for him to perform light duty work with restrictions on his lifting and carrying, but other supervisors conducted an unrelenting campaign of verbal and physical intimidation to keep him from enjoying the benefits of the accommodations.\(^{178}\) They continually gave him tasks beyond his restrictions and abused him when he complained.\(^{179}\) At one point, they assigned the plaintiff, who was six feet, seven inches tall, to a low table and chair that caused reaggravation of his back injury.\(^{180}\) They also blocked his transfer to a more suitable job in the company.\(^{181}\) The court focused on the ADA claim for hostile environment harassment in affirming judgment on the jury verdict in favor of the employee,\(^{182}\) but a claim for tortious interference would appear to be present as well.

4. Invasion of Privacy and Defamation

Torts that protect personal privacy and reputation seem particularly appropriate in some instances of disability discrimination.\(^{183}\) A court upheld a breach of privacy claim in the case mentioned above involving the lawyer with sickle-cell disease who alleged, among other things, that his partners entered his private office, removed items from the office, and opened personal mail belonging to him.\(^{184}\)

\(^{177}\) 247 F.3d 169, 179 (4th Cir. 2001).
\(^{178}\) *Id.* at 173–74.
\(^{179}\) *Id.* at 173.
\(^{180}\) *Id.* at 174.
\(^{181}\) *Id.* at 173–74.
\(^{182}\) *Id.* at 179.
\(^{183}\) See Ehrenreich, *supra* note 31, at 23 (“The intentional torts are not the only tort causes of action that protect dignitary interests. The torts of invasion of privacy and defamation do so as well.”).
\(^{184}\) Muwonge v. Eisenberg, No. 07-C-0733, 2008 WL 753898, at *10–12 (E.D. Wis. Mar. 19, 2008). The claim in Wisconsin was statutory, Wis. Stat. Ann. § 995.50 (West 2011), but the action would typically be one at common law in other places, see...
Defamation consists of communication to a third person of material exposing the plaintiff to reputational harm; in the modern era, the plaintiff must typically show that the material was false and that the defendant was at least to some degree at fault.\textsuperscript{185} Statements that a person is unfit for his or her trade or profession frequently provide the basis for defamation actions, and may support a damages award without proof of harm to reputation.\textsuperscript{186} Statements of this type include communicating a report that exaggerates the impact that a person’s disability has on his or her job performance. In one case applying California law, a federal court denied a motion to dismiss a claim for defamation when the plaintiff, who was recovering from a spinal tumor, alleged that his former supervisor had falsely and maliciously told others within the same company that the plaintiff was unable to perform the duties of a sales job for which he had applied, and that they should not even interview him.\textsuperscript{187} The valid defamation claim against the supervisor who, like the plaintiff, was a California citizen, destroyed diversity and led the federal court to remand the case to state court.\textsuperscript{188}

5. Assault and Battery

A surprisingly large number of cases involve physical abuse of people with disabilities, even in seemingly unlikely settings such as supervised workplaces and schools. One of the earliest cases in which a court upheld a claim under the ADA for harassment on the basis of disability is \textit{Haysman v. Food Lion, Inc.},\textsuperscript{189} which involved both verbal abuse and physical abuse by a supervisor who repeatedly struck a worker who had a severe back injury.\textsuperscript{190} The court ultimately found the plaintiff’s common law causes of action barred by operation of the workers’ compensation law, but it did not deny the applicability of common law relief had the statute not been interpreted to have preemptive effects.\textsuperscript{191} Cases in which courts have denied motions for summary judgment against plaintiffs include tort actions over abusive conduct by schoolteachers. For example, in a case involving a class of children with behavioral disorders, the teacher tied the students with bungee
cords and duct tape, left students restrained in chairs or on the floor for prolonged periods, hit students, and grabbed them.192

6. Intentional Infliction of Emotional Distress

The tort of intentional infliction of emotional distress is a relatively recent innovation in the law. Commentators in the first half of the twentieth century noted that courts were awarding relief in some intentional conduct cases where the harms were solely emotional and the facts did not meet all the elements of assault, defamation, or other well-recognized torts.193 In one highly influential case from the early 1950s, the California Supreme Court upheld a jury award of compensatory and punitive damages to a businessman who had been intimidated into agreeing to pay off a trade association and a competitor in exchange for being allowed to engage in the trash disposal business in Los Angeles.194 Although there were threats of violence, the association contended that the threats all concerned harm that might take place in the future, and were not threats of imminent harm that would support a claim for assault.195 The court said that did not matter: “[A] cause of action is established when it is shown that one, in the absence of any privilege, intentionally subjects another to the mental suffering incident to serious threats to his physical well-being, whether or not the threats are made under such circumstances as to constitute a technical assault.”196 The Second Restatement of Torts generalized this holding to declare that a claim exists whenever someone “by extreme and outrageous conduct intentionally or recklessly causes severe emotional distress to another” irrespective of whether threats are involved.197 The Restatement drafters famously commented that “Generally, the case is one in which the recitation of the facts to an average member of the community would arouse his resentment against the actor, and lead him to exclaim, ‘Outrageous!’”198

The intentional infliction cause of action is well suited to instances when someone inflicts severe emotional distress on a person with a disability by exposing that person to continual ridicule, cruel pranks, threats and intimidation, or

193 E.g., Calvert Magruder, Mental and Emotional Disturbance in the Law of Torts, 49 HARV. L. REV. 1033, 1058 (1936); William L. Prosser, Intentional Infliction of Mental Suffering: A New Tort, 37 MICH. L. REV. 874, 879 (1939). For a more elaborate history of the influences of scholarship, the American Law Institute, and the courts in the tort’s development, see Gergen, supra note 3, at 1705–07.
195 Id. at 284.
196 Id. at 284–85.
197 RESTATEMENT (SECOND) OF TORTS § 46 (1965).
198 Id. § 46 cmt. d.
Other abusive treatment.\textsuperscript{199} According to the Second Restatement (in language that may leave something to be desired), “The extreme and outrageous character of the conduct may arise from the actor’s knowledge that the other is peculiarly susceptible to emotional distress, by reason of some physical or mental condition or peculiarity.”\textsuperscript{200} Power disparities also support a finding that behavior is outrageous when the behavior inflicts severe emotional harm on a person with a disability. Examples of power disparity include those that exist between a student with disabilities and a teacher or nondisabled group of peers, or between a worker with a disabling condition and a supervisor or coworkers acting with the acquiescence of the supervisor.\textsuperscript{201}

There are a number of cases in which courts have allowed intentional infliction claims to proceed when the plaintiff alleged disability discrimination. Some are employment cases. In \textit{Martinez v. Monaco/Viola, Inc.},\textsuperscript{202} the plaintiff, a salesman with a strong record of success, missed six days of work.\textsuperscript{203} One of the principals of the company called the plaintiff’s physician and learned that the plaintiff had AIDS.\textsuperscript{204} After the plaintiff refused an order to resign, the principals of the company refused to speak with him.\textsuperscript{205} He learned a month and a half later that he was being moved from salary to straight commission pay, and when he demanded a meeting to discuss the compensation arrangement, he was fired.\textsuperscript{206} He refused to sign a liability release, and was told that as a result he would not receive back salary and commissions that were due him.\textsuperscript{207} The court, applying Illinois law, denied a motion to dismiss an intentional infliction claim.\textsuperscript{208}

---

\textsuperscript{199} There may be some divergence of opinion on this topic. Although noting that “the best candidate for situating a tort claim for discriminatory behavior has been the tort of intentional infliction of emotional distress,” Chamallas, \textit{supra} note 31, at 2125, Professor Chamallas believes that “conceptually, the harassment claim does not fit particularly well under either torts or civil rights,” \textit{id.} at 2139. She stresses that claims for pure emotional distress are at the periphery rather than the heart of tort law. \textit{Id.} at 2144. She further asserts that tort law does not operate in terms of social groups or, in general, enterprise liability. \textit{Id.} at 2146.

\textsuperscript{200} \textit{RESTATEMENT (SECOND) OF TORTS} § 46 cmt. f, illus. 9 (1965).

\textsuperscript{201} \textit{See} \textit{DOBBS, supra} note 34, § 304 (noting that intentional infliction liability applies when the “defendant uses the inequality [of position] to inflict emotional harm without regard for the plaintiff’s interests.”); \textit{see also RESTATEMENT (SECOND) OF TORTS} § 46 cmt. e (1965) (“The extreme and outrageous character of the conduct may arise from an abuse by the actor of a position, or a relation with the other, . . . or power to affect his interests.”). \textit{See generally RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL & EMOTIONAL HARM} §§ 45–46 (Tentative Draft No. 5, 2007) (embodies similar principles).

\textsuperscript{202} No. 96 C 4163, 1996 WL 547258 (N.D. Ill. Sep. 18, 1996).

\textsuperscript{203} \textit{Id.} at *1.

\textsuperscript{204} \textit{Id.}

\textsuperscript{205} \textit{Id.}

\textsuperscript{206} \textit{Id.}

\textsuperscript{207} \textit{Id.}

\textsuperscript{208} \textit{Id.} at *3.
Lehigh Valley Housing Authority, a worker with cerebral palsy alleged that a supervisor denied accommodation requests, called her stupid and a cripple, cursed her, mocked her condition in front of other employees, falsely told her that a male had called the office to say he wanted to visit her at home to “play games” with her, told clients she was no longer employed when she was on sick leave, demanded that she meet at the office on her day off, conducted conferences in overheated rooms, and left threatening messages on her voice mail. The court denied a motion to dismiss the intentional infliction claim.

Other cases arise in educational settings. In Baird v. Rose, the plaintiff, a high school girl, made a suicide attempt and was placed on a program of counseling and medication for depression. Her mother informed the school, and the next day the teacher in the plaintiff’s musical performance class told the class that she would be dropped from the next performance. When the mother complained to the teacher, the teacher said that she did not think someone with depression could be counted on to meet her responsibilities. The mother responded by submitting letters from a doctor and psychologist saying that the plaintiff could participate and would be harmed by being excluded. The teacher then removed her on the basis of absences, but was told by the principal of the school that she would have to exclude all children with the same number of absences. So the teacher told the class that she was being forced to drop three other students from the performance and asked the class if they understood why she had to enforce the attendance policy strictly. Other students responded by accusing the plaintiff, who left the class humiliated, and was crying uncontrollably and shaking by the time her mother arrived. Ultimately, she was kept out of all but a small part of the performance. The trial court dismissed the case, but the court of appeals applied Virginia law and upheld claims both under the ADA and under the tort of intentional infliction.

Still other cases involve public accommodations or public services. In Williams v. Tri-County Metropolitan Transportation District, a passenger with a physical disability who was using a service dog boarded a bus. The driver loudly questioned her about the dog, refused to look at papers authorizing use of the dog,

---

210 Id. at *1.
211 Id. at *5–6.
212 192 F.3d 462 (4th Cir. 1999).
213 Id. at 465.
214 Id.
215 Id.
216 Id.
217 Id. at 466.
218 Id.
219 Id.
220 Id. at 472–73.
222 Id. at 203.
wrongly demanded that the dog have an identification card, said that the passenger did not look disabled but was trying to illegally obtain a reduced fare, and finally ordered the passenger out of the bus at her destination, saying loudly that he did not have to let a dog on the bus. Although the trial court entered judgment on the pleadings on the claim for intentional infliction, the state appellate court emphasized that mistreatment directed to “historically disfavored personal characteristics” such as disability are more apt to be outrageous than other misconduct, and reversed the decision, remanding the case for trial.

One barrier to intentional infliction liability is the unwillingness of some courts to recognize conduct as outrageous. For example, in Costello v. Mitchell Public School District 79, a seventh-grader who had academic difficulties alleged that her band teacher tired of her requests for after-class help, and that every school day for a month told her in front of the class that she was “retarded, stupid, and needed to go to a school where retarded people were taught.” He said, “[i]f you’re so retarded, you don’t need to be in this classroom.” After grading her class notebook one day in front of the class, he called her stupid and threw the notebook at her, hitting her in the face. The student eventually quit the band, but had to take a music appreciation class from the same teacher; during the nine weeks she was in that class the teacher frequently called her “retarded” and “stupid” in front of her classmates, even after learning that she was undergoing mental health treatment. She developed major depression and suicidal thoughts and ultimately withdrew from school. Despite the argument that this conduct amounted to outrageous abuse of power by someone with a responsibility to nurture and protect the student, the federal district court granted summary

---

223 Id.
224 Id. at 204–05.
226 266 F.3d 916 (8th Cir. 2001).
227 Id. at 925 (Hamilton, J., concurring and dissenting in part).
228 Id.
229 Id.
230 Id. at 925–26.
231 Id. at 926.
judgment on the Nebraska common law intentional infliction claim, and the federal court of appeals affirmed.\textsuperscript{232} In another case, even though the Restatement provides as an example of intentional infliction the playing of a cruel prank on a person with a disability,\textsuperscript{233} a federal court refused to let the jury hear a claim in which an individual with multiple sclerosis alleged that on more than one occasion the heat was turned up in his work area to exacerbate his disability.\textsuperscript{234}

Juries appear to be more prone to view discriminatory conduct against people with disabilities as intolerable abuse than judges—especially federal judges—are. One federal court overturned a $150,000 jury verdict in an intentional infliction claim in which a worker with bipolar disorder put forward evidence that his employer’s supervisors knew about his disability, fired him because they believed his disability interfered with his work schedule, telephoned him at home while he was on vacation to tell him that he was being terminated, and falsely told him he was being fired for low productivity.\textsuperscript{235} The deception and abuse of power apparently registered more strongly with a cross-section of ordinary citizens than with a panel of federal appellate judges. In \textit{Sanglap v. LaSalle Bank, FSB},\textsuperscript{236} the same court of appeals affirmed a federal district judge’s decision to enter judgment against the plaintiff notwithstanding a jury verdict of $80,000 in the plaintiff’s favor, in an intentional infliction case over a bank’s closing the account of a customer after he had epileptic seizures on its premises.\textsuperscript{237} A common law action prosecuted in state court may have reached a different result in both cases.

7. \textit{A Tort of Disability Discrimination?}

The Canadian Supreme Court recently discussed but declined to rule on whether there exists a freestanding tort of disability discrimination in connection

---

\textsuperscript{232} \textit{Id.} at 924.
\textsuperscript{233} The Second Restatement of Torts provides the rather curious illustration:

\begin{quote}
A, an eccentric and mentally deficient old maid, has the delusion that a pot of gold is buried in her back yard, and is always digging for it. Knowing this, B buries a pot with other contents in her yard, and when A digs it up causes her to be escorted in triumph to the city hall, where the pot is opened under circumstances of public humiliation to A. A suffers severe emotional disturbance and resulting illness. B is subject to liability to A for both [emotional disturbance and illness].
\end{quote}


\textsuperscript{234} \textit{Shaner v. Synthes} (USA), 204 F.3d 494, 507–08 (3d Cir. 2000).
\textsuperscript{235} \textit{Van Stan v. Fancy Colours & Co.}, 125 F.3d 563, 569 (7th Cir. 1997).
\textsuperscript{236} 345 F.3d 515 (7th Cir. 2003).
\textsuperscript{237} \textit{Id.} at 517, 519.
with discharge from employment. The court expressed “concern that a tort of discrimination does not contain an effective limiting device” and that it would be inconsistent with legislative intent in establishing the Ontario Human Rights Code, which contains its own remedial scheme and restrictions on compensation for violations. The court determined that the facts involved in the employee’s discharge did not support such a claim in any instance.

Despite that court’s concerns, there may be stuff from which common law courts could fashion such a tort claim. A court, for example, found that it violates the implicit duties of a common carrier to refuse to sell a ticket to a person on the sole ground that he was blind. Several California courts have ruled that harassment in violation of the state’s discrimination laws necessarily constitutes the tort of intentional infliction of emotional distress. Just as intentional infliction liability itself grew out of traditional torts such as assault and negligent infliction of emotional distress, so too a specific tort of disability discrimination might emerge from negligence, violation of public policy, intentional infliction, and the other causes of action. If the existing avenues of relief prove adequate over time to remedy manifest unfairness and express community moral sentiment against disability discrimination, that may never happen, and never need to happen. The need for efforts to persuade courts to adopt such an approach may depend on the insufficiency of currently available common law claims in relation to progress on social attitudes about disability.

---

239 Id. at para. 65.
240 Id. at para. 63.
241 Id. at para. 67. Thirty years ago, the Canadian court ruled that alleged racial discrimination does not give rise to a common law tort. Seneca Coll. v. Bhadauria, [1981] 2 S.C.R. 181, 195 (Can.).
244 See supra text accompanying notes 42–48 (discussing relation of common law liability to moral sentiment and community sense of justice).
245 See Weber, supra note 38, at 1142–47 (discussing anticipated heightening of obligations of reasonable accommodation in light of disability rights movement and other social innovation).
IV. EFFECTIVENESS OF COMMON LAW CLAIMS

Thus, common law claims may lie in instances of disability discrimination. But can they furnish deterrence and compensation in cases where federal statutory claims prove inadequate? To answer that question, it is necessary to return to the obstacles to ADA liability. For some of these, such as statutes of limitations or size of employer, the efficacy of the common law alternative should be plain. But for others, a more nuanced analysis is required, one that considers the specific ADA theory and the common law’s ability to supplement it or substitute for it. Although some common law remedies have been blocked due to preemption by state statutes or unwillingness of courts to find employers liable for conduct of employees, those barriers are not insurmountable.

