ARTICLES

Landscape Fairness: Removing Discrimination From the Built Environment
Stephen Clowney

Turnover Actions and the “Floating Check” Controversy
David R. Hague

Reconciling Liberty and Equality in the Debate over Preimplantation Genetic Diagnoses
Jessica Knouse

Dissent into Confusion: The Supreme Court, Denialism, and the False “Scientific” Controversy over Shaken Baby Syndrome
Joëlle Anne Moreno & Brian Holmgren

Mapping, Modeling, and the Fragmentation of Environmental Law
Dave Owen

Defining Persecution
Scott Rempell

The Skeptic’s Guide to Information Sharing at Sentencing
Ryan W. Scott

NOTES

Timothy M. Bagshaw

Paving the Road Ahead: Autonomous Vehicles, Products Liability, and the Need for a New Approach
Kevin Funkhouser
The Utah Law Review was founded in 1948 with the purpose of serving the interests of students, the bench, and the bars of the state and surrounding areas. Since then, its scope has expanded to include legal issues of importance both domestically and internationally. The Utah Law Review is a student-run organization, with all editorial and organizational decisions made by student-editors enrolled at the S.J. Quinney College of Law at the University of Utah. The Utah Law Review publishes four issues per year, with each issue containing approximately 250 pages of legal scholarship.

The Utah Law Review makes every effort to conform to the following editing sources: The Bluebook: A Uniform System of Citation (19th ed. 2010) and The Chicago Manual of Style (16th ed. 2010). The Utah Law Review board of editors strives to maintain the text of published articles in its original format by editing only for accuracy and, where necessary, clarity. Views expressed herein are to be attributed to their authors and not to the Utah Law Review, its editors, the S.J. Quinney College of Law, or the University of Utah.

Copyright: © 2013 by the Utah Law Review Society. Printed Spring 2013 (ISSN 0042-1448). In addition to the rights to copy that are granted by law, the owner of the copyright of each article published in this issue, except as expressly otherwise noted, grants permission for copies of that article to be made and used for educational purposes by nonprofit educational institutions, provided that the author and journal are identified, that proper notice of copyright is affixed to each copy of the article, and that copies are distributed at or below cost.

Acknowledgements: The editors of the Utah Law Review wish to express their gratitude to the following: The University of Utah S.J. Quinney College of Law; the Journal Alumni Association for its support; Professor Amy J. Wildermuth, faculty adviser; Abby Thurman, editorial assistant; and the S.J. Quinney College of Law faculty and staff.

Printed in the United States of America by Joe Christensen, Inc. ISSN 0042-1448
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 cancellations of the *Utah Law Review*, please send notice and/or payment to: *Utah Law Review*, Attn: David Silva, 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.
BARADARAN, SHIMA (2013), Associate Professor of Law. B.S., 2001; J.D., 2004, Brigham Young Univ.
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), Professor of Law. Ph.D., 1992, Univ. of Michigan; J.D., 1996, Univ. of Utah.
DRYER, RANDY (2011), University of Utah Presidential Honors Professor and Visiting Professor of Law. B.S., 1972, J.D., 1976 Univ. of Utah.
FRANCIS, LESLIE P. (1982), Associate Dean for Faculty Research and Development and
Alfred C. Emery Professor of Law and Professor of Philosophy. B.A., 1967, Wellesley
College; Ph.D., 1974, Univ. of Michigan; J.D., 1981, Univ. of Utah.

GEORGE, ÉRIKA (2003), Co-Director, Center for Global Justice and Professor of Law. B.A.,

GUIORA, AMOS N. (2007), Co-Director, Center for Global Justice and Professor of Law.
B.A., 1979, Kenyon College; J.D., 1985, Case Western Reserve Univ.

Law School.

HESSICK, CARISSA (2013), Professor of Law. B.A., 1999, Columbia University; J.D., 2002,
Yale Law School.

of Utah.

HOLBROOK, JAMES R. (2003), Clinical Professor of Law. B.A., 1966, Grinnell College;

JOHNSON, CHRISTIAN (2008), Associate Den for Academic Affairs and Professor of Law.

KEITER, ROBERT B. (1993), Director, Wallace Stegner Center for Land, Resources, and the
Environment and Wallace Stegner Professor of Law. B.A., 1968, Washington Univ.;
J.D., 1972, Northwestern Univ.

KESSLER, LAURA T. (2001), Professor of Law. B.A., 1988, George Washington Univ.; J.D.,

KINGSBURY, BENEDICT (2011), Visiting Professor of Law. L.L.B, 1981, Univ. of
Canterbury, New Zealand; M. Phil., 1984, D.Phil., 1990, Oxford Univ.

KOGAN, TERRY S. (1984), Professor of Law. B.A., 1971, Columbia College; B. Phil., 1973,
Oxford Univ.; J.D., 1976, Yale Univ.

of Minnesota.

LUND, THOMAS (1978), Professor of Law. B.A., 1964, Harvard Univ.; LL.B., 1967,

MABEY, RALPH R. (2007), Professor of Law. B.A., 1968, Univ. of Utah; J.D., 1972,
Columbia Univ.

MALLAT, CHIBLI W. (2007), Presidential Professor of Middle Eastern Law and Politics.
B.A., 1982, Beirut Univ. College; Ph.D., 1990, Univ. of London; LL.M., 1983,
Georgetown Univ.

MARTINEZ, JOHN (1984), Professor of Law. B.A., 1973, Occidental College; J.D., 1976,
Columbia Univ.

MATHERSON, JR., SCOTT M. (1985), Hugh B. Brown Presidential Endowed Chair, Professor

McCORMACK, WAYNE (1978), W. Wayne Thode Professor of Law. B.A., 1966, Stanford
Univ.; J.D., 1969, Univ. of Texas.

McLAUGHLIN, NANCY ASSAF (2000), Robert S. Swenson Professor of Law. B.S., 1987,
Univ. of Massachusetts, Amherst; J.D., 1990, Univ. of Virginia.

MITCHELL, BONNIE L. (1987), Clinical Professor of Law. B.A., 1981, J.D., 1984, Univ. of
Utah.

MORRIS, JOHN K. (1979), Vice President and 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.

PETEKSON, CHRISTOPHER (2008), John J. Flynn Endowed Professor of Law. B.S., 1997,

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.


SCHWARTZ, JEFF (2013), Associate Professor of Law. B.A., 1999, Univ. of California at Los Angeles; J.D., 2002, Univ. of California at Berkeley School of Law.


STRAUBE, MICHELE (2012), Director, Wallace Stegner Center Environmental Dispute Resolution Program. B.A., 1974, Rice Univ.; J.D., 1979 Franklin Pierce Law Center.

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.

MURPHY, FELICITY (2011), Assistant Librarian and Adjunct Professor of Law. B.A., 2000, Brigham Young Univ.; J.D., 2011, Univ. of Utah.

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.

TABLE OF CONTENTS

ARTICLES

Landscape Fairness: Removing Discrimination From the Built Environment
Stephen Clowney 1

Turnover Actions and the “Floating Check” Controversy
David R. Hague 63

Reconciling Liberty and Equality in the Debate over Preimplantation Genetic Diagnoses
Jessica Knouse 107

Dissent into Confusion: The Supreme Court, Denialism, and the False “Scientific” Controversy over Shaken Baby Syndrome
Joëlle Anne Moreno & Brian Holmgren 153

Mapping, Modeling, and the Fragmentation of Environmental Law
Dave Owen 219

Defining Persecution
Scott Rempell 283

The Skeptic’s Guide to Information Sharing at Sentencing
Ryan W. Scott 345

NOTES

Timothy M. Bagshaw 409

Paving the Road Ahead: Autonomous Vehicles, Products Liability, and the Need for a New Approach
Kevin Funkhouser 437
LANDSCAPE FAIRNESS: REMOVING DISCRIMINATION FROM THE BUILT ENVIRONMENT

Stephen Clowney*

Abstract

At its core, this Article argues that the everyday landscape is one of the most overlooked instruments of modern race-making. Drawing on evidence from geography and sociology, the Article begins by demonstrating that the built environment inscribes selective and misleading versions of the past in solid, material forms. These narratives—told through street renamings, parks, monuments, and buildings—ultimately marginalize African American communities and transmit ideas about racial power across generations.

After demonstrating that the landscape remains the agar upon which racial hierarchies replicate themselves, this Article then pivots and examines current efforts to rid the built environment of discriminatory spaces. It argues that contemporary attacks on the landscape are doomed to fail. The approaches suggested by academics in law and geography either turn a blind eye to the political economy of local decisionmaking or fail to consider entrenched legal precedent.

The final section of the Article lays out a policy proposal that could spark a new focus on issues of “landscape fairness.” It argues in favor of a set of basic procedural requirements that would force jurisdictions to reconsider the discriminatory places within their borders. Procedural mandates would force government to internalize values it might otherwise ignore, allow citizen-critics to challenge dominant historical narratives, and push communities to view the past (and future) in much more diverse terms.

INTRODUCTION

During the last fifty years, African Americans have made appreciable gains in the struggle for political standing and economic status.¹ Legislative enactments and judicial pronouncements that once enshrined the edifice of white privilege
have been readily dismantled by activists and elected officials. Yet, despite many successes, the promise of full equality for African Americans remains elusive. Commentators express alarm that myths about race, forged during the era of chattel slavery, continue to distort how individuals view the world and construct their identities. In particular, observers worry that negative meanings associated with “blackness” twist African American life into something “qualitatively different from [that] enjoyed by Whites.” A host of recent empirical studies confirm the view that race remains a socially significant category of perception; black hospital patients receive less aggressive treatment than whites, banks award white loan recipients nearly twice as much investment capital as similarly situated blacks, governments sit toxic facilities in minority communities at alarming rates, and African Americans receive longer and harsher prison sentences than their white counterparts. The question we must ask is: Why? What accounts for


3 See, e.g., Johnson, supra note 1, at 381–82.


5 See Charles S. Cleeland et al., Pain and Its Treatment in Outpatients with Metastatic Cancer, 330 NEW ENG. J. MEDICINE 592, 595 (1994) (showing that African American patients are more likely to receive inadequate cancer treatment than white patients); AGENCY FOR HEALTHCARE RESEARCH & QUALITY, U.S. DEP’T HEALTH & HUMAN SERVS., AHRQ PUB. NO. 09-0002, NATIONAL HEALTHCARE DISPARITIES REPORT 176–87 (2008), available at http://www.ahrq.gov/qual/nhdr08/nhdr08.pdf (revealing that African Americans, as compared to their white counterparts, are more likely to receive potentially inappropriate prescription medication and experience poor provider-patient communication).