A. ADA Problems and Common Law Solutions

As noted above, there are various reasons the ADA may not be effective as an avenue of relief for disability discrimination. Three reasons to consider here are (1) the defendant’s conduct is considered insufficiently severe or pervasive under a standard borrowed from Title VII to establish a hostile environment claim under the ADA; (2) the plaintiff does not pass the test of whether he or she is a person with a disability under the ADA; and (3) the causal link between the disability and the discrimination is insufficient. Courts have found tort or contract liability even though the plaintiffs’ satisfaction of ADA standards might be viewed as deficient in these respects.

First, common law claims of intentional infliction of emotional distress, if pressed aggressively, may win in some instances when ADA claims for hostile environment disability harassment might lose for failure to show severe or pervasive discriminatory conduct. One might assume that conduct must be severe or pervasive in order to constitute extreme and outrageous behavior, and so cases that fail under the ADA standard borrowed from Title VII would also fail with regard to intentional infliction. In reality, the standards are different. An environment “permeated with discriminatory intimidation, ridicule, and insult,” might be extreme and outrageous within the meaning of the intentional infliction standard, but so might individual acts that do not constitute a pattern sufficient for the Title VII test, or so might pranks and cruelty that the court does not want to deem “discrimination.” Cases upholding intentional infliction claims based on single instances of conduct, such as discharge, are key examples of where the common law tort action might remedy something conventional interpretations of

\[246\] See supra text accompanying notes 11, 131 (discussing failure of many ADA harassment claims).

\[247\] Harris v. Forklift Sys., Inc., 510 U.S. 17, 21 (1993) (Title VII case). Although the context makes clear that this was supposed to be an example of how a workplace might be a hostile environment in violation of the law, many courts have treated the description as a minimum requirement for the claim. See Weber, supra note 87, at 249.
the ADA might not. For example, in *Soodman v. Wildman, Harrold, Allen & Dixon*, the court denied summary judgment for the defendant in a case involving a secretary who was discharged while she was on medical leave for a pregnancy that was unusually high risk due to her physical impairment, an incompetent cervix. The plaintiff suffered extreme distress, and gave birth prematurely. The court noted that although the defendant had always treated the plaintiff “in a cordial and professional manner,” rather than insulting or demeaning her, the evidence that she was selected for layoff in part on the basis of her medical condition was sufficient to support a claim for intentional infliction of emotional distress under Illinois law. In another case, a court affirmed a jury verdict for compensatory and punitive damages when an employer terminated an employee who had worked there for twenty-two years, after he had an apparent heart attack. Two representatives of the employer went to the plaintiff when he was lying in bed, partly undressed, and told him abruptly that he was fired. The trial court had previously dismissed the ADA claim and it was not on appeal, but the judgment on the claim for intentional infliction of emotional distress was affirmed.

Second, common law claims might win when the plaintiff does not meet the test for a person with a disability covered by the ADA. Many commentators have noted that the Supreme Court’s interpretation of the definition of a person with disability leaves many potential plaintiffs with seriously disabling conditions outside the circle of protection of the statute. One benefit of pursuing many claims under common law is the lack of any requirement, technical or otherwise, that the plaintiff shows he or she is a person with a disability. Once the plaintiff satisfies the elements of the claim, the focus is on the actions themselves, not whether the plaintiff falls into a given statutory definition. In cases involving persons with diabetes, heart trouble, and correctible visual impairments, all conditions courts have frequently found not sufficient to support the conclusion

---

249 Id. at *2.
250 Id. at *4.
251 Id. at *10.
253 Id. at 500.
254 Id. at 500–01.
257 Archer, 70 P.3d at 495.
that a person has a disability under the ADA, plaintiffs have prevailed on the tort claims they asserted. As noted above, however, the recent amendments to the Americans with Disabilities Act have the potential to reduce the need for common law remedies created by the Supreme Court’s previous interpretations of the statutory definition.

Third, some common law claims might succeed when an ADA action fails because the offending behavior, though severe, might have an insufficient causal connection to the disability. The point might seem superfluous. If the disability is not the cause of the abuse, why would the case be one of disability discrimination? Reality, however, often defies clear classifications of events and conditions. In Robel v. Roundup Corp., the Washington State Supreme Court reinstated judgment for plaintiff on a jury verdict on a state statutory claim and an intentional infliction claim when a delicatessen worker who had been injured on the job was subjected to a campaign of humiliation and profane abuse by coworkers after being moved to a light-duty assignment. The dissent, however, noted that hostility between the plaintiff and at least one of the coworkers preexisted the injury and so concluded that the statutory claim had to fail for the lack of a causal connection between disability and the mistreatment. What the dissent’s analysis ignored was that discriminatory conduct of this type is likely to focus on someone’s disability—or whatever makes the person different from and subordinate to others—even if the underlying motivation for the conduct is preexisting conflict, an evil disposition, or anything else. Thus the causation problem should not prevent statutory liability. Nevertheless, for the dissent it did. In any instance, with regard to the intentional infliction cause of action, facts going to the underlying motivation do not necessarily matter. The attention is on the actual conduct at issue and its real effects.

B. Potential Barriers to Common Law Claims

A potential problem with common law claims is preemption by state disability discrimination statutes and workers’ compensation laws. In the Sanglap case

260 59 P.3d 611 (Wash. 2002).
261 Id. at 613–14.
262 Id. at 623 (Bridge, J., dissenting).
263 See, e.g., Nagel v. Morgan Stanley DW, Inc., No. 07CV 3439, 2007 WL 3252915, at *3 (N.D. Ill. Oct. 31, 2007) (dismissing emotional distress action related to disability discrimination on account of preemption by state human rights law); Powell v. Greenwald Indus., Inc., No. CV095013578, 2010 WL 2383784, at *7 (Conn. Super. Ct. Apr. 29, 2010) (dismissing action for breach of covenant of good faith and fair dealing related to disability discrimination on account of preemption by state human rights law). Various courts have found intentional infliction claims based on sex or race harassment preempted by state discrimination statutes, though a larger number have found the claims not to be preempted. See Chamallas, supra note 31, at 2136–37 and accompanying notes (collecting cases). A prominent Iowa sex harassment case holding that the state civil rights statute preempted an
concerning the closing of the account of a bank customer after he had an epileptic seizure on the premises, an additional basis given by the Seventh Circuit for overturning the jury verdict was that the Illinois Human Rights Act was the sole avenue of state law relief. The court contended that the relevant facts regarding intentional infliction were inextricably linked to those that constituted the violation of the state statute.

The state case using the “inextricably linked” language in determining the scope of statutory preemption in fact rejected an argument that in a sexual harassment case the Act preempted common law claims against the coworker of a plaintiff for assault, battery, and false imprisonment when the coworker was the one accused of having committed the sexual harassment. The decision distinguished the common law intentional tort claims arising from the coworker’s conduct from a claim such as one of negligent hiring or retention against the employer, which would properly be preempted. If the reason that retention of the coworker is negligent is inextricably linked to the harassing behavior, then the negligent retention action cannot stand. But when common law claims “exist wholly separate and apart from a cause of action for sexual harassment under the Act,” and the elements of the tort are alleged “without reference to legal duties created by the Act,” there is no preemption. Other courts have interpreted the Illinois preemption standard as whether the plaintiff can prove the elements of the common law claim independent of the legal duties established by the state statute.

Applying this approach to, for example, an intentional infliction cause of action, the claim escapes preemption so long as the reason the underlying conduct was outrageous is not simply the fact that it is illegal under the state statute. Accordingly, a court found no preemption of an intentional infliction claim when the plaintiff was forced to work in an office building site suffused with harmful chemicals, dust, and debris, and then, after becoming sick, was fired on the basis of

intentional infliction claim, though not one for assault and battery, is Greenland v. Fairtron Corp., 500 N.W.2d 36, 38–39 (Iowa 1993).


265 Sanglap v. LaSalle Bank, FSB, 345 F.3d 515, 519–20 (7th Cir. 2003).

266 Id. at 519.


268 Id. at 23–24.

269 Creighton v. Pollman N. Am., Inc., No. 08 C 3241, 2008 WL 5377816, at *3 (N.D. Ill. Dec. 18, 2008) (“Therefore, the [state statute] does not preclude courts from exercising jurisdiction over IIED claims factually related to incidents of unlawful discrimination if the plaintiff can allege facts sufficient to establish the elements of IIED.”); see also Naeem v. McKesson Drug Co., 444 F.3d 593, 604 (7th Cir. 2006) (Title VII case) (“[T]he proper inquiry was not whether the facts that support Ms. Naeem’s intentional infliction of emotional distress claim could also have supported a discrimination claim, but instead whether Ms. Naeem can prove the elements of intentional infliction of emotional distress independent of legal duties furnished by the [state statute].”).
her perceived disability. The intentional infliction claim stood independent of the duties imposed by the state human rights law. This decision’s reasoning is more consistent with that taken in the controlling state supreme court case and the underlying goal of allowing tort relief unless the legislature actually directs a given category of claims elsewhere.

A few courts have found preemption of common law claims under workers’ compensation laws, but those holdings are, for good reason, rare. Disability discrimination is hardly one of the occupational injuries for which the workers’ compensation system is designed to provide a speedy, non-fault-based remedy. One case that recognized this fact is Bourbeau v. City of Chicopee. There a Massachusetts court considered whether the workers’ compensation law preempted intentional infliction claims based on a supervisor’s conduct that included failure to curb smoking by coworkers that exacerbated plaintiff’s seizure disorder, exclusion of the plaintiff from meetings after he complained about the smoking,

---

270 Creighton, 2008 WL 5377816, at *3.
271 Id.
272 In general, it would appear more plausible that a legislature would intend to preempt actions such as wrongful discharge in violation of public policy, particularly when the public policy is embodied in the statute that is doing the preempting, than that it would intend to preempt a freestanding action such as intentional infliction of emotional distress. Some courts have drawn such a distinction. See, e.g., Wright v. Cor-Rite, Inc., No. 3:05cv2431, 2007 WL 2907947, at *5 (M.D. Pa. Oct. 2, 2007) (finding claim for intentional infliction against employer by individual with mobility impairment not subsumed by state human rights law, but finding wrongful discharge claim barred); see also Jarod S. Gonzalez, State Antidiscrimination Statutes and Implied Preemption of Common Law Torts: Valuing the Common Law, 59 S.C. L. REV. 115, 122 (2007) (proposing approach whereby “[t]orts that do not have an existence separate and apart from employment discrimination are impliedly preempted if they were not recognized under a state’s common law prior to the enactment of the state antidiscrimination statute. Common law wrongful discharge torts that existed prior to the enactment of a state antidiscrimination statute are not impliedly preempted”). One court, construing a state statute that had previously been found to have broad preemptive effects, ruled that a tort claim could be brought against an employer and supervisor of a woman with an intellectual disability with whom the supervisor engaged in sexual conduct. Saucier ex rel. Mallory v. McDonald’s Rest. of Mont., Inc., 179 P.3d 481, 497 (Mont. 2008) (considering tort claims of battery, negligent and intentional infliction of emotional distress, negligent supervision, failure to provide a safe workplace, and breach of fiduciary duty).
273 One such case is Haysman v. Food Lion, Inc., 893 F. Supp. 1092, 1111–12 (S.D. Ga. 1995), which nevertheless upheld a damages claim under the ADA for disability harassment. Id. at 1108–11. Courts have gone in both directions on the issue whether workers’ compensation statutes preempt employment-related sex and race harassment claims for intentional infliction of emotional distress. See Chamallas, supra note 31, at 2137–39 and accompanying notes (collecting cases).
shouting at the plaintiff after the complaint, physically advancing on him, and ultimately locking him out of the office. The court rejected the preemption argument, concluding that intentional infliction of emotional distress is not one of the accepted risks of employment, unlike ordinary occupational injuries.

In general, preemption arguments depend upon the contention that a legislature intended enactments such as state antidiscrimination or workers’ compensation laws to eliminate previously existing rights to relief for wrongful conduct. Though in some instances, the legislature may wish to replace one remedy with another, the argument that a legislature intended to decrease the tort or contractual remedies available to a plaintiff by passing a broad human rights law or a law addressed to occupational injuries will often have difficulty passing a straight-face test. Claims that legislatures meant to weaken existing rights when they thought they were expanding protections should be viewed with skepticism.

A final issue that will arise in some cases and that may undermine effective relief is lack of respondeat superior liability for employers of individuals who discriminate, particularly when the basis for common law liability is an intentional tort. Respondeat superior is the familiar principle by which employers are liable for acts of employees committed within the scope of employment. Many courts take a broad view of the respondeat superior principle and accept as a matter of course the proposition that employers should be liable when employees intentionally inflict emotional distress on other employees or anyone else if the conduct occurs on the job. In Robel v. Roundup Corp., the court applied

---

276 Id. at 109.
277 Id. at 117. The court also rejected an argument that an intentional infliction claim was preempted by the state antidiscrimination law, noting that the Massachusetts Supreme Court had consistently found no preemption of intentional infliction by the antidiscrimination statute. Id. at 116–18; see also Creighton v. Pollman N. Am., Inc., No. 08 C 3241, 2008 WL 5377816, at *5 (N.D. Ill. Dec. 18, 2008) (finding no preemption of intentional infliction claim by Illinois workers’ compensation law in sick building case).
278 See Brooke v. Rest. Servs., Inc., 906 P.2d 66, 70 (Colo. 1995) (stating with regard to claim in connection with sex discrimination, “[W]e cannot conclude that the legislature intended the [state Anti-Discrimination] Act to preempt the remedies that are otherwise available to victims of sexually discriminatory conduct in the work place.”).
279 FRANKLIN ET AL., supra note 34, at 22.
281 59 P.3d 611 (Wash. 2002).
respondeat superior liability to the intentional infliction claim, pointing out that although the conduct was not authorized by company policy, it was nevertheless within the scope of employment. The employees “tormented Robel on company property during working hours, as they interacted with coworkers and customers and performed the duties they were hired to perform.” Not all cases have fallen in line with this view, however.

The underlying purposes of respondeat superior are to give incentives to employers to hire more carefully and use training and discipline to make sure their employees conduct themselves properly, to spread the costs of harm caused by employees when they fail to conform to standards of behavior, and to force enterprises to internalize the social costs of their activities. Employers need incentives to screen out and to not retain workers who engage in discriminatory conduct; moreover, disability discrimination represents real harm that should be compensated but may well not be if the employee-perpetrator is judgment proof, unable to be found, or otherwise unable to provide compensation. Accordingly, the courts that find respondeat superior liability have the better of the controversy.

With at least some forms of tort liability, respondeat superior will not necessarily be required to reach the employer. A cause of action for intentional infliction of emotional distress, for example, will apply if the defendant is reckless. Recklessness consists of proceeding without concern for the safety of others in light of known, serious risks. Applying these ideas, a court applying Illinois law refused to dismiss an intentional infliction claim against the employer in the case involving the employee forced to work in an unhealthy environment who was terminated after she became ill from the conditions, when she alleged that the actions were undertaken with recklessness.

\[282\] \textit{Id.} at 621.


\[284\] See DOBBS, \textit{supra} note 34, § 334.

\[285\] Both the older and the new Restatement include recklessness in their formulation of the tort. \textit{Restatement (Third) of Torts: Liability for Physical & Emotional Harm} § 45 (Tentative Draft No. 5, 2007) (“An actor who by extreme and outrageous conduct intentionally or recklessly causes severe emotional harm to another is subject to liability . . . .”); \textit{Restatement (Second) of Torts} § 46(1) (1965) (“One who by extreme and outrageous conduct intentionally or recklessly causes severe emotional distress to another is subject to liability . . . .”). Thanks to an unnamed participant in a workshop at Marquette University Law School for this point.

\[286\] DOBBS, \textit{supra} note 34, § 27.

CONCLUSION

If this Article demonstrates nothing else, it shows that the common law is both a fruitful and too often ignored means of remedying disability discrimination. The present era may be the “Age of Statutes,” but common law has a potential for vitality in the effort against disability discrimination and merits further development.

---

288 See generally Guido Calabresi, A Common Law for the Age of Statutes 1 (1982) (“The last fifty to eighty years have seen a fundamental change in American law. In this time we have gone from a legal system dominated by the common law, divined by courts, to one in which statutes, enacted by legislatures, have become the primary source of law.”) (footnote omitted).
E-VERIFY DURING A PERIOD OF ECONOMIC RECOVERY AND HIGH UNEMPLOYMENT

Emily Patten*

I. INTRODUCTION

In the last twenty years, there has been quantifiable growth in the number of unauthorized aliens1 in the United States. In March 2010, there were “11.2 million unauthorized immigrants living in the United States,” with “8 million unauthorized immigrants in the workforce”; these numbers are substantially greater than the “3.5 million unauthorized immigrants [who were] living in the United States in 1990.”2 Unauthorized aliens only accounted for 5.2 percent of the national workforce.3 However, this percentage varies from state to state, and states with large shares of unauthorized aliens typically have a proportionately larger share in their workforce.4 Because of this, national and state governments have addressed immigration reform, with states that have a higher proportion of unauthorized aliens in their workforce being most aggressive in reform. Unfortunately, under the United States’ current economic conditions, some states have taken reform measures that could actually further cripple the economy. One such measure is the use of an online status verification system called E-Verify.