8 See David B. Mustard, Racial, Ethnic, and Gender Disparities in Sentencing: Evidence from the U.S. Federal Courts, 44 J.L. & ECON. 285, 285, 312 (2001) (finding, in a study of 77,236 federal offenders that controlled for numerous variables, that African Americans received considerably longer prison sentences than whites and were more likely to be sentenced to prison rather than probation).
the staunch resolve of the color line, even as the law itself has become more and more “colorblind”?9

This Article contends that the law has largely ignored one of the most significant instruments of modern race-making: the everyday landscape. The landscape, put simply, “is not innocent.”10 It inscribes selective and misleading versions of the past in solid, material forms.11 These narratives ultimately marginalize certain communities—particularly African American communities—and transmit ideas about racial power across generations.12 This Article’s aim is not only to decry this duplicity, but also to focus attention on promising regulatory mechanisms for addressing landscape’s power to reinforce racial hierarchies. Specifically, legal insights on the value of procedural fairness have the potential to redress the absences, inclusions, and marginalizations currently buried in the landscape.

At first glance, attacking the hydra of modern racism with ideas about the built environment may seem like a rather mild and ineffectual approach. Skeptics will surely ask what a landscape-based approach can hope to accomplish where powerful antidiscrimination laws have already failed.13 This Article posits that several decades of careful scholarship in the social sciences demonstrate that the landscape operates much like a tectonic fault; although its presence is seldom scrutinized, its impact remains powerful.

Geographers in particular have worked tirelessly to expose the landscape’s meaning and muscle.14 At the most basic level, they define landscape as a tangible

---

9 This Article is not the first to ask this question. For decades, critical race theorists have argued that “[c]olorblind in law is not colorblind in fact.” Stephen Reinhardt, Remarks at UCLA Law School Forum on Affirmative Action: “Where Have You Gone, Jackie Robinson?,” 43 UCLA L. Rev. 1731, 1733 (1996). See generally CRITICAL RACE THEORY: THE KEY WRITINGS THAT FORMED THE MOVEMENT (Kimberlé Crenshaw et al. eds., 1995) (exploring the role of law in shaping racial hierarchy in America). The point here is only that the work begun by critical race scholars remains incomplete.
11 See infra Part I.
12 See id.
13 Matthew J. Lindsay, How Antidiscrimination Law Learned to Live with Racial Inequality, 75 U. Cin. L. Rev. 87, 88–91 (2006) (describing how the Supreme Court’s antidiscrimination decisions have failed African Americans).
arrangement of human and natural phenomena—an alchemy of man-made structures, land surface features, and vegetation. As such, it can be read and interpreted as a text. Artifacts placed on the land, like words inscribed in a book, tell stories about memory, identity, and history. The choices a locality makes as it designs buildings, consecrates memorials, and names streets reveal its tastes, values, and hopes in a highly visible form. Examples abound. For instance, the decision to erect a war memorial at the heart of the National Mall, rather than, say, a monument to an artist or writer, imparts a particular message about our social world; as a nation, we value wartime sacrifice and believe the work of soldiers merits special remembrance. The landscape, in effect, becomes our “unwitting autobiography.”

The geography literature offers another keen insight on the power of the built environment. Landscape does more than record a locality’s ideals; it also works


In the geography literature, “landscape” has a rather tortured definition. Professor Cosgrove states, “In geographical usage landscape is an imprecise and ambiguous concept whose meaning has defied the many attempts to define it with the specificity generally expected of a science.” COSGROVE, supra note 14, at 13; see also Mitch Rose & John Wylie, Animating Landscape, 24 ENV’T & PLAN. D: SOC’Y & SPACE 475, 475 (2006) (explaining that at times “the term ‘landscape’ itself [seems] beyond rescue”).


Fred Inglis, Nation and Community: A Landscape and Its Morality, 25 SOC. REV. 489, 489 (1977) (noting that landscape provides “the most solid appearance in which a history can declare itself”).

See Pleasant Grove City v. Summum, 555 U.S. 460, 470 (2009) (holding that placement of monuments in a public park constitutes government speech); Groth, supra note 14, at 1 (“Landscape denotes the interaction of people and place: a social group and its spaces, particularly the spaces to which the group belongs and from which its members derive some part of their shared identity.”); see also SANFORD LEVINSON, WRITTEN IN STONE: PUBLIC MONUMENTS IN CHANGING SOCIETIES 9–11 (1998) (addressing the contests over Civil War era public monuments); Peirce F. Lewis, Axioms for Reading the Landscape: Some Guides to the American Scene, in THE INTERPRETATION OF ORDINARY LANDSCAPES: GEOGRAPHICAL ESSAYS, supra note 14, at 11, 12 (arguing that all human landscapes have meaning); Richard H. Schein, Acknowledging and Addressing Sites of Segregation, 19 FORUM. J. 34, 38 (2005) (stating that the landscape reveals our values in tangible form).

See Lewis, supra note 18, at 12.

The landscape is more than just a pretty view awaiting our interpretation. See Richard H. Schein, Race and Landscape in the United States, in LANDSCAPE AND RACE IN
to “re-inscribe or perpetuate the very ideas . . . that we see reflected in it.” That is, communities mold their surroundings to reflect their worldview, and the physical durability of the landscape then carries those meanings forward and influences how future generations view events that confront them.\textsuperscript{22} Returning to the World War II Memorial as an example, it both reveals the values of its creators \textit{and} instills future visitors with a normative vision about the importance of patriotism and martial heroism.\textsuperscript{23} Thus, landscape and identity are mutually constitutive of one another—landscapes are shaped by and in turn influence the societies that create them.\textsuperscript{24}

This inherent capacity of landscapes to create structures of meaning has profound implications for the process by which racial categories are created, inhabited, and reproduced. Landscapes do not arise through the work of a divine hand to celebrate the deserving and promote truthful accounts of history.\textsuperscript{25} Rather, the built environment is shaped by the tastes of government leaders and ruling elites—the only groups with sufficient resources to organize costly building projects and install permanent memorials on the land.\textsuperscript{26} It takes little effort to see

\begin{flushleft}
THE UNITED STATES 1, 5 (Richard H. Schein ed., 2006) (arguing that “cultural landscapes are not simply just \textit{there} as material evidence in the service of observations about human activity”).
\end{flushleft}


\textsuperscript{22} KENNETH E. FOOTE, SHADOWED GROUND: AMERICA’S LANDSCAPES OF VIOLENCE AND TRAGEDY 33 (1997); Owen J. Dwyer, Symbolic Accretion and Commemoration, 5 SOC. & CULTURAL GEOGRAPHY 419, 422 (2004); Audrey Kobayashi & Linda Peake, Racism Out of Place: Thoughts on Whiteness and an Antiracist Geography in the New Millennium, 90 ANNALS ASS’N AM. GEOGRAPHERS 392, 394 (2000); Schein, supra note 18, at 38 (arguing that the landscape’s “existence and presence as a tangible, visible scene works to keep . . . memories alive”).

\textsuperscript{23} \textit{See}, e.g., Jo-Ann Moriarty, Monument a Tribute to Valor and Sacrifice, REPUBLICAN (Springfield, Mass.), May 23, 2004, at A1 (explaining that the World War II Memorial commemorates the valor and heroism of American soldiers).

\textsuperscript{24} \textit{See} Dwyer & Alderman, supra note 16, at 165; Kobayashi & Peake, supra note 18, at 394.

\textsuperscript{25} \textit{See} Dwyer, supra note 22, at 426.

that those with dominion over the land may use their power to teach the public their own desired political and historical lessons.\textsuperscript{27} The landscape, as a result, tends to either exclude the heritage and memories of subaltern groups or appropriate their stories for dominant-class purposes.\textsuperscript{28}

The landscape of Natchez, Mississippi captures the principle in full. Nearly 150 years after the conclusion of the Civil War, the city of Natchez remains widely celebrated for its “astonishingly determined” effort to preserve the heritage and imagery of the Old South.\textsuperscript{29} In particular, leading citizens have worked tirelessly to refurbish a collection of antebellum plantations—vast white-pillared estates surrounded by ornate gardens and lush magnolia trees.\textsuperscript{30} Scholars have demonstrated, conclusively, that the history of these properties is soaked in the sweat of African American slaves.\textsuperscript{31} Blacks picked the cotton that generated Natchez’s remarkable wealth and blacks labored to build the mansions of the planter elite. The landscape, however, has been thoroughly scrubbed of this presence. Property owners have removed the slave quarters—the most prominent material trace of black history—from these otherwise painstakingly restored

---

\textsuperscript{27} See Kobayashi & Peake, supra note 22, at 397 (arguing that “dominant power blocs . . . use racialized representations to maintain cultural and moral legitimacy” (citation omitted)).