Illegal immigration is fueled in part by labor pressures and “the need of U.S. employers for low-skill, low-wage labor, a need that is compounded by the shortage of legal workers as the ‘baby bust’ generation enters the labor market.”5

* © 2012 Emily Patten. J.D. Candidate 2013, S.J. Quinney College of Law, University of Utah.

1 An unauthorized alien refers to a non-citizen immigrant who “with respect to the employment of an alien at a particular time, that the alien is not at that time either (A) an alien lawfully admitted for permanent residence, or (B) authorized to be so employed by this chapter or by the Attorney General,” 8 U.S.C. § 1324a(h)(3) (2006). Throughout this Note the author uses the word “alien” and “immigrant” interchangeably to refer to individuals who reside within the borders of the United States but are not United States citizens. The author recognizes the sensitivity associated with these terms, and has chosen to use these terms because they correspond and appear in current statutes, legislation, and regulations. The terms are not meant to communicate any negative or derogatory connotation.


3 Id. at 17.

4 Id.

For this reason, it is not surprising that the United States experienced an immigration surge in the late 1990s because of the booming U.S. economy. Under this thriving economy, the growth in unauthorized aliens greatly increased along with correspondingly positive employment factors—including tumbling unemployment rates and notable wage gains. Apparently, the association between supply and demand in the United States labor market works exceedingly well. “In fact, trends in the flow of incoming Mexican migrants correlate very well with the growth in the U.S. economy and the U.S. national rate of employment.”

Thus, it is also not surprising that in 2007 the growth rate of unauthorized aliens began to slow, just as the United States recession began. In this era of globalization, the recession is not limited to the United States but encompasses many economies, including Mexico. Because of the imploding construction sector, both authorized and unauthorized workers in the United States were being laid off in high numbers even before the downturn was officially (and appropriately) labeled a “recession.”

Attitudes and strong opinions about immigration often have economic roots. Thus, the influx of unauthorized aliens in the years preceding the recession, coupled with the economic crisis and increasing unemployment rates, have caused the United States to re-evaluate existing immigration and citizenship policies. Moreover, some states, especially those with large representations of unauthorized aliens in their workforce, have taken immigration reform into their own hands. Arizona is one such state, with approximately 10 percent of its workforce made up of unauthorized aliens. In 2007, Arizona became frustrated with “the United States Congress’s failure to enact comprehensive immigration reform” and enacted the Legal Arizona Workers Act (“LAWA”). “Arizona [was] the first state in the nation to enact a law that penalizes businesses for knowingly hiring from the previous generation. See 1 ENCYCLOPEDIA OF IDENTITY 307 (Ronald L. Jackson II ed., 2010).
unauthorized immigrants.”\textsuperscript{16} LAW A requires employers to “verify the employment eligibility of the employee through [an online status verification system] . . . .”\textsuperscript{17} If the employer hires or contracts with any unauthorized aliens, and thus violates LAW A, severe sanctions ensue, including the potential loss of any licenses.\textsuperscript{18}

Utah followed Arizona’s lead and, as of July 1, 2009, public employers and independent contractors are required to register and participate in an online status verification system.\textsuperscript{19} Effective July 1, 2010, Utah expanded this legislation to require private employers with fifteen or more employees to be “registered with a status verification system to verify the federal legal working status of any new employee . . . .”\textsuperscript{20}

There are many issues associated with the correct implementation of immigration reform, and these issues become more convoluted during times of economic hardship and with advances in technology that offer the use of an online status verification system. This Note explores a few of the problems associated with mandating the use of an online status verification system. Specifically, it addresses how this mandate affects Utah businesses. Section II focuses on the legislation that implemented Utah’s current policy of requiring the use of online status verification and the proposed legislation calling for a national mandate. Section III examines some potentially harmful effects that enforcing an online status verification system could have on local businesses. Finally, Section IV proposes modifications to the current system and current legislation, and concludes that until proper modifications are made, we should not mandate the use of an online status verification system.

II. BACKGROUND

The United States is among the more generous nations in immigration policy, but immigration is considered a privilege and not an entitlement.\textsuperscript{21} “It is a fundamental right of a sovereign state to control its borders, define who can legally reside and seek employment within its borders, and define those who are citizens and those who are not.”\textsuperscript{22} As such, immigration law is a “nation’s set of rules designed to govern this important aspect of national sovereignty.”\textsuperscript{23} On one extreme, if the United States had entirely open borders there would be no illegal immigration, but this solution is unlikely as “movement across the borders of the United States has never been totally unrestricted.”\textsuperscript{24}

\textsuperscript{16} Id.
\textsuperscript{17} ARIZ. REV. STAT. ANN. § 23-214(A) (2005).
\textsuperscript{18} ARIZ. REV. STAT. ANN. § 23-212.01(F) (2005).
\textsuperscript{19} See Utah Code Ann. § 63G-11-103(3)(a) (West 2009).
\textsuperscript{20} Utah Code Ann. § 13-47-201(1)(a) (West 2010).
\textsuperscript{21} See GimpeL & EdwardS, supra note 12, at 1–5.
\textsuperscript{22} Id. at 5.
\textsuperscript{23} Id.
\textsuperscript{24} Id. at 11.
Immigration law in the United States has developed substantially throughout the years. Prior to 1952, a variety of statutes governed immigration law, but were not comprehensively codified in one text.\(^{25}\) In 1952, amidst some controversy, Congress passed the Immigration and Nationality Act of 1952 (“INA”),\(^{26}\) which remains the core statute governing immigration in the United States.\(^{27}\) The INA “revised, codified, and repealed nearly all existing immigration law.”\(^{28}\) Although “America is indeed a joining together of many streams which go to form a mighty river which we call the American way,” in 1952 it became clear that a change was needed as “millions [were] storming our gates for admission and those gates [were] cracking under the strain.”\(^{29}\)

INA has undergone several amendments since its 1952 enactment. In the early 1970s, illegal immigration became a forceful issue in the U.S. government.\(^{30}\) “The actual number of illegal aliens was a matter of speculation at this time but public attitudes to the phenomenon were clearly hostile.”\(^{31}\) With the surge of incoming Vietnamese, Cuban, and Central American refugees in the late 1970s and early 1980s, Congress actively pursued immigration reform.\(^{32}\) Thus, one important amendment to the INA was the Immigration Reform and Control Act (“IRCA”) of 1986, which makes it unlawful “to hire, or to recruit or refer for a fee . . . an alien knowing the alien is an unauthorized alien[,]”\(^{33}\) and it imposes “civil money penal[ties] for hiring, recruiting, and referral violations[.]”\(^{34}\) and criminal penalties for repeat or multiple violations.\(^{35}\)

IRCA “constituted a major statutory response to the vast tide of illegal immigration that had produced a ‘shadow population’ of literally millions of

\(^{25}\) See JACOBSON, supra note 5, at 44–48.


\(^{27}\) JACOBSON, supra note 5, at 49. Amidst conflict in Europe, President Truman initially vetoed the INA, stating:

Today we are ‘protecting’ ourselves . . . against being flooded by immigrants from Eastern Europe. . . . We do not need to be protected against immigrants from these countries—on the contrary we want to stretch out a helping hand, to save those who have managed to flee into Western Europe, to succor those who are brave enough to escape from barbarism . . . . In no other realm of our national life are we so hampered and stultified by the dead hand of the past, as we are in this field of immigration.

\(^{28}\) Id. at 49–50.

\(^{29}\) Id. at 48.


\(^{31}\) JACOBSON, supra note 5, at 55.

\(^{32}\) Id.

\(^{33}\) Id.


\(^{35}\) Id. § 1324a(e)(4).

\(^{36}\) Id. § 1324a(f)(1).
undocumented aliens in the United States.” The fact that Congress passed a bill restricting the entry of unauthorized aliens was remarkable because “[i]mmigration is believed to be a ‘no-win’ issue for most congressmen.” This is because most interest groups are opposed to immigration restriction, yet most of the American public is hostile toward liberal immigration legislation.

Under IRCA, employers are required to verify that their workers are not unauthorized aliens by examining “a document or combination of documents that reasonably appears on its face to be genuine . . . .” The IRCA guidelines established a paper-based compliance method for employers, commonly known as the I-9 system, which requires employers to attest to examining documentation and verifying that the employee is not an unauthorized alien. The passage of IRCA appeared to attack the demand for unauthorized aliens by “ending the magnet that lures them to this country[,]” and made “combating the employment of illegal aliens central to the policy of immigration law.”

Drafters of IRCA paid special attention to recognizing that discrimination might arise due to the legislation; therefore, provisions were enacted relating to “unfair immigration-related employment practice” and “discriminat[ion] against any individual . . . with respect to . . . hiring . . . .” If an employer is found to have “engaged in or is engaging in any such unfair immigration-related employment practice,” the employer will be required to cease and desist from the practice, and could potentially face additional civil penalties.

Additionally, to “improve deterrence of illegal immigration to the United States,” Congress enacted the Illegal Immigration Reform and Immigrant Responsibility Act of 1996 (“IIRIRA”), which increased penalties for documentation fraud and “improv[ed] the verification system for eligibility for employment . . . .” To improve the verification process under IIRIRA, the federal government established three pilot programs, with the Employment Eligibility

---

37 JACOBSON, supra note 5, at 55.
38 Id. at 55–56.
39 § 1324a(b)(1)(A)(ii).
40 § 1324a(b)(1)–(2).
43 See § 1324b(a)(1).
44 See § 1324b(g).
45 See 142 CONG. REC. 24,389 (1996).
Verification Program, more commonly known as “E-Verify,” being the only one still in existence.\textsuperscript{47}

E-Verify is a free online alternative to the I-9 form system.\textsuperscript{48} “E-Verify allows an employer to actually authenticate applicable documents rather than merely visually scan them for genuineness.”\textsuperscript{49} With the E-Verify system, the information usually entered on an employee’s I-9 form is compared with the government’s records, and within seconds the potential employee has a tentative confirmation, or tentative denial of eligibility to work in the United States. If a tentative denial is issued, “the employer must notify the employee, who has eight days to challenge the finding.”\textsuperscript{50} If an employee is unsuccessful in challenging the tentative denial of eligibility to work, or does not challenge it, termination is necessary; if the employer continues to employ the employee after receiving a final denial, the employer “is subject to a rebuttable presumption that it knowingly employed an unauthorized alien.”\textsuperscript{51}

Even though problems have become apparent in many situations, the use of E-Verify has increased. In 2003, the E-Verify program was expanded for use in all fifty states.\textsuperscript{52} As of December 10, 2011, “more than 307,000 employers [were] enrolled in the program, with over 17 million [queries being run through] the system in [the] fiscal year 2011.”\textsuperscript{53} Since 2005, every immigration reform bill has included a proposition to nationally mandate E-Verify.\textsuperscript{54} Congress currently makes

\textsuperscript{48} See Chicanos Por La Causa, Inc. v. Napolitano, 558 F.3d 856, 862 (9th Cir. 2009), cert. granted sub nom. Chamber of Commerce of U.S. v. Candelaria, 130 S. Ct. 3498 (2010). However, E-Verify does not replace the legal requirement to complete and retain I-9 form; the I-9 must still be retained and stored by the employer. Questions and Answers, U.S. Citizenship & Immigration Services, http://www.uscis.gov/portal/site/uscis/menuitem.eb1d4c2a3e5b9ac89243c6a7543f6d1a/?vgnextoid=51e6fb41c8596210VgnVCM100000b92ca60aRCRD&vgnextchannel=51e6fb41c8596210VgnVCM100000b92ca60aRCRD (last visited Feb. 28, 2012).
\textsuperscript{49} Lozano v. City of Hazleton, 620 F.3d 170, 200 (3d Cir. 2010), vacated, City of Hazleton v. Lozano, 131 S. Ct. 2958 (2011).
\textsuperscript{50} Chicanos, 558 F.3d at 862.
\textsuperscript{51} Id.
\textsuperscript{54} The Social Security Administration’s Role in Verifying Employment Eligibility: Hearing Before the H. Comm. on Ways and Means, Subcomm. on Social Sec., 112th Cong. 1 (2011) [hereinafter Hearings] (statement of Tyler Moran, Policy Director, National Immigration Law Center).
the use of E-Verify optional, but, seventeen of the fifty states, including Utah, have mandated its use by some or all employers.\footnote{See E-Verify FAQ, NAT’L CONFERENCE OF STATE LEGISLATURES, http://www.ncsl.org/?tabid=13127#table (last updated Nov. 4, 2011) (indicating that Alabama, Arizona, Georgia, Mississippi, North Carolina, South Carolina, Tennessee, and Utah have mandated use by all employers, while Colorado, Florida, Idaho, Indiana, Louisiana, Missouri, Nebraska, Oklahoma, and Virginia do so for only state agencies and/or contractors, and that Minnesota and Rhode Island rescinded their requirement for state agencies and contractors in 2011).}

Under LAWA, for example, Arizona requires employers to use E-Verify.\footnote{Chamber of Commerce v. Whiting, 131 S. Ct. 1968, 1970 (2011).} Shortly after LAWA was enacted, the United States Chamber of Commerce and various other business and civil rights organizations filed suit against those charged with administering the Arizona law, “argu[ing] that the Arizona law’s provisions allowing the suspension and revocation of business licenses were both expressly and impliedly preempted by federal immigration law, and that the mandatory use of E-Verify was impliedly preempted.”\footnote{Id. at 1977.} The District Court concluded that the state law was not preempted with respect to E-Verify, because “although Congress had made the program voluntary at the national level, it had expressed no intent to prevent States from mandating participation.”\footnote{Id.} Both the Ninth Circuit and the Supreme Court in \textit{Chamber of Commerce v. Whiting} concurred.\footnote{Id.} This controversial five-three decision\footnote{Justice Kagan took no part in the consideration or decision of the case. \textit{Id.} at 1968.} essentially stated that mandating E-Verify was not preempted. However, this does not mean that mandating E-Verify is the correct course of action to take to resolve America’s immigration problems.

Since the Supreme Court settled the constitutionality issue, a new Utah version of the Arizona law has been proposed.\footnote{David Montero, \textit{Lawmaker to Propose Bill that Mirrors Arizona’s E-Verify Law}, SALT LAKE TRIB. (May 31, 2011), http://www.sltrib.com/sltrib/politics/51900396-90/law-sandstrom-bill-arizona.html.csp?page=1.} Although E-Verify is already mandatory in Utah, the proposed bill will incorporate essentially the exact provisions as the Arizona bill.\footnote{Id.} The bill even purports to go a step further by establishing a “state-run verification program called U-Verify [that] would require businesses to register with it.”\footnote{Id.} Utah State Representative Stephen E. Sandstrom, the bill’s champion, wants “to act fast” in passing this legislation because “he worries undocumented immigrants in Arizona now will head to Utah.”\footnote{See Montero, supra note 61.} However, these stricter requirements do not come without controversy. Marty Carpenter, the
spokesperson for the Salt Lake Chamber, advocates a solution from the federal government, stating that “[j]ust because it’s legal doesn’t mean it’s good policy . . . . Businesses are in the business to create markets, transport and sell goods . . . [and] are not in the business to police or enforce immigration policy.”65

Texas Representative Lamar Smith and co-sponsor Utah Representative Jason Chaffetz66 have sponsored a bill in the United States Congress, known as the Legal Workforce Act, which would amend INA to nationally “make mandatory and permanent requirements” relating to the use of E-Verify.67 If enacted into law, the Legal Workforce Act will apply to employers on a rolling basis. Employers with over 10,000 employees will be affected first.68 Then, within twenty-four months of the date of enactment of the Legal Workforce Act, all employers—including those with only one employee—will be required to abide by its policies.69 Finally, under the Legal Workforce Act, employers’ fines will be significantly increased for violations relating to employment verification.70

By validating LAWA in Chamber of Commerce v. Whiting, the Supreme Court ruling seems to have prompted other bills and legislation relating to immigration reform, as many congressional representatives, including Utah congressional representatives, apparently waited until the outcome of this landmark case before sponsoring legislation. Although immigration reform may be a solution to the current immigration problems, the reform as currently presented is unlikely the best course of action.

III. DISCUSSION AND ANALYSIS

The use of E-Verify might appear to provide employers with a timely and efficient addition to the already required I-9 form. However, even though E-Verify is purportedly free-of-charge, it is far from free-of-censure. Instead of creating job growth, mandating the use of E-Verify has spurred criticisms of the system, which include: (1) the accuracy of the system, (2) the burdensome costs to employers, (3) the potential increase in employer discrimination due to these burdensome costs, and (4) the lack of employer participation. Each of these is discussed below.

65 Id.
66 Utah Representative Jason Chaffetz is one of approximately sixty-four co-sponsors of the Legal Workforce Act. See generally H.R. 2164, 112th Cong. (2011) (listing the sponsor and co-sponsors of the Legal Workforce Act, of which Jason Chaffetz is a part).
67 Id. §§ 1–2.
68 Id. § 2(b)(1)(E).
69 Id. § 2(b)(1)(E)(IV).
70 Id. § 8.
A. The Accuracy of E-Verify

“The overall success of E-Verify is closely tied to the accuracy of its findings.”71 Groups typically oppose the mandatory use of E-Verify because the program is not accurate72; this includes Utahns opposed to the federal bill sponsored by Utah Representative Jason Chaffetz that would mandate use of the E-Verify system nationally.73 Accurate data is a measure of successful implementation of E-Verify, whereas inaccurate data can add to employer burdens.74 Thus, “accuracy is . . . of paramount importance in gaining widespread acceptance of E-Verify.”75

According to an independent study, E-Verify is purported to be 99.2 percent accurate.76 Thus, 0.8 percent of individuals authorized to work in the United States will receive a tentative nonconfirmation.77 From this, one would assume that E-Verify is highly accurate. However, “U.S. employers hire approximately 60 million workers each year . . . [and] an 0.8 percent error rate means that 480,000 legal workers will be mistakenly classified as not qualified to work.”78

To be sure, an employee may challenge an inaccurate tentative nonconfirmation result from E-Verify.79 Yet this can be difficult for the approximately twenty-one million United States citizens who do not have government-issued photo identification and the thirteen million United States citizens who do not have access to passports, birth certificates, or naturalization papers.80 Further, those who do not have access to current identification documents are likely to be elderly, low-income, minorities, or individuals with recent name changes.81 Attempting to contest the nonconfirmation can be costly and time-consuming for these individuals. Of workers who were successfully able to correct

72 Id.
74 WESTAT, E-VERIFY PROGRAM EVALUATION, supra note 71, at 114.
75 Id.
76 See id. at 117–18.
77 Id.
79 Chicanos Por La Causa, Inc. v. Napolitano, 558 F.3d 856, 862 (9th Cir. 2009), cert. granted sub nom. Chamber of Commerce of U.S. v. Candelaria, 130 S. Ct. 3498 (2010).
81 Id. at 3.
the tentative nonconfirmation, 22 percent spent more than $50 and 13 percent spent more than $100.\textsuperscript{82} This does not include the lost wages of the 14 percent who had to take off two or more days to correct the mistake.\textsuperscript{83} Additionally, these numbers likely understate the actual cost to workers to correct E-Verify mistakes, because many individuals likely become discouraged and do not attempt to correct the problem.

Further, the inaccuracy rate for unauthorized aliens is roughly 54 percent,\textsuperscript{84} meaning that 54 percent of unauthorized aliens receive a confirmation of employment eligibility under E-Verify. This is likely attributable to the fact that many “unauthorized [aliens] are highly motivated to avoid having their work-authorization status detected and the fact that . . . E-Verify cannot identify most types of identity fraud.”\textsuperscript{85} Thus, “[the] difference [between the two error rates] is not surprising.”\textsuperscript{86} Like the Form I-9, E-Verify is highly susceptible to the challenges associated with identity fraud. In a 2006 U.S. Immigration and Customs Enforcement (“ICE”) raid, ICE found that “approximately 1,340 employees—all of whom ICE believes were processed through E-Verify—were not authorized to work in the United States.”\textsuperscript{87} Of the 1,340 unauthorized aliens, 274 were charged with identity theft, which included charges of using invalid Social Security numbers and document fraud.\textsuperscript{88}

\textbf{B. E-Verify’s Burdensome Costs to Employers}

Mandating E-Verify likely generates burdensome costs for employers. With the passage of IRCA, Congress attempted to enact policies that were the “least disruptive to the American businessman.”\textsuperscript{89} Although IRCA forbids the hiring of unauthorized aliens, the law mitigates the burden to employers by requiring only good faith compliance with the I-9 requirements.\textsuperscript{90} Congress further reduced the burden by not requiring employers to verify the eligibility of independent contractors and by making the I-9 system the national requirement.\textsuperscript{91} Finally, “Congress withheld from the Secretary of Homeland Security authority to require
private employers to utilize [the E-Verify system],” and “continues to decline to make it mandatory.”\textsuperscript{92} The Tenth Circuit Court of Appeals concluded that mandating E-Verify “disturbs the balance . . . deliberately crafted by Congress.”\textsuperscript{93}

Furthermore, although “there is no charge to employers” to use E-Verify,\textsuperscript{94} this claim ignores the fact that in order to implement the process, employers incur new or additional costs for data entry, computer maintenance, high-speed internet, filing cabinets and other office equipment, and staff training. The average total cost reported by employers using E-Verify was $100 to set up the system and $400 to maintain it.\textsuperscript{95} The most frequently mentioned set-up cost by employers who use the system was training, at a reported 17 percent of employers.\textsuperscript{96}

According to data compiled by Bloomberg, employers who voluntarily used E-Verify in the fiscal year ending September 30, 2008 spent about $43 million to access the Website.\textsuperscript{97} Bloomberg extrapolated this data to conclude that employers who voluntarily used E-Verify spent approximately $95 million in connection with the system in 2010.\textsuperscript{98} Moreover, if E-Verify had been mandated in the fiscal year 2010, it would have cost businesses a collective total of $2.7 billion.\textsuperscript{99} Unfortunately, $2.6 billion of this cost would have been disproportionately allocated to small businesses.\textsuperscript{100} This represents an astounding 99.7 percent of the costs of E-Verify being allocated to small businesses, since the fixed costs are spread over fewer hires.\textsuperscript{101} This unequal distribution costs small businesses an average of $147 per person to perform a new-hire check, compared to approximately $73 for larger firms.\textsuperscript{102}

The imposition of such unfairly distributed costs may undercut business competitiveness in the United States.\textsuperscript{103} Small businesses, which operate on lower margins, could potentially be driven out of business, while larger businesses may forgo other investment opportunities due to the cost of employment.\textsuperscript{104} These effects are not likely to contribute to the nation’s economic recovery. When the government claims that E-Verify is free, these costs are not taken into account. The process might be free, but if the government mandates the use of E-Verify, E-

\textsuperscript{92} Id. at 768.

\textsuperscript{93} Id.


\textsuperscript{95} WESTAT, E-VERIFY PROGRAM EVALUATION, supra note 71, at 183–84.

\textsuperscript{96} Id. at 183.


\textsuperscript{98} See id.

\textsuperscript{99} See id.

\textsuperscript{100} See id.

\textsuperscript{101} See id.

\textsuperscript{102} Id.

\textsuperscript{103} GIMPEL & EDWARDS, supra note 12, at 87.

\textsuperscript{104} Id.
Verify could be incredibly burdensome on businesses that may not be able to afford additional expenses.

Further, if an employer using E-Verify receives a tentative nonconfirmation and the employee chooses to contest the results, it typically takes between 7.6 and 12.5 days to resolve the nonconfirmation.\textsuperscript{105} During this time period, “[e]mployers may not terminate, suspend, delay training, withhold or lower pay, or take any other adverse action against an employee based on the employee’s decision to contest the [tentative nonconfirmation].”\textsuperscript{106} However, 86 percent of the tentative nonconfirmations typically result in a final confirmation of employment ineligibility,\textsuperscript{107} implying that the employer wastes valuable time and resources on an employee that ends up terminated as a result of the final confirmation. Some employers would rather not waste this valuable time and money, and may terminate a worker’s employment even though they cannot do so legally. In Arizona, a survey found that approximately 33.5 percent of the immigrants surveyed were fired after receiving a tentative nonconfirmation, and were not informed about the tentative nonconfirmation or given a chance to correct it.\textsuperscript{108}

It is unlikely that eight million unauthorized aliens are simply going to leave the country in response to mandating E-Verify.\textsuperscript{109} Approximately two-thirds of unauthorized aliens currently pay payroll taxes, which generated approximately $12 billion to the Social Security trust fund in 2007, creating a “net benefit of somewhere between $120 billion and $240 billion from unauthorized [aliens] by 2007.”\textsuperscript{110} However, in response to the costs imposed by immigration policies, some businesses have simply shifted employees off the tax rolls.\textsuperscript{111} After Arizona mandated E-Verify, The Arizona Republic reported that “[u]ndocumented workers and employers in Arizona are finding ways to circumvent the state’s employer-sanctions law by turning to the underground, or cash, economy.”\textsuperscript{112} Undocumented aliens are “performing services . . . on the side for cash . . . [or] borrowing the identities of citizens or legal residents to land jobs.”\textsuperscript{113} Thus, this phenomenon is

\begin{footnotesize}
\begin{itemize}
  \item \textsuperscript{105} Westat, E-Verify Program Evaluation, supra note 71, at 92.
  \item \textsuperscript{107} See Westat, E-Verify Program Evaluation, supra note 71, at xxix.
  \item \textsuperscript{109} See Hearings, supra note 54, at 1.
  \item \textsuperscript{110} Id. at 5.
  \item \textsuperscript{111} Manuel Pastor, By the Numbers: Business, E-Verify and the California Economy, Huffington Post (Oct. 6, 2011, 5:32 PM), http://www.huffingtonpost.com/manuel-pastor/by-the-numbers-business-e-b_998228.html.
  \item \textsuperscript{113} Id.
\end{itemize}
\end{footnotesize}
debilitating to the local economy because it “is lowering income tax revenues to
the state because the incomes generated through these informal means are
unreported and untaxed.”114

Although it is not possible to accurately estimate how many unauthorized
aliens are resorting to these measures,115 “[s]everal economists agree that a shift to
the underground economy . . . is worsening Arizona’s economic problems by
further deflating income-tax revenues.”116 “Arizona took in about $371 million less
in individual income-tax collections from January to September of [2008]
compared with the same period [in 2007], [which was] a drop of 13 percent,
according to the state’s Joint Legislative Budget Committee.”117

These findings were echoed in a study by the Public Policy Institute of
California, which found that newly enacted immigration legislation did lead to a
small exodus of unauthorized aliens, but the unintended consequence was that
many employees were pushed into the underground economy.118 Moreover, in
examining a 2008 bill that mandated E-Verify, the Congressional Budget Office
found that nationally mandating E-Verify would decrease federal revenue by more
than $17 billion over ten years—because it would increase the number of workers
who turned to the underground economy beyond the scope of the United States tax
system.119 This decrease in federal and state tax revenues will likely increase the
current federal budget deficits and further cripple our already fragile economy.

C. E-Verify’s Potential for Increased Discrimination

Mandating E-Verify likely encourages discrimination. Under IRCA, Congress
attempted to limit employment discrimination by “forbidding employers from
requesting more or different documents . . . or refusing to honor documents
tendered that on their face reasonably appear to be genuine . . . .”120 IRCA
incorporates many antidiscrimination provisions,121 but even with these provisions
there has been a noticeable pattern of discrimination against authorized workers.122

---

114 JUDITH GANS, ARIZONA’S ECONOMY AND THE LEGAL ARIZONA WORKERS ACT 18
115 Id.
116 González, supra note 112.
117 Id.
118 MAGNUS LOFSTROM ET AL., PUB. POLICY INST. OF CAL, LESSONS FROM THE 2007
119 Letter from Peter R. Orszag, Dir., Congressional Budget Office, to John Conyers,
Chair, Comm. on the Judiciary, U.S. House of Representatives (Apr. 4, 2008), available at
120 Chamber of Commerce of U.S. v. Edmondson, 594 F.3d 742, 767 (10th Cir. 2010).
122 See WESTAT, E-VERIFY PROGRAM EVALUATION, supra note 71, at 3 (summarizing
a 1990 General Accounting Office report following IRCA’s passage finding “that the
In contrast to IRCA, Utah’s statutory provision mandating E-Verify fails to protect authorized workers with antidiscrimination provisions.123

It is thought that mandatory E-Verify usage “increases discrimination against workers who look or sound ‘foreign,’” and that harsher sanctions for violations will further exacerbate the possibility of this type of discrimination by employers.124 Even though this argument may be “essentially speculative,”125 it is highly probable that when businesses consider the costs associated with a new-hire check, some may attempt to mitigate these costs by hiring someone who has a perceivably better chance of receiving a tentative confirmation of eligibility to work.

E-Verify’s purpose is not to function as a prescreening tool for job applicants. However, between 5 and 13 percent of employers admit to prescreening job applicants.126 Based on employee reports, this number is probably higher, between 19 and 40 percent.127 Using E-Verify, even the Social Security Administration (“SSA”) screened 25 percent of new hires prior to their actual hire date.128 Thus, some workers are never given an employment chance because of E-Verify, and it is highly likely that discrimination is exacerbated because of the system.

Finally, the actual E-Verify system itself can be discriminatory. E-Verify’s accuracy rate is diminished for individuals who are foreign-born citizens and legal immigrants.129 “[F]oreign-born workers are considerably more likely to receive [a tentative nonconfirmation] than are U.S.-born workers.”130 The error rate was only 0.3 percent for U.S. citizens in contrast to a one percent error rate for lawful permanent residents and a 5.3 percent error rate for aliens with authorization to work in the United States.131 Likely reasons for the increased error rates for foreign-born workers include

that employers are more likely to make mistakes when entering foreign-sounding names than in entering names with which they may be more
familiar, causing more nonmatches during the verification process for foreign-born workers . . . [and] despite instructions to the contrary, many foreign-born workers may list their date of birth in day-month order, resulting in nonmatches on that variable.\textsuperscript{132}

E-Verify is likely discriminatory on several levels. Discrimination may occur through employers and prejudicial hiring practices or through the system itself and inaccuracies related to foreign names. In an economy where many are currently struggling to acquire and maintain jobs, potential discrimination will likely hurt many prospective employees who are legally authorized to work in the United States.

\textbf{D. E-Verify Lacks Employer Participation}

Fourth, many employers, even those mandated to do so, are reluctant to participate in E-Verify. Barriers to participation include lack of knowledge or familiarity of the system,\textsuperscript{133} concerns with the system itself,\textsuperscript{134} and a lack of enforcement of the E-Verify requirements and related violations.\textsuperscript{135}

The primary barrier to employer participation in E-Verify is the lack of employer knowledge and familiarity of the system.\textsuperscript{136} In a study conducted by Westat, 63 percent of the 503 employers surveyed were not familiar with E-Verify.\textsuperscript{137} One employer stated “[they] have not used the verification system so far because this is the first time [they] have ever heard of it.”\textsuperscript{138} Large employers (those with over 100 employees) are significantly more likely to be familiar with the system than small employers are.\textsuperscript{139} However, given that small businesses comprise approximately 89 percent of the businesses in the United States,\textsuperscript{140} the amount of employers familiar with E-Verify in the United States is likely to be very small.

Of the employers who are aware of E-Verify, many have failed to implement the system because of concerns related to the system itself. Many feel E-Verify is not worth the implementation hassles, predominately stating concerns with costs, time consumption, and lack of perceived benefit.\textsuperscript{141} Some employers also believe

\begin{footnotesize}
\textsuperscript{132}Id. at 210.
\textsuperscript{134}Id. at 20–25.
\textsuperscript{135}Rosenblum, supra note 108, at 10–11.
\textsuperscript{136}WESTAT, PRACTICES AND OPINIONS, supra note 133, at 17.
\textsuperscript{137}Id. at 18.
\textsuperscript{138}Id.
\textsuperscript{139}Id. at 18–19.
\textsuperscript{140}Id. at 18.
\textsuperscript{141}Id. at 21.
\end{footnotesize}
that illegal immigration should be a government and not a business priority, and that “[o]ur government has a history of allowing illegal immigration to flourish and then relying on the businesses to bear the cost and exposure of controlling it.”\footnote{Id.} Other employers voice concerns that this practice will “encourage illegal activity [such as identity theft] as a means for survival” because “[t]here are hundreds of thousands of noncitizens who have been hard-working, productive, tax-paying contributors to our country under fraudulent work-authorization documents” and “[w]e must consider the implications for them and their families, and fair treatment before implementing such a program.”\footnote{Id.} Still, other small businesses have retained their employees for many years and feel no need for the system. Finally, there are employers who strongly resist E-Verify because they have experienced problems with government Internet systems in the past and have no confidence that the system will work properly: “[i]t’s just easier to complete the I-9 forms with supporting documentation, as we can control it at our local level rather than have issues with an Internet-based system.”\footnote{Id.}

A final barrier to employer implementation is the lack of monitoring and enforcement of the requirements of E-Verify. The United States Citizenship and Immigration Services (“USCIS”) created a Monitoring and Compliance Branch in 2007 to examine employer compliance and misuse of E-Verify.\footnote{Rosenblum, supra note 108, at 10.} Yet the USCIS lacks authority to enforce E-Verify requirements.\footnote{See GAO REPORT, supra note 87, at 30.} The USCIS only has the authority to terminate an employer’s access to the E-Verify system, and this punitive measure does not help the USCIS ensure that only authorized workers are being hired.\footnote{Id. at 22.}

Furthermore, ICE devotes very limited resources to enforcement.\footnote{Id. at 30–31.} Between December 2008 and August 2010, USCIS referred only three cases to ICE, and as these cases “were not significant threats or egregious violations,” senior ICE officials did not institute a full investigation.\footnote{Id. at 31 n.48.} In addition to this, USCIS is unable to determine from the E-Verify transaction date whether individuals who process E-Verify queries are the same individuals that are E-Verify registered users, because passwords to the system are easily borrowed.\footnote{Id. at 29.} An effective employer authorization system requires a credible enforcement and monitoring program to ensure that employers comply with the system, which is not currently available.
either nationally or locally. Because of the lack of enforcement mechanisms, employers’ lack of knowledge regarding the system, and concerns with the system itself, mandatory E-Verify implementation is likely to meet continued resistance.

IV. MODIFICATION OF THE CURRENT SYSTEM

Unless drastic measures are taken, such as building a fence along the United States border or authorizing border patrol agents to shoot any illegal immigrant attempting to enter the United States, illegal immigration will likely continue. It is logical to conclude that illegal immigration must be discouraged through different means. The problem of illegal immigration is likely exacerbated by the demand for low-wage workers. Thus, immigration law should attempt to reduce demand for unauthorized workers by reducing the availability of jobs within the United States.