\textsuperscript{28} Levinson, supra note 18, at 10 (showing that “those with political power within a given society organize public space to convey . . . desired political lessons’’); James E. Young, The Texture of Memory: Holocaust Memorials and Meaning 2 (1993) (“Memory is never shaped in a vacuum; the motives of memory are never pure.”); Dwyer & Alderman, supra note 16, at 169 (“Because of white opposition, African Americans have struggled and often been unsuccessful in commemorating the [civil rights] Movement within the traditional core of urban memorial space.”); Steven Hoelscher, The White-Pillared Past: Landscapes of Memory and Race in the American South, in Landscape and Race in the United States, supra note 20, at 39, 48 (“[L]andscapes are especially powerful media for dominant groups to present their case to the world.”); Kirk Savage, History, Memory, and Monuments: An Overview of the Scholarly Literature on Commemoration, Nat’l Park Serv., http://www.nps.gov/history/history/resedu/savage.htm (last visited Nov. 19, 2012).

\textsuperscript{29} Caroline Seebohm, Enshrining the Old South, N.Y. Times, Feb 10, 1991, at A15.


\textsuperscript{31} See Linda Lange, Time Leaves Natchez Undisturbed, Milwaukee J. & Sentinel, Nov. 9, 1998, at H8 (“In the antebellum era, the agriculture-based economy depended heavily on slave labor for work in the fields. The planter class also relied on slave craftsmen and brick masons in construction of opulent estates.”). See generally D. Clayton James, Antebellum Natchez (1968); Natchez Before 1830 (Noel Polk ed., 1989).
plantation estates. And tours of the Natchez museum-homes, which draw over a hundred thousand visitors every year, exclude evidence of human bondage in favor of discussions about antebellum furniture, architecture, and the patrician lineage of local families. As one commentator has noted, the “near complete absence of a black perspective . . . is astonishing.” Moreover, when landscape features do acknowledge the existence of antebellum African Americans, they present the slave system as a benign institution, universally referring to enslaved blacks as “servants” or “cooks” or “butlers.” Examined as a whole, it becomes clear that the landscape of Natchez is a subtle form of dominance. The concrete, three-dimensional reality of the municipality creates an ideological narrative that hides the contribution of blacks and exonerates whites from any complicity with the horrors of the past.

Although the built environment remains the agar upon which racial hierarchies replicate themselves, no legal scholarship has yet analyzed the interaction between landscape and race or suggested practical tools to disrupt the ongoing process of racialization. This Article sets out to fill the void. It begins, in Part I, by closely scrutinizing the landscape in one modern American city—Lexington, Kentucky—to demonstrate how the built environment continues to marginalize African Americans and ingrain ideas about racial difference. It shows that the city’s monuments, memorials, parks, and roads all conspire to broadcast a message of white superiority, convince African Americans to accept themselves as marginal players in Lexington’s history, and physically exclude blacks from important civic spaces.

After highlighting the effect of landscape unfairness on black communities, the Article turns toward devising a policy proposal that could prevent the further subordination and invisibility of minority groups. Part II of this Article explores the academic literature in both law and geography for potential solutions to the landscape problems facing African Americans. Geographers, for their part, counsel that oppressed groups should attempt to offer “progressive/emancipatory/liberatory” counternarratives that expose the symbolic meanings of the land. Legal academics, confronted with any question of racial discrimination, generally advocate erecting statutory proscriptions that either ban or criminalize the noxious activity in question. Although these reform traditions both pack some theoretical

32 Hoelscher, supra note 28, at 60 (stating that the mention of slavery in Natchez has been confined to a few specific sites and that tours of the mansions continue to venerate the antebellum period); Leonard Pitts, Jr., Bed, Breakfast . . . And a Denial of Black History, MIAMI HERALD, Mar. 9, 2003, at 1L (discussing the conversion of slave quarters on Natchez plantations into fancy bed and breakfasts).
33 Hoelscher, supra note 28, at 54.
34 Id.
35 Id.
36 Id.
punch, neither offers much hope of altering the built environment. The problem is this: current proposals to intervene in the landscape either rely on legal dead ends or disregard the political economy of local government decisionmaking.

Part III—the final section of this Article—searches for a new, more practical approach to transforming racialized landscapes. Rather than advocate for a sweeping ban on any landscape practice tinted with a hint of racism, this section pushes for something less grandiose and more achievable. It argues that the law should impose a set of basic procedural requirements on government bodies to ensure they thoughtfully consider the effects of any actions that significantly alter the landscape. The basic idea is that forcing government actors to contemplate landscape values they might otherwise neglect will lead inevitably to better informed, more rational, and racially sensitive decisionmaking. At the same time, stricter procedural requirements will open the landscape to greater public participation, thus enhancing democratic accountability and transparency.

I. THE LANDSCAPE OF LEXINGTON: AN EXAMPLE OF RACE-MAKING

The principal argument of this Article is that the American landscape has played a crucial, and largely unexamined, role in the creation and maintenance of racial hierarchies. To make this theoretical claim more concrete, the pages that follow present empirical evidence, on the microlevel of municipal development, of the tangled relationship between race and the built environment. This Article offers a contribution to the larger contemporary debate about the persistence of the color line by documenting how the landscape (1) communicates messages of white supremacy, (2) misrepresents the history of black communities, and (3) physically excludes African Americans from important civic spaces.

The subject of this research is Lexington, Kentucky. The city offers several advantages for a case study of this type. First, the power struggle between Lexington’s historically black population and its wealthier white community roughly parallels the political dynamic in the broader United States. Critics will point out that the focus on state action threatens to limit the scope of my proposed solution. Admittedly spaces constructed by private individuals, without any government action or intervention, would fall outside the reach of the suggested procedural remedy. Note, however, that government action can be defined broadly. California and New York have adopted procedure-based environmental protection statutes that apply to government decisions on land use permits, including projects financed or approved by government. See CAL. CODE REGS. tit. xiv, §§ 15000–15387 (2011); N.Y. COMP. CODES R. & REGS. tit. 6 §§ 617.1–617.20 (2011). A similarly broad scheme focused on landscape issues would capture most, but admittedly not all, significant building projects in a municipality.

See Eugene J. McCann, Race, Protest, and Public Space: Contextualizing Lefebvre in the U.S. City, 31 ANTIPODE 163, 163–64 (1999) (demonstrating that Lexington has traditionally been a racially divided city); Schein, supra note 10, at 210 (“Lexington is struggling to join the twenty-first century, especially in terms of its race relations or, more
city and nation, blacks constitute a significant—but not decisive—voting block. This demographic reality should lead to some useful conclusions about African Americans’ ability to resist a white majority’s influence over the physical setting. Second, Lexington sits at the crossroads of four distinct political/historical regions—the North, South, Midwest, and Appalachian Mid-Atlantic. It is therefore unlikely that current spatial practices can be attributed to particular sectional characteristics, cultural traditions, or ethnic practices. Finally, the citizens of central Kentucky have long cared about, and fought over, landscape issues. From the original dispossession of the Indians to the more recent crusade to preserve local farmland, ideas about land, property, and the environment have been stitched with heavy thread into the social fabric of the city. Though few places directly match Lexington’s obsessive concern with its physical milieu, scholars have noted, “The study of extreme instances often helps to illuminate the essentials of a situation.” In short, examining Lexington is fruitful not because it is perfectly representative of how Americans think about the built environment, but because it crisply reveals how landscapes work and why they matter.

Within Lexington, this research has concentrated on two sites: the old courthouse square and a recently constructed public park. These places represent a cross section of the types of spaces that average citizens move through in their daily routines. It is these kinds of everyday spaces that best reveal how the built environment uncritically smuggles ideas about race and racial hierarchy into the ether of everyday existence.

accurately, in terms of the structural imperatives of racism that are embedded in so many of the city’s . . . institutions.”).

40 The major difference between Lexington and the larger country is that the Bluegrass region has a much lower percentage of Latinos. Critics could charge that by focusing on the black-white binary, this Article is examining race through an inappropriately narrow lens. This criticism has some theoretical force, and my ideas about property and landscape have been shaped by living in places where the black-white social relations structure the spatial trajectories of everyday life.


42 See Schein, supra note 37, at 815.

A. Courthouse Square

1. A Landscape of White Supremacy

Perhaps the most obviously racialized space in Lexington is the Courthouse Square, a large public plaza in the heart of downtown. It takes little imaginative work to see that the square tells a narrow history of elite, white Lexington at the expense of African American memory and identity. Foremost, monuments honoring Confederate war heroes dominate the commons. A memorial to cavalry officer John Hunt Morgan, for example, commands the western edge of the plaza. Morgan, a hemp merchant from Lexington, became known as the “Thunderbolt of the Confederacy” for his freewheeling guerilla raids on Union positions in Kentucky and Tennessee. Supplementing the Morgan sculpture, a nineteen-foot statue of John Breckinridge, the Confederate Secretary of War and erstwhile Vice President of the United States, stands guard over the main entrance to the commons.

As Victor Turner has noted, “Objects speak,” and the statues that venerate the Confederacy convey an obviously racialized message. Take, for example, the

---

44 For more on the history and form of courthouse squares, see Edward T. Price, The Central Courthouse Square in the American County Seat, 58 GEOGRAPHICAL REV. 29 (1968).
45 There is a rich scholarly literature on Confederate statuary and memory of the Civil War. See, e.g., Kirk Savage, Standing Soldiers, Kneeling Slaves: Race, War, and Monument in Nineteenth-Century America 129–208 (1997); David W. Blight, “For Something Beyond the Battlefield”: Frederick Douglass and the Struggle for the Memory of the Civil War, 75 J. AM. HIST. 1156, 1178 (1989); Steven Hoelscher, Making Place, Making Race: Performances of Whiteness in the Jim Crow South, 93 ANNALS ASS’N AM. GEOGRAPHERS 657, 663 (2003).
47 The Breckinridge Statue: Something About Its Appearance and Where It Will Stand, LEXINGTON MORNING TRANSCRIPT, July 10, 1887, at 2. The statue is eight feet tall and sits on an eleven-foot-high pedestal.
50 There has never been any doubt in the black community about the meaning of statues dedicated to the Confederacy. Take, for example, Frederick Douglass’s rage at monuments to Robert E. Lee. Douglass “considered them an insult to his people and the Union.” Blight, supra note 45, at 1169; see also Jonathan Leib, The Witting Autobiography of Richmond, Virginia: Arthur Ashe, the Civil War, and Monument Avenue’s Racialized Landscape, in Landscape and Race in the United States, supra note 20, at 187, 191
reaction of author Mamie Fields to a sculpture of John C. Calhoun, the tireless defender of slavery and states’ rights, in Charleston, South Carolina. “Blacks took that statue personally,” Fields recalled, “Calhoun [seemed to] look[] you in the face and tell[] you, ‘Nigger, you may not be a slave, but I am back to see you stay in your place.’” Confederate memorials in Kentucky have faced similar criticisms. Lexingtonians—mostly black Lexingtonians—argue that the veneration of John Hunt Morgan conveys a message of “marked racism,” while a nearby monument to Jefferson Davis has faced protests because it “rekindles memories of hostile Southern resistance to African-Americans’ quest for freedom.”