A different approach, also aimed at discouraging illegal immigration from the sending countries, would be to provide economic assistance for countries that send the largest amount of unauthorized workers. This may be accomplished by reallocating a portion of the money currently spent on enforcing existing immigration laws, or by encouraging United States corporations to boost investments in those countries through government incentives.

Some advocates for E-Verify and its improvement propose using biometric identifiers, such as fingerprints, to reduce identity theft and fraud. These proponents suggest that incorporating biometrics into the E-Verify system will become more practical as technology progresses, and should be continuously explored. Possible methods include integrating fingerprints into an individual’s driver’s license or using a photo-screening tool. However, this too has the potential to burden employers with additional costs for equipment such as fingerprint readers. These drawbacks should be thoroughly examined prior to implementation. Experts estimate that it may cost up to $285 billion to issue biometric identifiers to current authorized workers. In addition to the costs of implementing E-Verify, biometric identifying costs could further drain already struggling businesses.

Further, even with hours of training, proper fingerprinting techniques may not be achievable. Typically, even with proper techniques, 20 percent of fingerprints taken are rendered unusable. This can be particularly true for

153 Id. at 180.
154 Hearings, supra note 54, at 7.
155 Westat, E-Verify Program Evaluation, supra note 71, at 245.
156 Id.
157 Id.
158 Hearings, supra note 54, at 7.
159 Id. at 8.
160 Id.
workers who perform manual labor because the tips of their fingertips may be worn. As many workers in manual labor fields tend to be immigrants, fingerprinting could add more discrimination to the already discriminatory E-Verify procedures. Thus, adding a biometric identifier will likely increase the obstacles with E-Verify and should only be examined and evaluated, but not yet implemented.

It has become clear that throughout the recent financial crisis, regulation of business markets—including the labor and employment market—is necessary to ensure these markets, and the economy as a whole, function efficiently and effectively. However, the particular design and implementation methods of such regulations can affect society, ultimately determining whether the regulations “benefit society or work against it.” Regulations that tend to work against society typically force compliance with certain practices and mandate a single approach to fulfilling a goal, inflicting burdensome costs on private companies. Businesses typically resist this type of regulation.

The various E-Verify legislation proposals discussed earlier in this Note appear to take this sort of all-or-nothing approach, by mandating online-status verification and forcing employer cooperation. The legislation has evoked controversy and received limited employer participation. The solution is to either alter the legislation itself or to delay the legislation to allow time to resolve many of the E-Verify system’s current problems. The next section discusses several proposed improvements to the E-Verify system or the program’s rules and procedures.

A. Suggested E-Verify System Improvements

As previously discussed, there are problems associated with the E-Verify system itself. Necessary improvements include: (1) USCIS should collaborate with other government databases to correct E-Verify inaccuracies; (2) an employer tutorial should be coupled with the system; (3) a complaint and redressability process should be paired with the system; (4) employer education efforts should be increased; and (5) incentives should be used to increase E-Verify usage.

First, USCIS should focus on reducing the inaccuracies correlated with E-Verify and should work collectively with the SSA and Department of Homeland Security (“DHS”) databases to correct inaccurate or inconsistent information

---

161 Id.
163 Id.
164 Id.
165 Id.
within their records to reduce incorrect nonconfirmations. In addition, USCIS should actively “disseminate information to [recently naturalized citizens] . . . on the potential for name mismatches and how to record their names consistently when providing name information to employers, SSA, and DHS.”167 If recently naturalized citizens are empowered with information regarding the potential for discrimination based on a foreign name, they hopefully will record their name in a consistent manner with multiple government agencies—thereby limiting the likelihood of discrimination based on the spelling or hyphenation of the name.168 These steps would aid in reducing concerns surrounding the accuracy of E-Verify.

Second, USCIS should “develop[] an analysis plan for the mastery test [currently used in connection with E-Verify], and use[] the analysis results to make fact-based decisions about whether and how to revise the test, the [E-Verify] tutorial, or both.”169 USCIS currently “administers the mastery test to newly registered users who have completed the E-Verify tutorial,” which records the number of correct and incorrect answers provided by users.170 However, USCIS has not analyzed the test data to determine if there are patterns in the data indicating that users did not understand certain topics in the tutorial.171

The tutorial currently “provides information on a variety of topics, including how to initiate and close a case, protect employees’ personal information, and properly verify employee eligibility through E-Verify, and on the potential consequences associated with discriminatory behavior and other misuse of E-Verify.”172 The tutorial should also include information about how to record names correctly—especially foreign names—which would assist in reducing the name-related nonconfirmations and potentially reduce the inherent discriminatory nature of the system.173 There should also be a readily accessible help line for employers who have specific concerns with the system or with how to conduct an accurate employee status verification check. “By developing an analysis plan for the mastery test, and using the analysis results to make fact-based decisions about whether and how to revise the test, the tutorial, or both, USCIS would be able to

166 See GAO REPORT, supra note 87, at 55. DHS has stated that it will attempt to ensure that “inaccuracies in various systems’ source data are corrected” and that “E-Verify provides the most accurate and up-to-date information on immigration status.” Id. at 58.
167 Id. at 55.
168 Id. at 55, 57.
169 Id. at 55.
170 Id. at 28.
171 Id. “USCIS has not done this because, according to senior E-Verify program officials, the computer program used to analyze responses to the test was not able to generate retrievable information on how test takers responded to individual questions.” Id.
172 Id.
173 See supra Part III.C.
better target its education efforts and ensure employer compliance with the E-Verify program.”

Third, a complaint and redressability process should be established in connection with E-Verify. This can be done at either the state or national level. There should be a process to assist workers who feel they have suffered discrimination or an illegal adverse employer action, since approximately “66 percent of workers face adverse action from their employer when they receive a [tentative nonconfirmation].” Coupled with this, employees should have the ability to access and correct personal information and inaccuracies in the governmental systems and databases operated by USCIS, the SSA and DHS. Currently, the USCIS is “developing an E-Verify Self-Check program which will allow individuals to check their own work authorization status against SSA and DHS databases prior to applying for a job.” These efforts should continue and will likely assist in reducing losses of valuable time and money by both the employer and employee due to tentative nonconfirmations and corresponding adverse actions.

Fourth, employer education efforts should be made to encourage greater employer compliance. This should include an information packet distributed when new businesses are organized or incorporated. For already established businesses, an information packet may be included with tax notices. The information packet should include information about the importance of using E-Verify, the benefits of E-Verify (including any tax or other monetary benefits implemented by the legislature in connection with compliance), and any potential penalties associated with misuse or nonuse of the system. As mandating E-Verify requires pro-action, not reaction, on the part of employers, it is especially important to increase awareness of the system itself. Increasing employer awareness will likely increase employer participation, decreasing the problems associated with lack of participation.

Finally, incentives should be used to increase employer participation—and to increase consistent usage of E-Verify by participating employers. These incentives may include tax benefits or public acknowledgment (and thus free positive publicity) of the businesses that use the system. These incentives can be

---

174 See GAO REPORT, supra note 87, at 29. However, even with these steps, there may still be problems with incorrect system use because “individuals who have not completed the tutorial and mastery test are able to borrow the user names and passwords of certified coworkers to gain access to the system.” Id.

175 Hearings, supra note 54, at 10–11.

176 Id. at 11.

177 See GAO REPORT, supra note 87, at 55.

178 Id. at 58. “While this should help individuals avoid receiving an erroneous TNC when they are hired for a job, it will not benefit newly hired employees who have already received such a TNC.” Id.

179 See supra Part III.B.

180 See supra Part III.D.
implemented through legislation, as discussed below, through state agencies, or through the E-Verify system itself.

**B. Suggested E-Verify Legislation Improvements**

E-Verify can also be improved by modifying legislation associated with it. Beneficial improvements include: (1) fairly disseminating the costs of E-Verify among businesses; (2) making E-Verify applicable only to new hires; (3) nationally mandating E-Verify on a rolling basis; (4) enacting legislation that prohibits the prescreening of applicants; and (5) creating an E-Verify oversight board.

First, USCIS should carefully analyze the actual costs to both large and small employers. Recognition of these costs can facilitate legislation that disseminates costs fairly among small and large employers. This could come in the form of tax incentives to smaller businesses that currently bear a disproportionate share of the costs.

Second, E-Verify should apply only to new hires for any state or national mandate. This would alleviate a significant burden private businesses would bear if forced to reverify the entire workforce. Yet it would still allow for timely verification of the majority of the labor force, since the nation’s typical employment turnover rate is approximately 40 percent per year—equivalent to approximately “50–60 million employees hired each year.”

Third, if E-Verify is mandated nationally, it should be implemented on a rolling basis—with businesses that have the most employees being affected first. As discussed previously, larger businesses face fewer costs related to the system because most large businesses have infrastructure currently in place. This would allow smaller businesses the opportunity to plan for and absorb the necessary start-up costs over a longer period, mitigating the initial financial impact.

Fourth, legislation should include provisions requiring employers to screen all future employees on a neutral basis, and prohibiting the prescreening of workers. To ensure that employers comply with this, an oversight and penalty structure should be enacted. This would discourage potential employer discrimination, a problem Congress has already deemed important enough to warrant protective legislation.

Finally, although USCIS created a Monitoring and Compliance Branch in 2007, it lacks authority to enforce E-Verify requirements. Thus, USCIS should amend its policies, or individual state legislatures should create oversight boards.

---

181 See *supra* notes 98–104 and accompanying text.
182 *Hearings, supra* note 54, at 9.
183 See *supra* notes 102–104 and accompanying text.
184 *Hearings, supra* note 54, at 10.
185 Id.
186 See *supra* notes Part III.C.
187 See *supra* notes 146–148 and accompanying text.
having the authority to conduct “random audits of the program that include . . . a
review of employer compliance with E-Verify requirements, a review of the adequacy of E-Verify rules and procedures to protect authorized workers, and a
review of whether the program is being managed in a way that appropriately
dresses civil rights and civil liberties concerns.”\textsuperscript{188} This would help ensure
compliance with the legislation, as an effective employer authorization system
requires credible enforcement and monitoring, and this is not available today either
nationally or locally.\textsuperscript{189}

States, such as Utah, that have already mandated E-Verify should re-evaluate
their current policies, as the current legislation may actually impede growth of the
local economy. Creating incentives to encourage employer participation and
reducing sanctions against nonparticipating employers will likely benefit the
legislation. In addition, the requirements currently surrounding E-Verify are not
comprehensively enforced within the state, and should remain that way until the
system can be improved.

Overall, until E-Verify is availed of its ills it should not be mandated on a
national level, especially during a time of economic uncertainty. With our current
fragile economy, implementing a system that could have sweeping detrimental
consequences will only hinder economic recovery.

V. CONCLUSION

E-Verify is an innovative, electronic-record search that has the potential to
meet employers’ needs in complying with future unauthorized-alien laws that may
be enacted throughout the next decade. Undoubtedly, as more and more businesses
transition to “paperless” operations, participation in E-Verify will become less
burdensome. Businesses will have the necessary tools to implement the program,
such as trained data-entry employees, Internet access, and computer hardware.

E-Verify has strong potential for businesses, but it is currently a work-in-
progress. The Form I-9 has performed well until now, and will likely continue to
perform well. Until more E-Verify problems can be adequately addressed,
businesses should be allowed to choose whether to use E-Verify or the standard I-9
form. Once some of the E-Verify issues are worked out, including its
discriminatory effects, a policy of mandatory requirement should be revisited.
However, until that time, and especially in our current ailing economy, Congress
should stop the pursuit of mandating E-Verify.

\textsuperscript{188} \textit{Hearings}, supra note 54, at 10.
\textsuperscript{189} \textit{See supra} notes 146–151 and accompanying text.
INCENTIVE FOR INNOVATION OR INVITATION TO INHUMANITY?:
A HUMAN RIGHTS ANALYSIS OF GENE PATENTING AND THE CASE OF MYRIAD GENETICS

Laurie E. Abbott*

I. INTRODUCTION

Eighteen years ago, the General Agreement on Tariffs and Trade (GATT) concluded its near decade-long Uruguay Round of multilateral trade negotiations.¹ Not only did the Uruguay Round transform the realm of international trade by creating the World Trade Organization (WTO), but, in a controversial move, it also brought intellectual property rights under the protective arm of the WTO.² In 1994, at the conclusion of the Uruguay Round, the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) was put into force, requiring all members of the WTO to adhere to measures providing significant protections for intellectual property rights.³

Though many applauded the creation of TRIPS, the agreement also saw instant backlash, not only by critics who saw this as a move away from the WTO’s trade mission,⁴ but also by human rights activists concerned that this new intellectual property regime would quash certain basic human rights,⁵ such as the right to health, food security and access to information.⁶ As a result of these human rights concerns, the WTO adopted the Doha Declaration on the TRIPS Agreement and Public Health in 2001, in which the WTO acknowledged the concerns that some intellectual property protections could harm human rights like access to healthcare and medicines. It therefore created some flexibility in the application of the TRIPS principles by allowing member states to overlook some patent rights that may impinge on access to essential medicines.⁷

* © 2012 Laurie E. Abbott. J.D. Candidate (2013), S.J. Quinney College of Law, University of Utah.

³ Id.
⁴ JAGDISH BHAGWATI, IN DEFENSE OF GLOBALIZATION 183 (2004).
Thus began the conflict between intellectual property rights and human rights. Though this conflict has mostly revolved around the clash between pharmaceutical patent rights and the right of access to medicines, the battle has entered additional fields like environmental law and food security. This Note examines one area of the conflict—the right to patent genes versus the human right of access to the human genome. This human rights perspective holds that the human genome and its countless components are the property of the entire human population, such that no person or entity can claim ownership to it. Specifically, this Note analyzes the key court decisions related to patenting of live, biological material to determine whether any human rights arguments have been made against such patents. Additionally, this Note analyzes from a human rights perspective the recent Federal Circuit Court of Appeals case involving Myriad Genetics, which questioned a company’s ability to patent a gene related to breast and ovarian cancers. This analysis considers the human rights issues related to genetic research and evaluates possible human rights violations that could result from patenting genetic material.

This Note proceeds in four Parts. Part II begins with an overview of the background and history of the human rights-intellectual property rights conflict, including a discussion of the arguments from both camps. Part III discusses international actions taken to establish a human right to the human genome and the implications this has on the ability to patent genes. Part IV analyzes some of the past landmark cases related to the patenting of biological material to determine whether any human rights implications were considered when granting those patents. Included with this final section is a discussion of the Myriad Genetics case to determine whether the recent decision of the Federal Court of Appeals acknowledged any human rights concerns with human gene patenting and what impact the human rights perspective may have had on the outcome of the case. Part V concludes.

---


9 One well-devised definition of a human rights approach to evaluating the effects of intellectual property rights explains that “[a] human rights approach takes what is often an implicit balance between the rights of inventors and creators and the interests of the wider society within intellectual property paradigms and makes it far more explicit and exacting. A human rights approach is predicated on the centrality of protecting and nurturing human dignity and the common good. . . . Or to put the matter another way, from a human rights perspective, intellectual property protection is understood more as a social product with a social function and not primarily as an economic relationship.” Audrey R. Chapman, *The Human Rights Implications of Intellectual Property Protection*, 5 J. INT’L ECON. L. 861, 867 (2002).
II. BACKGROUND AND HISTORY OF THE CONFLICT BETWEEN INTERNATIONAL HUMAN RIGHTS AND INTELLECTUAL PROPERTY RIGHTS

A. The History of the Conflict

As mentioned above, the WTO’s TRIPS Agreement was the first important step in creating international regulations for intellectual property rights. Prior to the creation of the TRIPS agreement, efforts to harmonize intellectual property protections were generally unsuccessful, and efforts that did succeed were typically bilateral treaties, resulting in a disorganized system of intellectual property protections that lack uniformity and consistency. Beginning in the 1980s, the need for greater international intellectual property protections became increasingly evident, as rampant piracy “was undermining key industries that relied on intellectual property protection, such as the pharmaceutical industry, the music industry, the computer software industry, the publishing industry, and the motion picture industry.”

Initial efforts by the World Intellectual Property Organization (WIPO) to create multilateral intellectual property agreements and update prior intellectual property conventions broke down in the 1980s. WIPO’s inability to succeed at updating international intellectual property protections was due to several factors, including the increased number of developing nations participating in WIPO, the “ideological split between socialist block and the non-socialist,” fear of American dominance in intellectual and cultural production, and the United States’ unwillingness to make any concessions that would require it to change its own intellectual property laws.

In the end it was the GATT that was able to achieve what WIPO could not—the creation of a multilateral treaty on intellectual property rights protection—the Agreement on Trade-Related Aspects of Intellectual Property Rights. As a creation of a trade organization, the TRIPS Agreement was intended to only regulate those aspects of intellectual property that affect trade, though many have criticized TRIPS for distorting the WTO’s trade mission, and instead turning it into

---

11 See id.
12 Id. at 24.
14 Oman, supra note 10, at 19.
15 Id. at 20–21.
16 TRIPS Agreement, supra note 2.
a “royalty collection agency.”17 Despite this apparently odd marriage of trade and intellectual property, TRIPS remains the most influential international intellectual property rights agreement, “mak[ing] protection of intellectual property rights an integral part of the multilateral trading system . . . .”18

TRIPS obligations require the highest level of intellectual property protection seen in any international agreement, and since all WTO members are bound by the TRIPS obligations, the Agreement provides robust protection for intellectual property rights throughout the world. Furthermore, the binding power of the WTO’s Dispute Settlement Body decisions means that members who violate the TRIPS Agreement could face devastating repercussions such as being required to pay compensation to or possible retaliation by the injured country.19

TRIPS obligations parallel many of the general WTO free trade requirements, including requiring each WTO member to grant all other WTO members “treatment no less favourable than that it accords to its own nationals with regard to the protection of intellectual property,”20 as well as most-favored-nation treatment21 with regard to the protection of intellectual property.22 Additionally, TRIPS provides for numerous intellectual property-specific requirements including extensive copyright protection,23 protection of trademarks,24 integrated circuit design, and most importantly for the purposes of this Note, protection of patents for essentially all types of inventions with a minimum patent term of twenty years.25 Although the TRIPS agreement affords members discretion to exclude some types of inventions from patentability,26 it requires that “patents shall be available and patent rights enjoyable without discrimination as to the place of

17 BHAGWATI, supra note 4, at 82–83, 182–83.
19 RALPH H. FOLSOM, INTERNATIONAL BUSINESS TRANSACTIONS (PRACTITIONER TREATISE) § 9.6 (2011).
20 TRIPS Agreement, supra note 2, at art. 3, para 1.
21 The WTO defines most favored-nation-treatment as follows: “Under the WTO agreements, countries cannot normally discriminate between their trading partners. Grant someone a special favour (such as a lower customs duty rate for one of their products) and you have to do the same for all other WTO members.” Principles of the Trading System, WORLD TRADE ORG., http://www.wto.org/english/thewto_e/whatis_e/tif_e/fact2_e.htm (last visited Feb. 27, 2012).
22 TRIPS Agreement, supra note 2, at art. 4.
23 For example, TRIPS provides fifty years of copyright protection for sound recordings and motion pictures. See id. at art. 12.
24 Id. at art. 15.
25 Id. at arts. 27–34.
26 Specifically, WTO members may choose to exclude from patentability “(a) diagnostic, therapeutic and surgical methods for the treatment of humans or animals; (b) plants and animals other than micro-organisms, and essentially biological processes for the production of plants or animals other than non-biological and microbiological processes.” Id. at art. 27, para. 3(a)–(b).
invention, the field of technology and whether products are imported or locally produced.” Upon its passage in 1994, the TRIPS Agreement applied immediately to all WTO members. However, in perhaps its greatest concession to the disgruntled developing countries, the Agreement provided a transition period of four years (until January 1, 2000) for developing countries and transition economies, and an eleven-year transition period (until January 1, 2006) for least-developed countries.

Though many lauded the TRIPS Agreement as an important and crucial step toward the protection of intellectual property rights, TRIPS was also scrutinized as an unfair policy toward developing countries, resulting in human rights violations against some of the world’s poorest people. Specifically, human rights advocates criticize TRIPS for putting private interests above basic rights “including the right of everyone to enjoy the benefits of scientific progress and its applications, the right to health, the right to food, and the right to self-determination . . . .” Recognizing the tension that was developing between the human rights and intellectual property rights regimes, several international organizations began trying to ease this tension. For example, on November 9, 1998, WIPO organized a daylong panel discussion titled, “Intellectual property and human rights,” at which world intellectual property leaders discussed the causes of the conflict, and proposed various solutions.

The most prominent advocate of advancing the human rights agenda in the intellectual property discourse was the United Nations. Beginning in 1999, the United Nations made several gestures intended to promote greater adherence to international human rights principles by intellectual property regimes. Specifically, in both its 1999 and 2000 Human Development Reports, the United Nations Development Programme “identified circumstances attributable to the implementation of the TRIPS Agreement that constitute contraventions of international human rights law . . . .” Furthermore, in November 1999, the

27 Id. at art. 27, para. 1.
30 Singh, supra note 6.
United Nations Committee on Economic, Social and Cultural Rights presented statements at the Third Ministerial Conference of the WTO in Seattle, Washington. The Committee’s statement broadly encouraged the WTO “to assess the impact that trade liberalization may have on the effective enjoyment of human rights . . . .” and it specifically signaled “a strong warning against the negative consequences of the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS), particularly on food security, indigenous knowledge, bio-safety and access to health care . . . .”

As a culmination of its efforts to bring human rights to the forefront of the intellectual property dialogue, the United Nations Commission on Human Rights put forth a declaration in 2000 titled Intellectual Property Rights and Human Rights. To show that it was not a total rejection of intellectual property rights, the Declaration affirmed “the right to protection of the moral and material interests resulting from any scientific, literary or artistic production of which one is the author” pursuant to article 27, paragraph 2 of the Universal Declaration on Human Rights. However, the Declaration also revealed concerns about global intellectual property regimes by announcing that “the implementation of the TRIPS Agreement does not adequately reflect the fundamental nature and indivisibility of all human rights, including the right of everyone to enjoy the benefits of scientific progress and its applications, the right to health, the right to food, and the right to self-determination . . . .”

By acknowledging that a right to intellectual property protections exists, while at the same time pronouncing the harm that those protections may impose on other human rights like access to proper health care, the United Nations recognized the apparent conflicts “between the intellectual property rights regime embodied in the TRIPS Agreement, on the one hand, and international human rights law, on the other . . . .” In other words, this Declaration expressed the United Nations’ fear that protecting the intellectual property rights of technology related to essential medicines, food production, or other basic health provisions, would cause prices of such health provisions to increase and access to those health products to decrease in poor countries, thereby violating the human right of health of millions of people. As a result of this apparent conflict, the report called on the WTO and particularly the Council on TRIPS to “take fully into account the existing State obligations . . . .”

35 Id. at para. 4.
36 Resolution 2000/7, supra note 31.
37 Id. at para. 1.
38 Id. at para. 2.
39 Id.
under international human rights instruments . . . .

Thus, the United Nations called out the WTO by letting it know that decades before it obliged its member states to protect intellectual property rights under the TRIPS Agreement, the United Nations had obliged many, if not all, of those same member states to protect international human rights.

In response to the United Nations’ call to action, the WTO members adopted a special Ministerial Declaration during the Doha Round of trade negotiations in 2001 to clarify some of the concerns about the TRIPS Agreement’s impact on public health. Though the Declaration did not go so far as to discuss all the human rights implications and violations resulting from the TRIPS Agreement, it did make some important concessions to an area of human rights that seemed particularly threatened by the TRIPS regime: public health. The Declaration recognized the threat that intellectual property rights may have to the prevention and treatment of diseases like HIV/AIDS, tuberculosis and malaria, and in a bold statement declared that the WTO member states

agree that the TRIPS Agreement does not and should not prevent Members from taking measures to protect public health. Accordingly, while reiterating our commitment to the TRIPS Agreement, we affirm that the Agreement can and should be interpreted and implemented in a manner supportive of WTO Members’ right to protect public health and, in particular, to promote access to medicines for all.

The Declaration goes on to allow important concessions for developing and least-developed countries, particularly in the realm of intellectual property relating to pharmaceuticals. In addition to granting governments a substantial amount of discretion in their use of patented pharmaceuticals, the Doha Declaration also extended the TRIPS Agreement transition period for least-developed nations from 2006 to 2016. Though this Declaration may not have gone as far as human rights advocates would have hoped, the Doha Declaration was a significant WTO step toward acknowledging the negative implications of global intellectual property rights protections, and that in some cases, the negative impact on human rights outweighs the benefits that the intellectual property rights may have on private interests and the potential for future innovation.

As demonstrated above, the years following the TRIPS Agreement’s international debut brought the human rights-intellectual property rights debate to

40 Id. at para. 7.
42 Id. at para. 4.
43 Id. at paras. 5–7.
center stage. Though both human rights and intellectual property rights advocates have evidently recognized the validity of each other’s arguments, and though several praiseworthy negotiations, compromises and concessions have been reached, the debate between the two camps continues. In particular, disagreements remain about how far the intellectual property rights regimes should go in accommodating human right concerns. What follows is a discussion of the arguments by both the human rights and intellectual property rights advocates to show the seemingly impossible task of creating a perfect compromise.

B. The Human Rights Argument

An initial evaluation of the human rights arguments against international intellectual property rights protections presents somewhat of a conundrum, since the United Nations Universal Declaration on Human Rights, at least vaguely, recognizes intellectual property rights as a basic human right. Not only does the Declaration recognize a right to own property, but it also specifically declares that “[e]veryone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author.” Without explicitly calling this an intellectual property right, this article of the Declaration essentially defines intellectual property and declares it a universal right. Within that same article however, the Declaration also says that “[e]veryone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits.” Thus within the same article, the Declaration grants protection of both intellectual property rights and a right to freely benefit from all scientific advancement.

Clearly at odds with itself, the human rights community, in its arguments for the protection of certain human rights like basic health and food security over intellectual property rights, essentially makes a judgment call about which human rights rank higher than others. Although human rights advocates have not necessarily ignored that intellectual property rights are in themselves a basic human right, they argue that it is necessary to sacrifice some levels of intellectual property protections to guard other, more essential human rights. Specifically, as the United Nations has argued, some intellectual property protections ought to be forgone in order to reduce

impediments to the transfer of technology to developing countries, the consequences for the enjoyment of the right to food of plant variety rights and the patenting of genetically modified organisms, “bio-piracy” and the reduction of communities’ (especially indigenous communities’) control over their own genetic and natural resources and cultural values,

45 Universal Declaration of Human Rights, supra note 5, at art. 17.
46 Id. at art. 27(2).
47 Id. at art. 27(1) (emphasis added).
and restrictions on access to patented pharmaceuticals and the implications for the enjoyment of the right to health . . . . 48

As evidenced by this United Nations statement, there are several human rights that are threatened by global intellectual property protections. However, the particular right that human rights advocates most strongly defend as needing protection over intellectual property rights is that of access to essential medicines. Access to medicines is considered a central component of the basic human right to health, 49 yet nearly two billion people throughout the world lack access to essential medicines. 50 Although gaining access to medicines has always been a challenge for developing countries, after the TRIPS Agreement went into force, which expanded the “protection of pharmaceutical products and processes,” the cost of patented essential medicines like HIV/AIDS medications became exorbitantly high for developing countries, thus exacerbating the already pandemic-sized problem of treating HIV/AIDS patients in those poor countries. 51 As explained by one human rights advocate, “[t]he exclusive monopoly that owners of intellectual property enjoy for a period of time with respect to the manufacture, marketing, sale, and distribution of the medicines they produce permits pharmaceutical corporations to demand higher prices for their products.” 52 These higher prices of patented pharmaceuticals “place essential medicines beyond the reach of many developing countries.” 53

After the enactment of the TRIPS Agreement but prior to the Doha Declaration, governments of developing nations often took measures into their own hands to retaliate against the restrictive effect that pharmaceutical patents had on access to essential medicines. For example, in 1997 the South African government enacted the Medicines and Related Substance Control Amendment Act, which made essential medicines more affordable by permitting South Africa’s Minister of Health to allow the parallel importation of patented drugs and the

51 Id. at 169.
52 Id. at 180.
53 Id.
manufacturing of generic HIV/AIDS medicines, and by allowing the government to enact price-controlling measures.54

Pharmaceutical companies responded to this Act by filing suit against the South African government for patent infringements.55 In retaliation to this lawsuit, HIV/AIDS and human rights activists ignited a social movement against the pharmaceutical companies and generated enough support and influence that the pharmaceutical companies eventually dropped their suit against the South African government.56 The activists framed their movement around human rights arguments, asserting that the South African government had an obligation to “ensure adequate access to healthcare.”57 The South African conflict between health activists and pharmaceutical companies provides a shining example of how human rights movements have succeeded in containing the detrimental impact of international intellectual property rights regimes on the right to basic healthcare.

While access to essential medicines has been the most successful argument against the advancement of intellectual property protections, human rights advocates have presented additional anti-intellectual property arguments. For example, human rights advocates argue that if intellectual property rights regimes are not forced to consider human rights, researchers and inventors will focus their efforts on the most lucrative innovations, which are not necessarily the innovations that are most beneficial to the world’s poorest populations. As stated by the 1999 Human Development Report, “[i]n defining research agendas, money talks, not need—cosmetic drugs and slow-ripening tomatoes come higher on the priority list than drought-resistant crops or a vaccine against malaria.”58 Thus, because of the profit-generating motives behind intellectual property rights, innovations that will result in higher profits will take precedent over innovations that would result in greater promotion of human rights.

Furthermore, human rights advocates argue that the increased protection of intellectual property rights has intensified the technological inequalities between the developed and developing worlds. In 1999, “[i]ndustrial countries [held] 97% of all patents worldwide” and around the same time “more than half of global royalties and licensing fees were paid to the United States, mostly from Japan, the United Kingdom, France, Germany and the Netherlands.”59 Not only are the wealthy countries deciding what areas of research and development will get most attention and funding, but also the profits generated from intellectual property

55 Pharm. Mfrs.’ Ass’n of S. Afr. v. President of the Republic of S. Afr., Case no. 4183/98, High Court of South Africa (Transvaal Provincial Division).
56 George, supra note 50, at 186.
59 Id. at 68.
rights are staying in those wealthy countries. Additionally, the human rights camp argues, “patent laws pay scant attention to the knowledge of indigenous people, leaving it vulnerable to claim by others. These laws ignore . . . diversity in views on what can and should be owned . . . . The result is a silent theft of centuries of knowledge from developing to developed countries.”

Therefore, the world’s poorest people are benefiting the least from intellectual property protections, and at the same time are the recipients of the greatest harm that intellectual property protections inflict on other essential human rights.

C. The Intellectual Property Argument

Viewed by human rights advocates as defending the private interests of powerful and wealthy industries like pharmaceutical and high-tech industries, intellectual property rights advocates often face an uphill battle when defending the protection of intellectual property over the protection of other basic human rights. However, the intellectual property camp is not without sound arguments as to why intellectual property rights do not need to be sacrificed to protect and promote other basic human rights. An initial argument intellectual property advocates often bring up in the context of human rights is that the right to the protection of one’s ideas and inventions is in and of itself a human right recognized by the United Nations. Thus, while rights to health, food security, etc. are also important human rights, they should not necessarily trump the right to intellectual property. One conceptual path that has been used to support this argument is that “rights that protect the connection between a creator of an information product and the information product belong in the category of human rights because they protect the personality of the creator.”

One of intellectual property advocates’ strongest arguments asserts that in the long run, protection of intellectual property rights will actually be more beneficial to the protection of other human rights like health and food security because intellectual property protections provide incentives for innovation. The argument follows that as technology for medicines, food production, and environmental protection becomes more innovative, they will benefit people throughout the world by finding cures to more diseases, increasing the quantity and quality of food production, and creating more environmentally friendly technologies. Thus, by limiting intellectual property protections now, we are only doing a disservice to the promotion and advancement of human rights in the future. Indeed, one of the basic principles of the TRIPS Agreement is that, as a result of the Agreement, “intellectual property protection should contribute to technical innovation and the

---

60 Id.
61 Universal Declaration of Human Rights, supra note 5, at art. 27(2).
transfer of technology. Both producers and users should benefit, and economic and social welfare should be enhanced . . . .

Still other intellectual property rights advocates argue that strong intellectual property policies can actually promote development and reduce poverty. These advocates claim that “over time there are dynamic gains from the introduction of new products, information, and creative activities.” Additionally, countries with stronger intellectual property rights protections attract multinational companies, thereby increasing foreign direct investment and eventually increasing overall growth. An economic case study of Lebanon serves as an example of the potentially positive effect of intellectual property rights protection on growth and development. The study found that if the country were to strengthen its intellectual property rights it would first, benefit from increased foreign direct investment, second, it would “experience increases in product development by local firms,” third, Lebanese firms would “find it easier to enter into joint ventures and technology-sharing or product-licensing agreements with foreign firms,” and fourth, “the average quality of products and services on the market should rise.”

A final argument expressed by the intellectual property camp is that countries with stronger intellectual property rights protections experience less “brain drain” because the educated members of those countries have a stronger incentive to stay in their native country as opposed to immigrating to find better jobs. India provides an impressive example of how greater intellectual property protections improved the country’s high-tech employment opportunities. As one researcher described, “[i]n the years before strong IP laws and policies encouraged innovators in India, much of the country’s educated cadre of information technology workers left the country to work elsewhere.” However, now that India enjoys “stronger IP institutions and legislation . . . to encourage and reward local innovation, the Indian software industry employs some 500,000 software engineers, and Indian-produced software is used worldwide.”

---

65 Carsten Fink & Keith E. Maskus, Why We Study Intellectual Property Rights and What We Have Learned, in INTELLECTUAL PROPERTY AND DEVELOPMENT: LESSONS FROM RECENT ECONOMIC RESEARCH 1, 2 (Carsten Fink & Keith E. Maskus eds., 2005).
66 Id. at 3.
67 Id. at 7.
71 Id.
As evidenced by the many strong arguments on both sides of the human rights-intellectual property debate, neither side is definitively in the right, nor is there an easy solution that will lead to an obvious overall benefit. Instead the pros and cons of both the staunch protection of human rights and the expanding protection of intellectual property rights must be balanced against each other. This balancing effort has been ongoing since the signing of the Doha Declaration, as advocates on both sides of the debate have come to realize the need for some compromise.72

However, as intellectual property rights expand into ever increasing realms, the need once again to take a step back and evaluate the potentially harmful effects of those expanding rights becomes necessary. A recent example of a controversial expansion of intellectual property rights is that of gene patenting. Once considered purely natural subject matter, genes have slowly crept into the realm of intellectual property due to advanced genetic research that has enabled research and development companies to claim property rights and obtain patents for those genes. This new patenting phenomenon has, of course, been subject to great criticism.73 Critics have emerged from a variety of paradigms, one of which is the human rights paradigm. What follows is a discussion of argument of the human right to access to the human genome and what, if any role it has played in scrutinizing the right to patent genes.