That these statues stick like a bone in the throat of the black community should not surprise. The monuments’ use of traditional commemorative forms coupled with the enormous outlays of capital entailed in such installations infuses the landscape with a luster of civic authority and cultural legitimacy. Moreover, the artistic composition of the statues cements the notion that Lexington intentionally moored ideas about racial stratification in the physical geography of the plaza. The design of the Morgan memorial, in particular, undermines any notion that Courthouse Square presents an impartial narrative above racial bias.

(discussing the resistance of the black citizens of Richmond to the Lee monument in 1887 and 1890); Alexander Tsesis, The Problem of Confederate Symbols: A Thirteenth Amendment Approach, 75 TEMP. L. REV. 539, 554 (2002) (“Confederate symbols express a nostalgic longing for the Old South.”). But see GEORGE SCHEDLER, RACIST Symbols & REPARATIONS 43 (1998) (arguing that “the more pervasive [the Confederate flag] becomes in popular culture the more it loses the racist overtones”).

51 Karen Fields, What One Cannot Remember Mistakenly, in HISTORY AND MEMORY IN AFRICAN AMERICAN CULTURE 150, 156 (Geneviève Fabre & Robert O’Meally eds., 1994).


54 Owen J. Dwyer, Location, Politics, and the Production of Civil Rights Memorial Landscapes, 23 URB. GEOGRAPHY 31, 32 (2002).

55 Critics have charged that such reactions are overwrought, and that the Morgan and Breckinridge statues merely reflect Lexington’s sympathies and historical experiences during the Civil War. See, e.g., Donald G. Shelton, Column Attacking John Hunt Morgan Insults Southerners, LEXINGTON HERALD-LEADER, Dec. 3, 2001, at A10. On close inspection, this argument burns off like a thin morning fog. In truth, the vast majority of citizens and soldiers in Kentucky, which did not secede, sided with the cause of the Union. WALL, supra note 41, at 63. Union generals such as Thomas Leonidas Crittenden, Thomas John Wood, Kenner Garrard, Barton S. Alexander, and John Buford all represented Kentucky during the war. EZRA J. WARNER, GENERALS IN BLUE: LIVES OF THE UNION COMMANDERS 100, 569, 167, 581, 52 (1964). Even Lexington, more outwardly Confederate than the rest of the state, had stronger ties to the North—300 Confederate and 1,300 Union veterans lie buried in Lexington Cemetery. Telephone Interview with Daniel Scaife, President & Gen. Manager of Lexington Cemetery (Oct. 10, 2012). Yet, despite this
The twenty-foot monument depicts a dignified incarnation of Morgan sitting calmly astride a well-muscled stallion. Yet historians agree that in life, he rode a small mare and forged a reputation for daring and roguish bravado. This discrepancy—the gap between history and memory—did not occur spontaneously through either accident or artistic whimsy. As Professor Kirk Savage has demonstrated, the portrayal of the Confederate general as a disciplined officer, riding with easy control of his animal was a common feature in the sculptural iconography of the time and carried a well-understood meaning. The equestrian in equipoise, Savage writes, was an “overripe metaphor” that worked beautifully “as a model of leadership for a white supremacist society trying to legitimate its own authority. [It] is at once a retrospective image of the benevolent master . . . and a prospective image of a postwar white government claiming to know what is best for its own black population.” Architecture critic Catherine Bishir agrees, claiming that Morgan’s sculptural avatar is the embodiment of the “confident leader who calmly controls not only the horse but the racially stratified society in which he operates.” In short, the Confederate sculptures nestled in the heart of Lexington definitively encode the landscape with a racialized presence—a perpetual reminder that public space belongs to the white majority, both now and into the future.

Those who defend the project of Confederate remembrance will, of course, take issue with any claim that the city government has endorsed antebellum attitudes about racial difference. The potential lines of attack are easy to spot. First, some will argue that Lexington has no moral responsibility for memorial projects that were organized by private groups. This argument is easily dismissed. Local newspaper accounts reveal that city officials worked closely with the sculptor of the Morgan statue and helped fashion “every detail” of the final design. The same is true of the Breckinridge monument. Moreover, the prominent placement of the memorials on sacred municipal land—near Main Street and in the shade of history, the landscape of Courthouse Square masks the strength of unionism and promotes a romantic ideal of Confederate honor and gentility.


57 See SAVAGE, supra note 45, at 133–35.

58 Id. at 135.

59 Bishir, supra note 46, at 80.

60 Id. at 78.

61 See Louise Taylor, Historical Statue Earns Interest: A Colorful Century at Fayette Courthouse, LEXINGTON HERALD-LEADER, Dec. 13, 2002, at A14 (showing that the Breckinridge statue was funded by a $10,000 gift from the government).
the Courthouse—further suggests an officially sanctioned longing for the Confederacy, and helps claim civic space for the Lost Cause tradition.62

Skeptics will also contend that contemporary Lexington should not be tarred with decisions made by city fathers during the early twentieth century. Although this claim carries some theoretical force, on close inspection it, too, crumples under the weight of evidence. In 2010, Lexington redeveloped Courthouse Square with a goal of making the area more attractive and pedestrian friendly.63 To implement this plan, the city built a new glass and steel pavilion, embedded LED lighting in the walkways, and reworked the walking paths.64 The city council also made a conscious decision to retain the Confederate memorials and chose to move the statue of John Breckinridge from the interior of the square to a more visible location along the city’s primary traffic artery.65 The result is a thoroughly modern landscape that continues to inscribe a “white pillared” version of the past.66

Lexington’s Courthouse Square is not unique. Across the South, similar memorial spaces sit at the heart of both large cities and small towns. That these scenes remain so commonplace does not render them harmless. The prominent position of Confederate sculptures along Main Street and the deliberately misleading interpretations of history conspire to ingrain ideas about racial hierarchy, cement conclusions about racial difference, and send messages that African Americans are not full members of the polity.

2. Black History, Concealed

In addition to encouraging fidelity to a social order founded on racial terror, Courthouse Square also conceals the narratives of Lexington’s African American community.67 Before the war, the northwest corner of the commons, known regionally as Cheapside, was the site of the largest slave auction in the state, and one of the most important slave-trading districts in the South.68 Locals claim that the Cheapside moniker derives from the area’s reputation for selling older and cheaper slaves—the more valuable chattels were taken directly from the

---

62 See Dwyer, supra note 22, at 420 (noting that courthouse lawns commonly “attract a plethora of memorials, all of them seeking to legitimate the cause they represent via close association with the seat of government”).
64 Id.
65 Id.
66 Hoelscher, supra note 28, at 39.
67 See Foote, supra note 22, at 322–32 (noting that the landscape conceals much African American history and that more could be done to bring that history to light).
68 See Kellogg, supra note 41, at 25 (discussing Lexington’s role as an important slave trading market); McCann, supra note 39, at 170 (referring to Lexington as a “major regional slave market”).
countryside to auctions in the Deep South. Although this history, staked in blood, permeates the area, no heroic statuary commemorates the local African American community. The landscape fails to honor the sacrifice of the twenty-four thousand black Kentuckians who fought in the Civil War; it fails to recognize the role of slave labor in fostering regional prosperity; and it fails to enshrine a single African American politician, thinker, or artist in the pantheon of the city’s heroes. Moreover, no public memorial records that after emancipation, the plaza was the site of violent mob attacks against blacks who transgressed socially enforced codes of behavior. Like a shadow on a cloudy day, black history has silently disappeared from the environs of Lexington’s Courthouse Square.

Admittedly, there is one landscape feature that acknowledges what Toni Morrison might call an “Africanist Presence.” A historical marker erected in 2003 concedes that Courthouse Square hosted Lexington’s slave auction block. But this single exception to the dominant-class rendering of the landscape proves the larger rule—Lexington still seems to favor the mythology of the Confederacy over the gritty history of its black community. Note, for example, that during the

---

69 See Schein, supra note 18, at 37.
70 Owen J. Dwyer, Interpreting the Civil Rights Movement: Place, Memory, and Conflict, 52 Prof. Geographer 660, 663 (2000) (arguing that American memorial landscapes have routinely marginalized the presence of African Americans).
71 ANNE E. MARSHALL, CREATING A CONFEDERATE KENTUCKY: THE LOST CAUSE AND CIVIL WAR MEMORY IN A BORDER STATE 20 (2010). The invisibility of blacks in Civil War statuary is not unusual. African Americans also lack a place in the post–Civil War era fictional literature. As Professor Daniel Aaron writes, “Although black slavery was the root cause of the War, it is hard to name one War novel containing a fully realized Negro character.” DANIEL AARON, THE UNWRITTEN WAR: AMERICAN WRITERS AND THE CIVIL WAR 333 (1973); see also DAVID W. BLIGHT, BEYOND THE BATTLEFIELD: RACE, MEMORY, & THE AMERICAN CIVIL WAR 112 (2002) (“Slavery, the war’s deepest cause, and black freedom, the war’s most fundamental result, remain the most conspicuous missing elements in the American literature inspired by the Civil War.”); Savage, supra note 28 (discussing how African American soldiers were removed from mainstream public memory).
72 See McCann, supra note 39, at 170 (describing the lack of memorials to the subjugation of black slaves, or to the bravery of those involved with the underground railroad).
2010 redesign of the plaza—the same renovation that lifted John Breckinridge into
the spotlight—the city placed the slave auction marker in a peripheral corner
behind the Courthouse building. The location, far from the main pedestrian
thruway, is so marginal that the local paper initially insisted, incorrectly, that the
city removed the plaque altogether.