III. HUMAN RIGHTS AND THE HUMAN GENOME

“The human genome underlies the fundamental unity of all members of the human family, as well as the recognition of their inherent dignity and diversity. In a symbolic sense, it is the heritage of humanity.”74 This introductory article to the Universal Declaration on the Human Genome and Human Rights captures the United Nations’ general sentiment that there is something almost sacred to humanity about the human genome. The human genome is our core connection to every other person on the planet, and according to this Declaration, it must be guarded, and research and experiments relating to it must be carefully kept in check.

The Declaration on the Human Genome was created at a time when human biological research was reaching exciting breakthroughs, and the mapping of the human genome was coming to fruition. In 1990, the Human Genome Project began

with a goal to identify and map the 20,000–25,000 genes that constitute the human genome. Once considered lofty, futuristic ideas, processes like cloning and genetic altering were becoming a reality, thereby igniting ethical concerns about genetic research. The United Nations’ response to the ever-growing bioethical issues relating to human genome research came from its Educational, Scientific and Cultural Organization (UNESCO), whose constitution requires it to promote “collaboration among the nations through education, science and culture . . .”

After four years of preparation, UNESCO’s International Bioethics Committee completed what it considered a sufficiently credible document about this controversial topic to present it as a resolution to the General Assembly. The Declaration was well received by the General Assembly, and on March 10, 1999, the General Assembly adopted the Declaration, thereby endorsing a bold bioethical statement intended to influence the future research decisions of its member states.

Several sections of the Declaration are meaningful when considering the ethical and human rights issues relating to gene patenting. Beginning in Article 10, the Declaration establishes that “respect for the human rights, fundamental freedoms, and human dignity of individuals” should prevail over any other concern involving the human genome.

Article 12(a) thus stresses the importance of making genetic information freely available to all people. This assertion apparently contradicts the logic of

77 United Nations Educational, Scientific and Cultural Organization Const. art. 1(1), Nov. 16, 1945, 4 U.N.T.S. 275, 278.
78 Lenoir, supra note 76, at 546.
80 Declaration on the Human Genome, supra note 74, at art. 10.
81 Id. at art. 12 (emphasis added).
gene patenting, which allows researchers who discover genes to have exclusive rights to research and profit-generating activities involving those genes. Upon closer observation of Article 12(b), its mandates potentially cut both ways when it comes to the ethics of patenting genes. On one hand, patents are often considered crucial for the furtherance of scientific research, thus gene patenting could promote the goals of 12(b) by encouraging advancements in human genome research. On the other hand, gene patents create limitations as to who is allowed to perform certain types of research on those patented genes, thus directly violating the call for freedom of research. Furthermore, though patenting genes may lead to more cures and remedies for genetic disorders in the long run, thereby offering “relief from suffering and improving the health of individuals and humankind as a whole,” in the short run, gene patents can limit the amount of research doctors can conduct on behalf of specific patients and on the amount of information they can convey to their patients. This result would seem to violate 12(b) by potentially causing more present suffering and reducing the likelihood of improving individuals’ health.

Finally, Article 18 also conveys affirmations pertinent to the ethics of gene patenting. This Article declares that “States should make every effort . . . to continue fostering the international dissemination of scientific knowledge concerning the human genome . . . [and] to foster scientific and cultural cooperation, particularly between industrialized and developing countries.” As mentioned above, one human rights argument against international intellectual property rights protections is that wealthy nations own the vast majority of patents, thereby causing developing nations to suffer from inaccessibility to high-priced patented goods and a lack of transmission of scientific information. Therefore, in assessing the obligations contained in Article 18, it seems that patenting genes could restrict the “international dissemination of scientific knowledge concerning the human genome”—especially to developing countries.

Apart from this United Nations Declaration, other individuals and organizations have expressed their human rights concerns about the impact of gene patents. For example, an official at the International Center for Technology Assessment to the Commissioner of Patents and Trademarks expressed his concern that “[t]he public has a strong interest in keeping the human genome freely accessible not only for the use of scientific research, but also because it is the foundation of human life and thus should not be patented by any one person.” Though this concern was about patenting the entire human genome, and not just one gene, the same concerns about keeping information about particular genes freely accessible exists.

82 Id. at art. 12(b).
83 Id. at art. 18.
84 See supra Part II.B.
85 Declaration on the Human Genome, supra note 74, at art. 18.
Similarly, one British woman, Wendy Watson, who became a health rights advocate after being diagnosed with breast cancer, helped organize the Hereditary Breast Cancer Foundation (HBCF). Watson was motivated to create the HBCF after finding out that Myriad Genetics had gained patent rights to genes related to breast cancer. Watson lobbied before the European Parliament, arguing she had been one of the research subjects that enabled Myriad to isolate the gene, and that “[n]o company should benefit commercially from that kind of research.” Watson’s group made ardent human rights arguments against gene patenting, asserting that “claims on behalf of intellectual property are becoming more and more commercial, thereby linking them with the powerful forces of pharmaceutical corporations and placing them on uneven par with the human rights claims of isolated individuals.” It appears that people like Watson may have had some influence on European policy makers because, though the United States has been somewhat more open to the idea of patenting genes, the European Union has shown greater reservations in its approach to opening patent law to the human genome. For example, as early as 1993, an E.U. Proposed Directive explicitly found that parts of the human body are unpatentable. This early legislation has resulted in the E.U. being more cautious about granting patents related to genetic material.

Despite the concerns about the human rights implications of gene patenting, U.S. patent law has gradually moved toward increased acceptance of the patentability of human genes. This steady acceptance began in 1980, when the U.S. Supreme Court affirmed the patentability of living organisms, continued in 1995 by allowing patents on human cells, and has reached the current era where over twenty percent of all human genes have been patented. In the United States’ most recent stamp of approval of gene patenting, the United States Court of Appeals for the Federal Circuit held for the first time in July of 2011 that genes are patentable subject matter. What follows is a review of U.S. jurisprudence regarding genetic

---

88 Id.
89 Id.
95 Ass’n for Molecular Pathology v. U.S. Patent & Trademark Office, 653 F.3d 1329 (Fed. Cir. 2011), reh’g denied (Sept. 13, 2011), reh’g denied (Sept. 16, 2011), cert. granted, judgment vacated sub nom. Ass’n for Molecular Pathology v. Myriad Genetics,
patenting to determine if U.S. courts have considered any human rights arguments as part of their decisions. Most importantly, the following section will evaluate what, if any, human rights arguments have been made in recent case against Myriad Genetics.

IV. HUMAN RIGHTS CONSIDERATIONS IN U.S. CASE LAW RELATED TO GENETIC AND BIOLOGICAL PATENTING

A. Human Rights in Pre-Myriad Cases

Often considered the genesis of U.S. jurisprudence related to the patentability of biological material, *Diamond v. Chakrabarty* set forth a seemingly lenient standard for such a controversial subject matter. The central question in *Chakrabarty* was whether a live, human-made bacterium was patentable. While the U.S. Patent and Trademark Board of Appeals found that as a living organism, the bacterium was not patentable subject matter, both the U.S. Court of Customs and Patent Appeals and the U.S. Supreme Court disagreed—holding that “anything under the sun” that has a man-made component is patentable subject matter. Focusing most of its energy on statutory interpretation, the Court in *Chakrabarty* paid little attention to the petitioner’s ethical arguments. The Court referenced the “parade of horribles” described in petitioner’s and amici briefs and acknowledged concerns that “genetic research may pose a serious threat to the human race” and that “genetic research and related technological developments may spread pollution and disease, that it may result in a loss of genetic diversity, and that its practice may tend to depreciate the value of human life.” However, the Court did little more than mention these concerns, and instead deferred responsibility by stating that “[w]hatever their validity, the contentions now pressed on us should be addressed to the political branches of the Government, the Congress and the Executive, and not to the courts.” Therefore, though the Court did not reject the ethical/human rights arguments, it did not factor them in to its final decision in *Chakrabarty*.

---


100 *Chakrabarty*, 447 U.S. at 309.

101 For the Court’s brief discussion of petitioner’s ethical concerns, see *id.* at 316–17.

102 *Id.* at 316.

103 *Id.*

104 *Id.* at 317.
Despite the Chakrabarty Court’s congressional call to action, Congress was surprisingly quiet after the controversial decision.\textsuperscript{105} Perhaps due to the legislature’s uncertainty about how to deal with this new scientific matter, Congress did not make any relevant legislation for seven years after the Chakrabarty decision.\textsuperscript{106} It was not until the Patent and Trademark Office (PTO) approved patents on genetically altered animals\textsuperscript{107} that Congress decided to act.\textsuperscript{108} In response to the PTO’s decision and the public outcry that followed, the House Subcommittee on Courts, Civil Liberties, and the Administration of Justice held hearings in 1987 relating to animal patents.\textsuperscript{109} Though obviously animal rights issues led the discussions, Representative Robert Kastenmeier from Wisconsin, the congressman responsible for organizing the hearings, expressed his concerns that the course that the PTO had taken might eventually lead to patents of human beings.\textsuperscript{110} Thus it is evident that Congress was anticipating future human rights concerns that would result from the continued acceptance and expansion of patents of live organisms. These hearings may have opened the dialogue and exposed many concerns about biological patents, but they failed to produce any real legislation.\textsuperscript{111}

Over a decade later, Congressional interest in biological patenting re-emerged due to concerns about the patentability of human embryos and human genes.\textsuperscript{112} One particularly interesting statement made by the director of the PTO at a 2000 House Subcommittee Hearing entitled “Gene Patents and Other Genomic Inventions” stated

\begin{quote}
the USPTO does take notice of the legitimate concerns regarding access to genomic inventions. Clearly, inventors and owners of genomic patents need to be acutely aware of the heavy responsibility inherent in that ownership; their licensing and other technology transfer practices need to strongly account for the powerful public desire to ensure that the
\end{quote}

\textsuperscript{106} Id.
\textsuperscript{107} Patent and Trademark Office Notice: Animals – Patentability, 1077 Off. Gaz. Pat. & Trademark Office 24 (April 21, 1987). Interestingly, this announcement by the PTO explicitly excluded human multicellular living organisms, thereby showing the PTO’s hesitation about patents related to human bodily parts. Id.
\textsuperscript{108} Lumelsky, supra note 105, at 660–62.
\textsuperscript{109} Id. at 661 (citing Patents and the Constitution: Transgenic Animals: Hearing on Supplemental Appropriations Act Before the H. Subcomm. on Courts, Civil Liberties, and the Admin. of Justice, 100th Cong. 2 (1987).
\textsuperscript{110} Id. at 662.
\textsuperscript{111} Id. at 663.
\textsuperscript{112} Id. at 670.
use of these inventions for the greater good of all humankind is not unduly burdened.\textsuperscript{113}

This cautionary statement provides evidence that, though the PTO was moving forward with genomic patents, it was aware of concerns that the restrictive nature of patents might hinder general accessibility to the positive outcomes of genetic research. While various congressional hearings were held during this time and human rights arguments were made, the legislature was unable to pass any bills that would limit the scope of human patenting.\textsuperscript{114}

Between these seasons of bioethical debates at Congress, the judiciary received another opportunity to weigh in on the genetic patenting controversy with the 1990 case of \textit{Moore v. Regents of the University of California}.\textsuperscript{115} In this landmark case, the California Supreme Court held that a patient whose cells had been extracted and patented by researchers had no legal rights to his patented cells or to any economic gain that resulted from the patents.\textsuperscript{116} The majority opinion gave little, if any, consideration to the ethical implications of its decision. However, in a fiery dissenting opinion, Justice Mosk expressed grave concerns with the majority’s decision.\textsuperscript{117} In a particularly poignant statement connecting the exploitation of patented genetic material with other forms of human rights violations, Justice Mosk asserted that

\begin{quote}
our society acknowledges a profound ethical imperative to respect the human body as the physical and temporal expression of the unique human persona. One manifestation of that respect is our prohibition against direct abuse of the body by torture or other forms of cruel or unusual punishment. Another is our prohibition against indirect abuse of the body by its economic exploitation for the sole benefit of another person. The most abhorrent form of such exploitation, of course, was the institution of slavery. Lesser forms, such as indentured servitude or even debtor’s prison, have also disappeared. Yet their specter haunts the laboratories and boardrooms of today’s biotechnological research-industrial complex.\textsuperscript{118}
\end{quote}

By comparing genetic patents to various forms of slavery, Justice Mosk provided a strong human rights argument against genetic patenting. This case provides further

\begin{footnotes}
\footnote{Lumelsky, \textit{supra} note 105, at 671–72.}
\footnote{\textit{793 P.2d 479} (Cal. 1990).}
\footnote{\textit{Id.} at 488–97.}
\footnote{\textit{Id.} at 506–23 (Mosk, J., dissenting).}
\footnote{\textit{Id.} at 515.}
\end{footnotes}
evidence that, while human rights arguments have continually made their way into decisions relating to genetic patents, they have generally been the minority view, and have thus had no limiting effect on the expansion of such patenting.

In 2003, the U.S. District Court for the Southern District of Florida received its chance to weigh in on the genetic patenting debate in the case of Greenberg v. Miami Children’s Hospital Research Institute, Inc.119 This case involved a dispute over a patent for an isolated gene that causes Canavan disease.120 A research physician who had received tissue from a patient suffering from the disease discovered and patented the gene.121 The court in this case, like the court in Moore, granted the defendant’s motion to dismiss, holding that the plaintiffs had no legal rights to the patented genes.122 Unlike in Moore, there was no dissenting opinion and no mention of any human rights or ethical arguments.

From these three influential cases and the congressional responses that followed, it is evident that though human rights arguments have made their way into some decisions, they are the minority view. As such, the expansion of genetic patenting has in no way been restrained by human rights concerns. During the past decade, the number of gene patents has continued to grow with little resistance. However, the smooth sailing of genetic patenting hit a rough patch of water in the form of the American Civil Liberties Union’s (ACLU) case against Myriad Genetics. The following section will review this most recent case in the fight against genetic patenting to determine if human rights have had any influence on the courts’ decisions.

B. The Myriad Genetics Case from a Human Rights Perspective

On May 19, 2009, the ACLU and the Public Patent Foundation filed a lawsuit against Myriad Genetics and the PTO asserting the unconstitutionality of patenting two human genes, the mutations of which are associated with most cases of hereditary breast and ovarian cancers.123 The main allegations listed in the complaint asserted that “human genes are products of nature, laws of nature and/or natural phenomena, and abstract ideas or basic human knowledge or thought, [therefore] the challenged [patent] claims are invalid under Article 1, section 8,

---

120 Id. at 1066. Canavan disease is caused by the lack of a certain enzyme that leads to a build up of a particular acid in the brain. This build up results in severe mental disabilities. See Canavan Disease, PUBMED HEALTH (Nov. 14, 2011), http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002553/.
121 Greenberg, 264 F. Supp. 2d at 1067.
122 Id. at 1077–78. The only claim that the court did not dismiss was plaintiff’s claim of unjust enrichment.
clause 8 of the United States Constitution and 35 U.S.C. § 101.” 124 These allegations were of consequence not just to Myriad, but to hundreds of other biotechnology companies who also hold patents to human genes.125

The plaintiffs alleged that, as a consequence of Myriad’s patents to the two genes, BRCA1 and BRCA2, and also to technology related to testing for those genes, ease of access to information and diagnosis involving breast and ovarian cancer was being restricted.126 Furthermore, plaintiffs alleged that, because Myriad “chooses not to license the patents broadly . . . [w]omen are thereby prevented from obtaining information about their health risks from anyone other than the patent holder, whether as an initial matter or to obtain a second opinion.”127 Finally, contrary to the goals of patent rights, plaintiffs asserted that “[g]ene patents can serve as a disincentive to innovation in molecular testing because they deny access to a vital baseline of genomic information that cannot be invented around.”128 Therefore, other scientists are limited in the amount of research they can perform on the patented genes because “threat of enforcement from a patent holder and ensuing litigation costs lead to a chilling effect as clinical laboratories are reluctant to develop new tests, even when new tests could directly benefit patients.”129

In their memorandum in support of motion for summary judgment, plaintiffs framed their arguments against Myriad as violations of federal and constitutional law. Their first argument asserted that Myriad’s patent claims violate 35 U.S.C. § 101130 because “human genetic sequences and the scientific inquiry of looking at a gene or comparing two human genes constitute natural phenomena, laws of nature, and abstract ideas and thus are not patentable subject matter.”131 Second, “[t]hey also constitute patents on thought, knowledge, and ideas in violation of the First Amendment.”132 Finally, “[b]ecause they patent basic scientific principles, not inventions or discoveries, they have impeded rather than advanced science and thus also violate the U.S. Constitution, Article 1, Section 8, Clause 8 (the


126 Complaint, supra note 124, ¶ 2.

127 Id.

128 Id. ¶ 88.

129 Id.

130 35 U.S.C. § 101 (2006) (“Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title.”). Plaintiffs claimed that Myriad patents violated this provision because they were not for the discovery of any new item, but instead were items that occurred in nature.


132 Id. at 1.
Plaintiffs asserted that because Myriad’s patents resulted in restrictions to rather than the progress of science and they limit thought and knowledge, those patents were in violation of the Constitution.

Though plaintiffs’ arguments relied on constitutional and federal law as opposed to international human rights law, some common threads between constitutional rights and human rights exist that make parts of these arguments consistent with human rights arguments. Namely, plaintiffs’ arguments that Myriad’s patents restrict rather than encourage scientific advancements and limit patients’ abilities to benefit from information about the patented genes are similar to the assertions made in the Universal Declaration on the Human Genome and Human Rights encouraging openness in research and access to scientific innovations. Thus, while not directly relying on human rights law to state a claim against Myriad, plaintiffs used arguments that are consistent with the human rights arguments against genetic patenting.

Defendants responded to the plaintiffs’ complaint by trying to get the suit dismissed for lack of subject matter jurisdiction, standing, and personal jurisdiction. However, the district court disagreed with the defendants, and it refused to dismiss the case. Therefore the suit moved forward, and by the time the court heard arguments on the merits in 2010, eleven parties had submitted amicus briefs either defending or refuting plaintiffs’ claims of the invalidity of the suit.

Of those eleven amicus briefs, only one made a significant human rights argument. That brief was filed by five public interest groups: the National Women’s Health Network, Asian Communities for Reproductive Justice, Center for Genetics and Society, Generations Ahead and Pro-Choice Alliance for Responsible Research. In their brief, the groups made two strong human rights arguments. First, they argued that “[g]ene patents cause harm to patients, in particular to women, by inappropriately stifling innovation and competition and interfering with health access.”Within that argument, they stated that such

133 Id. The relevant constitutional provision grants Congress the power to “promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries.” U.S. CONST. art. I, § 8, cl. 8.

134 See Declaration on the Human Genome, supra note 74.

135 See Memorandum of Law in Support of Defendants’ Motion to Dismiss, Ass’n for Molecular Pathology, 702 F. Supp. 2d 181 (No. 09 Civ. 4515); Defendant USPTO’s Memorandum of Law in Support of Motion to Dismiss, Ass’n for Molecular Pathology, 702 F. Supp. 2d 181 (No. 09 Civ. 4515).


137 See Brief for Amici Curiae Nat. Women’s Health Network, Asian Communities for Reproductive Justice, Center for Genetics and Society, Generations Ahead, Pro-Choice Alliance for Responsible Research, Ass’n for Molecular Pathology, 702 F. Supp. 2d 181 (No. 09 Civ. 4515).

138 Id. at 9.
restrictions are particularly detrimental to racial and ethnic minorities because researchers more often exploit their unique genes. Furthermore, by reducing competition and thus increasing costs of medicine, gene patents are also more detrimental to the poor. Second, the groups argued that “[h]uman genes are part of the common heritage of humanity and should not be removed from the public domain.” Within this argument, the brief discussed the Universal Declaration on Human Rights and the Human Genome, and asserted that patenting human genes is contrary to international law by violating the notion that the human genome is part of the “common heritage of humanity,” and therefore cannot be owned by any individual or corporation. Such commercial ownership would deprive the rest of humanity of its right of access to the human genome. Though these human rights arguments did not necessarily make a significant impression on the court’s decision, as will be discussed below, the court was at least provided with this information such that it could contemplate the human rights implications of its decision.

In March of 2010, the district court came down with a controversial ruling—holding that both the patents for the BRCA1 and BRCA2 genes and the patents for methods of analyzing the gene sequences were invalid. In its discussion of the parties’ arguments, the court acknowledged that there “exists a deep disagreement between the parties concerning the effects of gene patents on the progression of scientific knowledge.” Although the court did not necessarily explicitly address the human rights implications of these genetic patents, the court indirectly agreed with the possibility that if such patents do inhibit the progress of scientific knowledge, there would be an overall detriment to humanity. While the human rights arguments certainly played a minor role in the preparation for and outcome of the district court decision, that they were at least recognized indicates that human rights concerns about gene patenting remain relevant.

Despite the court’s indirect acknowledgement of the implications that gene patents have for scientific progress and the impact that has on fundamental human rights, the court based its holding on findings that Myriad’s gene patents were for phenomena of nature, and that the isolated DNA sequences were not “markedly different from native DNA as it exists in nature.” Therefore, the genes did not constitute patentable subject matter under 35 U.S.C. § 101. The court similarly held that, “because the claimed comparisons of DNA sequences are abstract mental processes, they also constitute unpatentable subject matter under § 101.” Since

References:

139 Id. at 9–11.
140 Id.
141 Id. at 16.
142 Id. at 20.
143 Ass’n for Molecular Pathology, 702 F. Supp. 2d at 184.
144 Id. at 207–08.
145 See id. at 207–11.
146 Id. at 232.
147 Id.
148 Id. at 185.
the court was able to find Myriad’s patents invalid under the natural phenomena and abstract mental processes exceptions to 35 U.S.C. § 101, the court dismissed plaintiff’s constitutional claims by way of constitutional avoidance. By dismissing plaintiffs’ constitutional claims, the court failed to consider those arguments that were most related to the human rights arguments against genetic patents.

As would be expected, the district court’s ruling stunned the biotechnology community. Whereas most indications seemed to be pointing in the direction of increased acceptance of patent rights for genetic material, the Myriad case threw a significant wrench in the progress of genetic intellectual property rights. Prior to this decision, the U.S. patent office had allowed for the patenting of many isolated and purified genes. Following the district court’s holding, many concerned biotechnology companies question whether the nearly 2,000 already existing gene patents were still valid.

What, if any, were the human rights implications of the district court’s decision? As previously mentioned, the district court acknowledged the argument that gene patents may restrict the flow of relevant scientific knowledge instead of promoting it, which the Universal Declaration of Human Rights to the Human Genome explicitly condemned. However, the court’s holding relied on the natural phenomenon exception to patent rights, and gave no real consideration to a human rights interest in gene patents. Indeed, the court chose not to address the plaintiffs’ constitutional arguments, which are perhaps the closest that the plaintiffs made to a human rights argument.

As scholars and practitioners tried to unpack the holding and implications of the district court’s decision, several made at least passing reference to the human rights implications of the decision. For example, one scholar wrote a paper entitled Interpreting Myriad: Acquiring Patent Law’s Meaning Through Contemporary Jurisprudence and Humanistic Viewpoint of Common Heritage of DNA. In that paper, the author felt that one important lesson from the district court’s opinion is that

Myriad reminded us that whenever exclusive rights are conferred upon a selected few, the majority suffers. Myriad’s product of nature doctrine is a fervent reminder that a more positive outcome might result if patenting is forever foreclosed on human genes, thereby ensuring that majority is

149 Id. at 237–38.
151 Id.
152 Id.
never deprived of a common heritage of mankind. Perhaps, the days may
not be too far, when patenting of human genes is not legally possible.\textsuperscript{154}

Thus, while the district court may not have necessarily made any human rights
connections to its decision against Myriad, many human rights advocates
interpreted the decision as a victory for its movement.

That victory was, however, short lived. Following the district court’s decision,
Myriad and its fellow defendants filed an appeal. In their appellate brief, the
appellants argued that the case should be reversed for two main reasons.\textsuperscript{155} First,
the appellants argued that the patented materials do not fit within the § 101 natural
phenomena, laws of nature, or abstract ideas exception, but rather that they are
“undisputedly compositions of matter.”\textsuperscript{156} Second, the patented research methods
that the district court also found invalid should be revalidated because “methods
that include ‘transformations’ of a human sample” and that require “extracting,
processing, and analyzing a human tissue” are patent-eligible subject matter.\textsuperscript{157}

In their response, the appellees stuck to their original arguments and the
findings of the district court, that Myriad’s patents are invalid because they
encompass materials that are natural phenomena, products of nature, and abstract
ideas.\textsuperscript{158} Though the district court had rejected appellee’s constitutional arguments,
the appellees chose to include those same arguments in their brief.\textsuperscript{159} By including
those constitutional arguments, the appellees kept intact their only arguments that
had any resemblance to a human rights argument.

In addition to the parties’ briefs, twenty-five third parties submitted amicus
briefs. Two of those briefs, both in support of the appellees, made some mention of
human rights. Interestingly, the groups that filed a human rights-influenced amicus
brief to the district court—National Women’s Health Network, the Asian
Communities for Reproductive Justice, the Center for Genetics and Society,
Generations Ahead, the Pro-Choice Alliance for Responsible Research and
Alliance for Humane Biotechnology—filed a similar appellate brief, but dropped
their “common heritage of humanity” argument.\textsuperscript{160} Instead the group focused on
what seemed to be the winning “natural phenomena” arguments, with the addition

\textsuperscript{154} Id. at 539.
\textsuperscript{155} Brief for the Appellants, Ass’n for Molecular Pathology v. U.S. Patent &
Trademark Office, 653 F.3d 1329 (Fed. Cir. 2011) (No. 2010-1406).
\textsuperscript{156} Id. at 17.
\textsuperscript{157} Id. at 53, 55.
\textsuperscript{158} Brief for the Appellees, Ass’n for Molecular Pathology, 653 F.3d 1329 (No. 2010-
1406).
\textsuperscript{159} Id. at 60.
\textsuperscript{160} Brief of Amici Curiae Nat. Women’s Health Network, the Asian Communities for
Reproductive Justice, the Center for Genetics and Society, Generations Ahead, the Pro-
Choice Alliance for Responsible Research and Alliance for Humane Biotechnology, Ass’n
for Molecular Pathology, 653 F.3d 1329 (No. 2010-1406).
of their argument about the disproportionate harm that gene patents have on women of color and low-income women. ¹⁶¹

A different amicus brief, filed by the International Center of Technology Assessment, the Indigenous Peoples Council on Biocolonialism, Greenpeace, Inc., Friends of the Earth and the Council for Responsible Genetics, picked up the human rights arguments. ¹⁶² Specifically, this brief maintained that gene patents have “significant negative scientific, social, cultural and environmental consequences.” ¹⁶³ Of particular human rights significance, the groups argued that “[p]atents on indigenous people’s genes facilitate the exploitation of indigenous peoples and violate international law.” ¹⁶⁴ The brief provided examples of remote and relatively homogenous indigenous people whose genes have been viewed as “treasure troves” by genetic researchers. ¹⁶⁵ Such indigenous groups have often felt that researchers have exploited their genes, and that their religious, cultural, and legal rights have been violated as a result. ¹⁶⁶ The brief affirmed the idea that “properly excluding gene sequences as impermissible subject matter pursuant to the product of nature doctrine would serve to protect the rights, under international and federal law, of Indigenous peoples, that are currently being violated.” ¹⁶⁷ This brief provided the court of appeals with a strong human rights concern that gene patenting could lead to the exploitation of certain susceptible groups.

The amicus brief submitted by the Southern Baptist Convention presented similar concerns about the impact of genetic patents on humanity. It began its argument by proclaiming “[t]he patenting of human genes is an affront to humanity. The possibility of obtaining a patent on a person’s genes also encourages physicians and researchers to treat people in a dehumanizing way. For many people, the patenting of genes also violates their religious beliefs.” ¹⁶⁸ While this brief was clearly religiously motivated, it also put forward arguments concerning genetic patents turning people into commodities and negatively impacting the human right to proper healthcare by “alter[ing] the relationship between individuals and researchers.” ¹⁶⁹

Beyond those two briefs, the remaining interested parties generally ignored human rights concerns, and following in suit, the July 2011 decision of the Court

¹⁶¹ Id. at i–ii.
¹⁶³ Id. at 12.
¹⁶⁴ Id. at 24.
¹⁶⁵ Id. at 25.
¹⁶⁶ Id.
¹⁶⁷ Id. at 29.
¹⁶⁸ Brief of Amici Curaie S. Baptist Convention at 2, Ass’n for Molecular Pathology, 653 F.3d 1329 (No. 2010-1406).
¹⁶⁹ Id. at 5–6.
of Appeals for the Federal Circuit made no reference to human rights. In a two-to-one decision, with Judge Bryson dissenting, the court affirmed in part and reversed in part. Specifically, the court affirmed that the method claims for comparing or analyzing the isolated DNA sequences were not patentable. However, the court reversed the district court’s decision in a significant part by holding that the actual isolated DNA sequences themselves were patent-eligible subject matter, in addition to the method claim for screening potential cancer therapeutics. To refute the claim that the isolated DNA sequences were natural phenomena and therefore not patent eligible, the court found that the DNA sequences were removed from their native environment and substantially manipulated, thereby producing a “molecule that is markedly different from that which exists in the body.”

Though it may seem that this ruling was a partial victory for both parties, in terms of the human rights implications, the very act of finding genes patentable made the outcome an outright defeat for human rights advocates. Despite this defeat, there has been little, if any, audible outcry from the human rights community following the court of appeals’ decision. Perhaps activists are waiting for the U.S. Supreme Court to determine the Myriad’s ultimate fate.

This wait may be longer than expected, however, considering the tumultuous subsequent history of the Federal Circuit’s opinion. First, in October 2011, the ACLU announced its decision to appeal the Federal Circuit opinion. After the ACLU submitted its cert petition on December 7, 2011, nine amicus briefs were filed in support of the petition. While several of these briefs focused on the standing, First Amendment, and patentable subject matter issues, some added distinctively human rights arguments. For example, one such amicus brief argued that patenting genes reduces both access to and quality of essential medical care for women—particularly minority and socio-economically disadvantaged women.

---

171 Id. at 1358.
172 Id. at 1352.
174 Petition for Writ of Certiorari, Association for Molecular Pathology v. Myriad Genetics, Inc., 653 F. 3d 1329 (Fed Cir. 2011) (No. 11-725) 2011 WL 6257250.
175 See e.g., Brief of Amici Curiae Kali N. Murray and Erika R. George, Ass’n for Molecular Pathology, 653 F.3d 1329 (No. 11-725).
176 Brief of Amici Curiae National Women’s Health Network, et al., Ass’n for Molecular Pathology, 653 F.3d 1329 (No. 11-725).
The Supreme Court eventually granted cert in March 2012, but merely to vacate the Federal Circuit opinion and remanding the case “for further consideration” in light of the Court’s decision in Mayo Collaborative Services v. Prometheus Laboratories, Inc. Justice Breyer wrote the unanimous Prometheus opinion, which held that a particular diagnostic test related to autoimmune diseases was not eligible for patents because the test was a mere recitation of a law of nature. The day the Prometheus decision came out, Myriad’s stock took a plunge—signifying fears that this ruling meant the Supreme Court would overrule the Myriad decision. Once the case was remanded, however, patent supporters felt more optimistic because the case would return to the more patent-friendly Federal Circuit.

Regardless of the outcome of the Federal Circuit’s second attempt at resolving the Myriad dispute, the case may still make its way to the Supreme Court. If the Supreme Court does eventually hear the case, Myriad may have cause for concern. Kenneth Chahine, a visiting professor at the University of Utah College of Law who has followed the case closely, stated, “I would say the ACLU probably has a better chance at the Supreme Court than they do at the Federal Circuit. . . . It’s not unusual for the Supreme Court to disagree, or at least partially disagree, with the Federal Circuit.” Furthermore, as the highest court in the United States, the Supreme Court’s decisions attract more international attention and tend to give more consideration to international law than do lower U.S. courts. As such, it is possible that the Court will be influenced by the European Patent Office’s 2004 decision to revoke Myriad’s patents on the breast/ovarian cancer genes.

Indeed, in recent years, the Court has given greater deference to international perspectives. For example, Justice Kennedy expressed in a 2010 case that “[t]he judgments of other nations and the international community are not dispositive as to [the Court’s constitutional interpretation]. But ‘[t]he climate of international opinion concerning the acceptability of [the Court’s decision]’ is also ‘not irrelevant.’” Thus with the Court’s willingness to consider international and foreign trends, perhaps the human rights groups should not raise their white flag just yet.

178 Prometheus, 132 S. Ct. 1289.
180 Id.
181 Id.
IV. CONCLUSION

Though international organizations like the United Nations have made definitive declarations regarding the human rights implications of intellectual property rights and specifically rights to the human genome, the international outcry about patenting isolated human genes has been minimal. Specifically, it is yet to be determined how the international human rights community will respond to the view of U.S. courts that genes are patentable subject matter. If the U.S. Supreme Court eventually hears the Myriad case, its decision will likely draw international attention as a landmark intellectual property decision. If the Court does hear the case, chances are the international human rights implications will not be the Justices’ primary consideration. But regardless of what decision the Court makes and what legal arguments it uses to support its decisions, it is likely that there will be a response from the international human rights community, especially if the Court finds Myriad’s patents valid.

The conflict between human rights and intellectual property rights is far from black and white, especially in the realm of gene patenting where there are obvious positive implications for encouraging the advancement of genetic research. Thus, if legal decision makers continue to move toward greater protection of gene patenting, there will be positive outcomes like discovering cures for diseases, allowing people to better understand their genetic make-up, and perhaps even completely eliminating certain genetic disorders. However, these advancements may come at a cost, both financially and in terms of the cost of limited access to valuable information. Though clearly there are economic motivations behind gene patenting, hopefully those financial incentives will not overshadow the underlying purpose for allowing patents in the first place—to motivate innovative research that will benefit all of humanity.