Less visible but just as telling, the design and construction of the slave auction
marker received no direct support from local or state government, a stark contrast
to the tens of thousands of dollars in public funds spent on projects dedicated to
Confederate memory. Only the bootstrapping organization and fundraising of a
local black fraternity ensured that the landscape of Courthouse Square holds any
reminder of the solemn turns in Lexington’s history. And even with the marker in
place, taking in the full sweep of the built environment leads a casual observer
straightforwardly to the conclusion that the two-hundred-year presence of African
Americans in the state’s history barely deserves mention.

3. Physical Exclusion

So far, this subsection has attempted to demonstrate that the landscape injects
ideas about race and power into the scenes of everyday life. Before moving on, this
Article explores more concretely how the “tradition of white . . . ownership” of
public space imposes actual, material costs on black communities. Foremost, this
Article argues that racialized landscapes physically exclude African Americans
from important public arenas and deny them the opportunity to fully participate in
civic and economic life.

In Lexington, for instance, interviews with local stakeholders establish that
some members of the black community feel compelled to entirely avoid
Courthouse Square out of respect for the memory of the slave auctions. Civic
leaders are uniformly aware of the area’s history and described the space as “not
inviting,” “not a center for all,” and “not a comfortable place.” Others,

77 Id.
78 See Bishir, supra note 46, at 77 (relaying that half of the funds for the John Hunt
Morgan statue came from government sources).
79 SHERRILYN A. IFILL, ON THE COURTHOUSE LAWN: CONFRONTING THE LEGACY OF
80 My research assistant, a Lexington native, originally identified a handful of
stakeholders. He then used the discussants’ informal social networks to arrange further
interviews with different segments of the community. I make no claim that the interviewees
represent a cross section of the black community in Lexington.
81 Telephone Interview with Patrice Muhammad, Radio Host, Key Conversations
Radio (July 22, 2011).
82 Telephone Interview with David Cozart, Adm’r of Dev., Urban League of Fayette
Cnty. (July 22, 2011).
especially in the older generation, refuse to patronize nearby businesses as a protest against the racial hierarchy that remains embedded within the landscape. At the very least, many blacks do not feel part of the plaza’s community and fail to make free and efficient use of the commons. In this regard, Lexington’s black community is not unique. A burst of scholarly research has established that African Americans in other parts of the country feel similarly “constrained,” “shutout,” “traumatized,” and “frozen” when confronted with mainstream civic landscapes and employ strategies to avoid “the indignities of racialized space.”

The subtle exclusion of African Americans exacts a heavy toll. In Lexington, blacks end up ostracized from Courthouse Square—the social and political hub of the Bluegrass region. Local African Americans may fail to exploit the farmers’ market or the civic festivals that draw thousands of revelers into downtown each week. And, more pertinently, they may avoid the cascade of public lectures, demonstrations, memorial celebrations, and political rallies that the square

---

83 Interview with Sonia Price, Dir. of the African Am. Studies & Research Program, Univ. of Ky. (June 22, 2011).
85 See McCann, supra note 39, at 173; see also OWEN J. DWYER & DEREK H. ALDERMAN, CIVIL RIGHTS MEMORIALS AND THE GEOGRAPHY OF MEMORY 26 (2008) (noting that Lexington’s African American community is well aware of the racialized history of Courthouse Square).
86 See IFILL, supra note 79, at 22 (“In the contemporary physical landscape, continued white control of the terms in which public space in multiracial communities is defined and developed is a vestige of white supremacist notions of racial place.”); McCann, supra note 39, at 168 (“While it is widely acknowledged that public spaces often constrain the actions of women, cultural studies literature has begun to show how bourgeois public space has similar effects on people of color . . . .” (citation omitted)); Susan Ruddick, Constructing Difference in Public Spaces: Race, Class, and Gender as Interlocking Systems, 17 URB. GEOGRAPHER 132, 136 (1996) (“[S]cholars in black and cultural studies . . . have noted the ways in which roles of visible minorities have been scripted in and through public space.”).
87 Robert R. Weyeneth, The Architecture of Racial Segregation: The Challenges of Preserving the Problematical Past, PUB. HISTORIAN, Fall 2005, at 11, 33 (“One way that African Americans could minimize some of the indignities of racialized space was simply to avoid these places.”); Nanda R. Shrestha & Wilbur I. Smith, Geographical Imageries and Race Matters, in GEOGRAPHICAL IDENTITIES OF ETHNIC AMERICA: RACE, SPACE, AND PLACE 279, 280 (Kate A. Berry & Martha L. Henderson eds., 2002) (explaining that blacks feel “traumatized” and “frozen” in certain geographies).
88 See McCann, supra note 39, at 170 (arguing that the area around Courthouse Square is the “heart of the contemporary downtown financial area”); Cheapside Park One of the Most Historic Places in Lexington, LEXINGTON LEADER, June 30, 1938, § 3, at 47 (“A detailed history of Cheapside would be a history of Lexington itself.”).
regularly hosts. Such outcomes inevitably corrode the entire city’s public life. As Berman argues, “The glory of modern public space is that it can pull together all . . . different sorts of people . . . . It can both compel and empower all these people to see each other, not through a glass darkly but face to face.” Hierarchical landscapes, like Courthouse Square, set fire to this important crucible of participatory democracy.

The physical exclusion that results from racialized space also wreaks havoc on the economic prospects of black Americans. Professor Sherrilyn Ifill’s ethnographic study of Maryland, for example, uncovered that African Americans purposively spurned real estate opportunities along the Eastern Shore because the landscape in those communities was stamped with memories of racial violence and intimidation. Ifill sums up that “decisions about where to look for a new home, where to stop late at night for gas or for a bite to eat, and where to send one’s children to school” are all informed by whether a community’s public spaces broadcast a message of racial hierarchy or fosters rigid power relations.

In Lexington, the message transmitted by Courthouse Square is clear. The Confederate statues that abut Main Street proclaim, quite loudly, the subordinate position of local African Americans. These memorials, and the negative emotions they evoke, push blacks out of the public square, denying them access to important municipal rituals and increasing feelings of invisibility within the community. Thus, Lexington demonstrates that stakes in debates over the built environment remain high. Contests over the landscape are about more than pretty sculptures; they fundamentally affect the way that racial attitudes reproduce themselves across time.

B. Thoroughbred Park

The message of racial difference carved into Courthouse Square spirals out into parts of the city removed from the urban core. Take, for example, Thoroughbred Park, a three-acre triangle of land that serves as the southern gateway to downtown. The park, at first glance, is a marvelous space. A lush rolling hill, which echoes the surrounding Bluegrass countryside, towers over a

---

90 See McCann, supra note 39, at 170–71 (discussing physical constraints on African American’s participation in the life of the city). For more on the square’s importance to Lexington, see Schein, supra note 84, at 390–91.

91 Marshall Berman, Take It to the Streets: Conflicts and Community in Public Spaces, 33 DISSERT 476, 482 (1986).

92 See Ifill, supra note 79, at 20–21; see also Shrestha & Smith, supra note 87, at 280 (stating it is not uncommon among some racial minorities to avoid areas marked by events rooted in racism).

93 Ifill, supra note 79, at 21.

94 Kevin Nance, Architects Reveal Plans for New Park: Themes Will Be Bluegrass, Horses, LEXINGTON HERALD-LEADER, Oct. 28, 1989, at A1 (stating that goal of park was to “bring the quality of the Kentucky countryside directly into downtown Lexington”).
superbly crafted stone wall, a fountain, and a group of bronze racehorses that careen toward an imaginary finish line. Although the grassy knob and winding footpaths of Thoroughbred Park may initially seem far removed from the Confederate stronghold of the courthouse, all is not as it appears. On closer inspection, this setting, too, crafts myths about white achievement, elides African American history, and excludes blacks from central Lexington.

I. White Presence, Black Absence

Any critique of Thoroughbred Park must begin by tracing the history and meaning of the landscape’s many statues and memorial signs. In the early 1990s, the city government of Lexington and its allied private foundations agreed that the grounds of the park would record the history of horseracing in the Bluegrass area. The site would symbolize, in tangible form, Lexington’s international reputation as an equine mecca. To achieve that end, the grounds of the park “commemorate outstanding people and horses in the thoroughbred industry.” Statues of famous jockeys and celebrated horses stand sentinel on the park’s grass, and the pathways brim with plaques dedicated to the champions of the horse industry, from George Washington (a renowned horse breeder) and Bing Crosby (a technical innovator) to the contemporary financiers who continue to push the business forward.

Although aesthetically satisfying, the narrative of the horse industry, carved in bronze and presented as gospel, is profoundly incomplete—and profoundly racialized. Thoroughbred Park presents the public with a misleading version of the region’s history that ultimately champions the deeds of elite whites and writes out the labor and lives of African American workers. The park, for example, offers no mention of the (mostly black and Hispanic) trainers, stable hands, and barn foremen who have provided the unseen manpower and moxie essential to the industry’s survival. Perhaps more egregious, Thoroughbred Park also fails to commemorate any of the black jockeys who dominated horseracing in the late nineteenth century, even though the park’s memorial apparatus repeatedly references that era of the sport. The space makes no room for even prominent historical figures like Oliver Lewis, the first jockey to win the Kentucky Derby, or Isaac Murphy—the winningest jockey of all time, a well-known Lexingtonian, and a member of the Horseracing Hall of Fame. That a rider of Murphy’s stature fails

general background on the Bluegrass landscape, see Karl Raitz & Dorn Van Dommelen, Creating the Landscape Symbol Vocabulary for a Regional Image: The Case of the Kentucky Bluegrass, 9 LANDSCAPE J. 109 (1990).

95 Nance, supra note 94.
96 Id.
97 Randy Romero, Pat Day, Bill Shoemaker, Jerry Bailey, Don Brumfield, Chris McCarron, and Craig Perret are all represented in the statuary.
98 See Schein, supra note 37, at 821.
99 For more on Murphy, see WALL, supra note 41, at 110.
to appear in the narrative suggests that the omission of black figures stems not from benign neglect, but an active attempt to conceal the role of African Americans in horseracing’s origin story.

It is worth repeating that the ability of dominant (white) groups to inscribe their version of history into the built environment has enduring consequences. The power of landscape—and its insidiousness—lies in its ability to make selective accounts of the past seem normal. Lexington’s Thoroughbred Park, emplaced like a geological outcropping in the landscape, asserts that the contributions of the black community are insignificant. Visitors, by and large, must accept this account as part of the natural order of things because they cannot see the contested process of selecting, forgetting, and reassembling narratives that underlies the site. Moreover, the park’s physical permanence allows it to project forward the myth that blacks have not supported the horse industry—the one endeavor that largely defines the region’s identity. The landscape, in short, is not a passive vessel, but an active vehicle that produces and shapes the future.

2. Physical Exclusion

Beyond the misleading historic narrative, the easiest proof that Thoroughbred Park has marginalized the black community can be found in the story of the park’s unique, hilly topography. In 1989, when Lexington first announced the plans to convert a string of auto-related businesses into Thoroughbred Park, the site demarcated a boundary between two contrasting socioeconomic areas. On one side, the parcel fronted a rich, white section of the city. Across the property, and clearly visible to motorists entering downtown, loomed the black neighborhood of

100 See Dwyer & Alderman, supra note 16, at 168 (“The subtle power of memorials is that they often communicate seemingly authentic and unproblematic representations of history.”); Hoelscher, supra note 45, at 661 (arguing that landscape’s duplicity stems from its ability to “project a sense of timelessness and coherency”).

101 As some scholars argue, “Publics have been trained to view monuments and historical markers, . . . [such as] plaques, as carrying an aura of unity, universality, and timelessness.” Lisa Maya Knauer & Daniel J. Walkowitz, Introduction to MEMORY AND THE IMPACT OF POLITICAL TRANSFORMATION IN PUBLIC SPACE 1, 5 (Daniel J. Walkowitz & Lisa Maya Knauer eds., 2004). It is also worth examining Catherine Bishir’s take on Lexington’s John Hunt Morgan Statue. “I didn’t know much about the famed Confederate cavalry officer,” she writes, “but from his statue I knew he was important and figured he was probably good.” Bishir, supra note 46, at 62.

102 See SAVAGE, supra note 45, at 4 (noting that memorials “remain a fixed point, stabilizing both the physical and the cognitive landscape”).

103 See Bishir, supra note 46, at 81 (arguing that dominant groups shaped the landscape to claim the past and “set the terms by which they meant to shape the future”).


105 See Schein, supra note 10, at 214.

106 Id.
...the northeast—a place represented in the local news media as poor, violent, and overrun by disaffected teenagers.

Local business interests argued, sometimes forcefully, that the view was not conducive to Lexington’s redevelopment efforts and, as a result, the large rolling hillside of Thoroughbred Park was built. The mound was “literally built for the park to effectively hide the [African American residential district] from view.”

For anyone approaching downtown from the interstate highway, Lexington’s black neighborhood—and black bodies—remain firmly out of sight, tucked neatly behind the grassy partition. An editorial in the local paper succinctly captured the dynamic; “Though aesthetically pleasing, the park is historically false. The ‘rolling bluegrass hills’ are manmade. . . . [T]he park not only ignores [the black neighborhood], but also screens it from view. It is a whitewash. It is telling that almost every African American . . . instantly recognizes this racial effect.”

Thus, the construction of the artificial hillside in Thoroughbred Park stands easily as an emblem for the mischief that built environments have unleashed in black communities. The park normalizes and cements borders between socially constructed racial zones. It physically corrals, through division and enclosure, the scope of the local African American neighborhood. And, finally, it proclaims an easily intelligible message about who exactly belongs on which side of Main

---

107 Id.
108 See McCann, supra note 39, at 171.
109 See, e.g., Dottie Bean, Humana Pledges $500,000 for Thoroughbred Park, LEXINGTON HERALD-LEADER, Feb. 20, 1990, at B1 (describing the view as “singularly unattractive”); Robert Kaiser, Work to Begin Next Week on Thoroughbred Park, LEXINGTON HERALD-LEADER, Apr. 11, 1991, at A1 (describing site as a “scene of urban decay”); Nance, supra note 104 (quoting mayor saying that park would “act as an entrance to the city that will tell people they’re coming into something nice”); Shelia M. Poole, Foundation Unveils Its Plan for East Downtown Park, LEXINGTON HERALD-LEADER, Oct. 27, 1988, at B1 (complaining about the view); see also Bettye Lee Mastin, Panel Sought Referendum on Urban Renewal Plan in ’64, LEXINGTON HERALD-LEADER, Sept. 13, 1984, at D5 (reporting that the city had attempted to demolish the neighborhood in question during the height of the urban renewal period).
110 Schein, supra note 10, at 214; see also McCann, supra note 39, at 179 (making the case that contemporary public spaces are “designed to keep the frequency of uncomfortable encounters to a minimum and to maintain a rigid power relation between Whites and people of color”); Weyeneth, supra note 87, at 13 (arguing that built environment has long been designed to “manage[] contact between whites and blacks”).
112 Spatial control has long been a strategy used by dominant classes to fix racial difference and “‘demonstrat[e]’ the superiority of one race over the other.” Hoelscher, supra note 45, at 671.
113 See McCann, supra note 39, at 171; Schein, supra note 10, at 217.
Street; blacks, the park seems to say, are unwelcome and “out of place” in the white-dominated world of downtown.\textsuperscript{114}

In sum, Thoroughbred Park is not an obviously racialized space. No statues or plaques proclaim the innate inferiority of African Americans or celebrate white supremacist heroes. And yet the park’s space still retains a strong racial meaning. In particular, the decisions to hide a black neighborhood and erase black jockeys from the history of the local thoroughbred industry convey a message about the relative worth and standing of African Americans in the local community.

\textbf{C. Wrapping Up}

At this point, it should come as no surprise that revolutionaries, as one of their first tasks, often seek the physical destruction of spaces that symbolize the regimes they have replaced.\textsuperscript{115} In 1985, after the death of Albania’s brutal dictator, Enver Hoxha, protesters smashed statues of their former leader and vandalized government property.\textsuperscript{116} Similar scenes have engulfed Afghanistan, Russia, and Iraq after recent regime changes. The reason is clear: landscapes influence the societies that produce them. Like the chemicals in a dark room, the built environment develops ideas about community values, notions of personal identity, and feelings of belonging.\textsuperscript{117}

In the United States, the landscape has played a particularly active role in shaping black lives. Again and again, the built environment has either physically excluded African Americans or presented a weaponized version of history that symbolically annihilates their accomplishments. While this Article has focused primarily on sites of race-making in central Kentucky, it is a somber fact that thousands of other locations remain complicit in reproducing outdated notions about the black-white binary. In some places, the machinery of race is easily visible. The French Quarter of New Orleans, for instance, offers an obelisk dedicated to the White League, a paramilitary organization that led resistance to African American enfranchisement.\textsuperscript{118} In a similar vein, the government of North Carolina has built twenty-four memorials on the capital grounds in Raleigh, including half-a-dozen dedicated to the Confederacy—but not one African

\textsuperscript{114} McCann, supra note 39, at 171 (arguing that the state produces space to maintain hierarchies, and that groups who are not included are continually made to feel “out of place”); see also Kobayashi & Peake, supra note 22, at 393 (explaining that “whiteness” is a “profoundly geographic phenomenon” based on the control of space).

\textsuperscript{115} See Levinson, supra note 18, at 12.

\textsuperscript{116} See id. at 15.


\textsuperscript{118} See Dwyer, supra note 22, at 421. To the credit of New Orleans, the Liberty Monument has been moved to a slightly less visible location than it occupied for many years.
American stands among the honored. And across the South, schools still bear the name of the Klan’s founder. At other sites, race lingers inconspicuously. Scholars claim the design of contemporary public spaces, historic preservation districts, national parks, and everyday road signs all subtly ingrain ideas about racial power.

There is, however, some good news; unlike social problems rooted in the currents of culture and history, the troubles that stem from landscape unfairness can be fixed swiftly by dedicated and creative local governments. We have the capability, in ways both large and small, to expose the absences in the landscape and alter its physical framework and meanings.

II. CURRENT EFFORTS TO REFORM THE LANDSCAPE

Drawing on extensive theoretical work about landscape, the previous Part sought to identify how the built environment marginalizes African American communities and reproduces ideas about racial power. This Article now pivots and tries to find solutions to the knotty problems posed by landscape unfairness. How, exactly, can communities remove the taint of racial hierarchy from the surroundings they inhabit? Is it possible for localities to devise a principled framework to guide decisions about racialized spaces, or, on the contrary, is each place so context dependent that overarching theories provide little traction?

Rather than reinventing the wheel, it may be useful to begin by asking whether scholars already claim any solutions to the landscape problems facing

119 Austin, too, drips with reminders of the slaveholding past. For example, the Memorial to Confederate Veterans reads, “Died for states rights guaranteed under the Constitution[.] The people of the South, animated by the spirit of 1776, to preserve their rights, withdrew from the federal compact in 1861. The North resorted to coercion. The South, against overwhelming numbers and resources, fought until exhausted.” Levinson, supra note 18, at 55. Monument Avenue in Richmond is another example of a prominent racialized space. See Leib, supra note 50, at 190.

120 There are at least two public high schools named after Nathan Bedford Forrest (a Confederate general and founder of the Klan): one in Jacksonville, Florida and another in Chapel Hill, Tennessee. See Name Dropping: Should School Names That Honor Supporters of Slavery Be Changed?, CURRENT EVENTS, Dec. 8, 2008, at 7 (describing controversy over school in Florida named after Nathan Bedford Forrest). Forrest’s name also adorns other buildings and spaces throughout the South. For example, in Memphis, there is a park named after Forrest, while the Reserve Officers’ Training Corps (ROTC) building at Middle Tennessee State University bears Forrest’s name. The National Center for Education Statistics also shows school districts routinely name schools after Robert E. Lee, Stonewall Jackson, and Jefferson Davis.

121 See McCann, supra note 39, at 179.

122 See Schein, supra note 20, at 10.


124 See Schein, supra note 20, at 10–11; Dwyer & Alderman, supra note 16, at 166.
African Americans. As mentioned earlier, the legal literature is of little help; scholars of law rarely examine the spatial trajectories of everyday life and they have largely ignored the problems that the built environment imposes on black communities. Geographers, in contrast, have long offered suggestions to mitigate the harms caused by landscape unfairness. Although details vary, the thrust of the geography literature is that minority groups should alter the landscape by building their own counterdiscourses. According to this scholarly tradition, African Americans should craft their own monuments, mark their own histories, and contribute to the visible form of the American scene.

After kicking the tires and checking under the hood, this Article argues that geographers’ attempt to inject impartiality into the landscape has major shortcomings, which virtually ensures that the real world condition of African Americans will not improve. More specifically, the following Section contends that the suggestion that blacks impose counternarratives on the landscape fails to grapple with the political realities of decisionmaking at the local level and the stunted economic power of African American communities.

A. The Potential Power of Counternarratives: Kelly Ingram Park

Geographers routinely argue that subaltern groups should resolve landscape imbalances by crafting their own counterdiscourses. Professors Kobayashi and Peake, for instance, write that minority communities must take it upon themselves to “[d]isrupt established attributes of place and the confining boundaries that have literally allowed whiteness to take place” and “engage in coalition building in order to resist the creation of racialized hierarchies.” The potential of this approach is clear. Staking a tangible presence in places that have traditionally reflected the hopes of white men would signal that African Americans, too, have been an integral part of the national story. A more inclusive landscape could also reduce minority groups’ sense of alienation, spark new discourses about the legacy of racial oppression, and bring the achievements of outsiders into mainstream historical narratives.

To put flesh on the bones of this theory, the geography literature has repeatedly pointed to the success of Kelly Ingram Park in Birmingham, Alabama as empirical proof that campaigns to emplace counternarratives can dismantle the
implicit bias in the American landscape. The park, once ensconced in the national imagination as the site of Bull Connor’s attack on young civil rights protesters, was remade in the early 1990s to commemorate the struggle to desegregate Birmingham’s public facilities and commercial enterprises. Local activists convinced the city to rededicate the space as a “Place of Revolution and Reconciliation” and commissioned a series of steel sculptures to capture iconic moments of the civil rights campaign in Alabama. Visitors to the park must now wind their way through large-scale installations that depict protestors standing firm against snarling dogs, mechanized water cannons, and mass arrests.

The overall effect of the change has been nothing less than transformative. Most obviously, the landscape of Kelly Ingram Park validates, in stone and steel, the historical contributions of African Americans to the country’s freedom struggle and publicly scrutinizes the city’s white power establishment. As one visitor remarked on seeing the space for the first time, “I was witnessing the transformation of Birmingham’s official history.” Priscilla Cooper, the institute’s education program consultant, said, “[Whites from the suburbs will finally] have to face the truth. They must tell the students what really happened.” The statues’ visceral presence along the park’s pathways has also encouraged scenes of cathartic emotional release. A place once touched by tragedy, shame, and violence now sparks an outpouring of community pride and symbolizes the new standing of the city’s black population.

---

129 See, e.g., DWYER & ALDERMAN, supra note 85, at 18–19; Thomas H. Cox, From Centerpiece to Center Stage: Kelly Ingram Park, Segregation, and Civil Rights in Birmingham, Alabama, 18 S. HISTORIAN 5, 5 (1997); Dwyer, supra note 54, at 38; Schein, supra note 20, at 8; Dell Upton, Commemorating the Civil Rights Movement, 40 DESIGN BOOK REV. 22, 22–23 (1999).
130 See Dwyer, supra note 70, at 662.
131 See Dwyer, supra note 54, at 36–38.
132 Id. at 38 (quoting Frederick Kaimann, Kelly Ingram Park Has Look to Go with Institute Design, BIRMINGHAM NEWS, Nov. 15, 1992, at 36P).
133 See DWYER & ALDERMAN, supra note 85, at 19.
134 Inscriptions on the sculptures proclaim key slogans from the 1963 campaign: “Segregation is a sin,” “Warriors of a Just Cause,” and “We Ain’t Afraid of Your Jail.” Another sculpture, of two children behind jail bars, encourages viewers to both stand with the children on the inside of the cell and then outside as oppressor.
135 See Owen J. Dwyer, Interpreting the Civil Rights Movement: Contradiction, Confirmation, and the Cultural Landscape, in THE CIVIL RIGHTS MOVEMENT IN AMERICAN MEMORY 5, 8 (Renée C. Romano & Leigh Raiford eds., 2006).
137 Id.
138 For more on the power of memorials to provide a focus for emotional release, see FOOTE, supra note 22, at 179.
139 See Whitehouse, supra note 136.
B. Political and Economic Realities

The happenings in Birmingham demonstrate that the geographers’ claim about the importance of counternarratives carries water. A large-scale remaking of the American landscape could, in theory, bestow real and lasting benefits on black populations. The geography literature, however, consistently fails to recognize the practical difficulties of replicating the success of Kelly Ingram Park at a larger spatial scale or in different localities. In particular, geographers seem reluctant to acknowledge that any significant change to the built environment requires vast stores of political capital and economic resources—assets that local black communities traditionally lack.

Even in Birmingham, landscape change occurred only as a result of an unusual confluence of three demographic and political events—a model difficult to reproduce at the national level. First, the legacy of the civil rights struggle, still fresh in the minds of many locals, left behind a core of African American activists with intense motivation to record their experiences in the built environment. Second, as a result of declining white population in the 1970s, Birmingham became one of the few American cities where blacks constitute a majority of registered voters and control the levers of political power. Third, in the early 1990s, the city’s white business community—a potential adversary to any major change in the downtown landscape—was seeking fresh ways to address the city’s reputation as a hotbed of racism. As a result, the local chamber of commerce avidly supported the redesign of the park and helped procure $4 million in private funding. The story of Birmingham then, demonstrates the potential power of large-scale landscape change, but also reveals that emplacing counternarratives in the built environment requires a far more complex mix of sustained activism, political will, and financial support than geographers have indicated. In few other locations will the perceived interests of all stakeholders line up quite so neatly to allow African Americans to infuse the land with their memories and accounts of history.

The empirical research on the politics of renaming civic infrastructure in honor of Martin Luther King, Jr. makes this point explicitly. Social scientists have discovered that, across the country, attempts by African American communities to rename landscape features after King often face “insurmountable” opposition from white majorities who believe that the legacy of the civil rights leader “is too
narrowly racial” to be commemorated in integrated public spaces.\textsuperscript{144} As Professor Derek Alderman has written, “Many whites do not personally identify with King . . . . The renaming of major thoroughfares is frequently disrupted by the protests of the street’s business owners and operators, who cite the financial costs of changing their address and the social stigma . . . of being associated with the black community.”\textsuperscript{145}

Indeed, in cities both large and small—from Phoenix to Danville, and from San Diego to Statesboro—whites have either successfully resisted attempts to name streets after the civil rights leader or substituted smaller roads for the major thoroughfares originally proposed by black activists.\textsuperscript{146} Efforts to honor King have faced particular difficulty in localities without large black communities. On average, African Americans constitute nearly 40% of the population in locations with a renamed street.\textsuperscript{147} In over a third of the successful locations, African Americans make up more than 50% of the population.\textsuperscript{148} This statistical portrait demonstrates that any call for African Americans to mold their own landscapes overlooks the political reality on the ground: black communities often lack the power to impose their vision of the future on white majorities.

African Americans have also encountered difficulties in other quests to remake the built environment with their own counternarratives. For example, large memorial projects dedicated to the civil rights movement, such as statues and museums, have often met resistance from whites, who view these endeavors as displacing white history and memory.\textsuperscript{149} As one scholar summarized, “Main Street,

\textsuperscript{144} Dwyer, supra note 54, at 48.

\textsuperscript{145} Derek Alderman, Street Names and the Scaling of Memory: The Politics of Commemorating Martin Luther King, Jr. Within the African American Community, 35 AREA 163, 164 (2003). Comedian Chris Rock famously lampooned the stigma of streets named in honor of Martin Luther King. “Martin Luther King stood for non-violence. Now what’s Martin Luther King? A street. And I don’t . . . [care] where you live in America, if you’re on Martin Luther King Boulevard, there’s some violence going down.” Bring the Pain (HBO television broadcast June 1, 1996).

\textsuperscript{146} See Dwyer & Alderman, supra note 16, at 169 (“Many of America’s roadways named for Martin Luther King, Jr., are side streets or portions of roads located within poor, black areas of cities . . . .”).

\textsuperscript{147} DWYER & ALDERMAN, supra note 85, at 51.

\textsuperscript{148} Id.

\textsuperscript{149} See, e.g., IFILL, supra note 79, at 20 (“Many whites in Talbot County viewed the plans to create a memorial to Frederick Douglass as a threat to their conception of the courthouse lawn as a ‘sacred’ space set aside to honor white heroes.”); Dwyer, supra note 70, at 667 (arguing that civil rights museums fail to examine issues of contemporary racism as a result of pressure from corporate and government donors); Dwyer & Alderman, supra note 16, at 169 (“Because of white opposition, African Americans have struggled and often been unsuccessful in commemorating the [civil rights] Movement within the traditional core of the urban memorial space.”); Lorri Helfand, Not the Time for King Honor, ST. PETERSBURG TIMES, Sept. 13, 2007, at 1 (discussing white opposition to a plan to honor Martin Luther King with a memorial plaza in Largo, Florida).
the county courthouse, and city hall remain devoted to remembering white-dominated historical narratives.”150 Similarly, the campaign to recognize the role of African Americans in the Civil War has encountered fierce opposition. The design of the Freedman’s Memorial—a monument dedicated to capturing the spirit of emancipation—was hijacked by white political interests and twisted to portray African Americans not as active agents in the struggle for their own liberation, but as passive automatons who received the gift of freedom from white leaders.151 Moreover, recent efforts by African Americans to expand displays about the black experience at Civil War battlefields “have encountered resistance both inside and outside” the National Parks Service.152 At Manassas Battlefield, for example, archaeological evidence of African American life has been deliberately removed from the park, meaning that “issues of slavery, oppression, racism, Reconstruction, and Jim Crow are prevented from becoming part of the national story.”153 The cascade of examples reveals, again and again, that when African Americans have taken it upon themselves to challenge dominant historical narratives and paint the landscape with more diverse memories, they have found themselves politically outgunned and outmaneuvered.

Further complicating the quest for counternarratives, many landscape projects—especially monuments and memorials—are financed through private donations.154 Thoroughbred Park in Lexington, for instance, required over $8 million in gifts from individual donors before construction commenced.155 This dynamic creates difficulties for memorial activists concerned about honoring African American heroes and creating spaces that record the diverse experiences of minority groups. The problem, in a nutshell, is that compared to whites, African Americans control far less wealth and, therefore, face larger obstacles navigating the bramble of inflexible costs necessary to remake the built environment.156

150 Dwyer & Alderman, supra note 16, at 169.
151 Monuments devoted to capturing the spirit of emancipation have been particularly ripe targets for intervention by white elites. The most famous example is Ball’s Freedman’s Monument. Other examples include Ward’s Henry Ward Beecher Monument and Scofield’s Emancipation Panel. For a thorough treatment of the African American presence of Civil War sculpture, see Boime, supra note 41, at 171–79 (arguing that images of emancipation, including the famous Freedman’s Monument “were designed to emphasize the dependence of the emancipated slaves upon their benefactors”). See generally Savage, supra note 45.
152 Savage, supra note 28.
153 Paul A. Shackel, Memory in Black and White: Race, Commemoration, and the Post-Bellum Landscape 180 (2003).
Indeed, in 2007, the financial wealth \footnote{Financial wealth is a household’s net worth minus the equity it has accrued in owner-occupied housing. This statistic seems especially relevant in the landscape context, as it is unlikely that families would mortgage their homes to make a contribution to a monument campaign or memorial project.} of the median white family hovered around $43,000, while the equivalent African American family’s financial wealth was closer to $500.\footnote{G. William Domhoff, Who Rules America? Challenges to Corporate and Class Dominance (2006). For updated statistics, see G. William Domhoff, Information for Students (and Others) Reading Who Rules America? Challenges to Corporate and Class Dominance, http://www2.ucsc.edu/whorulesamerica/book.html (last visited June 8, 2013).} Given these vast disparities in access to capital, it is little wonder that when black communities have won permission to alter the landscape, they often experience difficulty finding the resources necessary to effect change. The North Carolina Freedom Monument, a large-scale project intended to “recognize and honor the African American experience” on the statehouse grounds, remains unbuilt after nearly fifteen years of fundraising.\footnote{The North Carolina Freedom Monument Project was initiated in 2001 and is still working diligently to raise needed funds. See Janet Kagan, The North Carolina Freedom Monument Project: A Freedom Grove For All, Indy Week (Apr. 19, 2006), http://www.indyweek.com/indyweek/the-north-carolina-freedom-monument-project/Content?oid=1197171. For more information, see the project’s website at http://www.ncfmproject.org/.} Likewise, the campaign to construct a memorial to African American Civil War veterans in Hagerstown, Maryland has gained very little traction, despite support from the city government.\footnote{See Hagerstown-Washington Cnty. Convention & Visitors Bureau, Fighting for Freedom: Remembering African-American Veterans, Kickstarter, http://www.kickstarter.com/projects/1318805847/fighting-for-freedom-remembering-african-american (last visited June 23, 2013).} Even the well-publicized effort to erect a Martin Luther King, Jr., Memorial on the National Mall languished for years under the weight of weak fundraising efforts.\footnote{See George E. Curry, Money and the MLK Memorial, St. Louis Am. (Aug. 25, 2011 12:15 AM), http://www.stlamerican.com/business/local_business/article_8d0eddfe-ce9a-11e0-b1a0-001cc4c002e0.html (discussing details on fundraising and noting the few black donors).} African Americans were either unable or unwilling to contribute major gifts to the project, and the sponsors were forced to rely on large corporate donations to fuel construction efforts.\footnote{Id.; see also Stephon Johnson, Black Philanthropy Lacking in MLK Memorial Fundraising, N.Y. Amsterdam News, Jan. 13, 2011, at 1, 34 (“[W]hen it comes to donations by millionaire or billionaire African-Americans they seem to be nowhere among the major donors.”).}

At the end of the day, the process of assembling the landscape remains a ruthlessly political endeavor. As the first Part of this Article described, rival interest groups have continuously struggled to embed visual representations of
their histories, their ideals, and their worldviews into the built environment. For 250 years, elite white men have been the decisive winners of this contest, controlling the narratives told through the land and, in the process, erasing some stories from public display. To counteract the problem of landscape unfairness, geographers stress that African Americans and other subaltern peoples should pry open the whiteness of these sites by emplacing their own marks on the landscape. Although this solution to the problem of landscape unfairness has a strong theoretical underpinning, relying on market mechanisms to balance the landscape seems doomed to fail. In addition to the danger of creating a semiotic jumble of monuments and memorial plazas, most African American communities simply lack the political power and financial resources to effect sweeping changes to the built environment. Thus, academics and activists concerned about the representation of African Americans in the landscape should not continue to grasp at the scholarly vapor put forth by geographers, but rather accept the hard demographic realities and seek new levers to help black communities remake their surroundings.

III. LAW & LANDSCAPE: CAN WE BAN RACIALIZED SPACES?

The law remains the most systematic method of providing “high-level, generalized, rules” to manage complex social conflicts. It is tempting, therefore, to inquire if the law, as currently practiced and interpreted, can provide an antidote to the problem of racially charged landscapes. What legal strategies could litigants deploy? And, more importantly, would their arguments find any traction in a court of law?

The most direct, and potentially powerful, approach would call upon the law to ban landscapes that perpetuate racist ideologies or alienate minority populations. Such a tactic has obvious promise. Law, unlike geography, offers minority groups a robust stash of weapons to protect themselves against the tyranny of majority decisionmaking. During the last fifty years, litigants have proven especially adept at using federal civil rights guarantees to chisel away at the grievous racial harms inflicted on black communities. Courtroom battles to achieve greater school desegregation, integrated housing, and access to public accommodations have

---

163 Levinson, supra note 18, at 75.
164 See, e.g., Vikram David Amar, California Constitutional Conundrums—State Constitutional Quirks Exposed by the Same-Sex Marriage Experience, 40 Rutgers L.J. 741, 763 (2009) (“[A] constitution protects minorities from unreasonable and oppressive acts imposed on them through government by the majority—that is, from the potential ‘tyranny of the majority.’”).
all been won under the aegis of federal law.\textsuperscript{167} Given this history, it seems plausible that patches of the constitutional text could also help African Americans roll back the worst abuses enshrined in the built environment.

Despite the inherent advantages, pursuing a substantive ban on racialized landscapes would be a mistake. On close inspection, the approach suffers from both practical and conceptual infirmities. As an initial matter, the current shape of American law, as announced by the Supreme Court, would make it difficult for activists to raze racialized spaces through the legal process. During the last two decades, judges have considerably narrowed the ability of litigants to sue government actors under the First and Fourteenth Amendments—the civil rights protections most relevant to the landscape context.\textsuperscript{168} A more theoretical difficulty also looms. Even if courts do eventually show an increased willingness to smoke out race-based injustice, it would be unwise to vest judges with the power to interpret culturally contested landscapes. Courts, as a general matter, lack the institutional competence to tease out the meaning and implications of the built environment. Moreover, it is questionable whether those individuals sitting behind the bar—almost invariably members of the cultural elite—are best positioned to ascertain the meaning of spaces they may never see or use.

\textit{A. The Practical Problems with Pursuing a Ban on Racialized Spaces}

The animating thesis of this section is that activists interested in upending racialized landscapes should rethink any strategy that depends on judicial action. These concerns are chiefly instrumental. Although the text of the Constitution provides an obvious staging ground to contest government actions that shape the meaning of the built environment,\textsuperscript{169} in recent years courts have erected “substantial roadblocks to the prosecution of civil rights cases.”\textsuperscript{170} These roadblocks will make it difficult, if not impossible, to effectuate a ban on racialized spaces.\textsuperscript{171} In fact, the pathway to legal victory appears so constricted that activists’ time, energy, and money seem better spent on political organizing, mobilization, and direct action at the local level.


\textsuperscript{168} See City of Richmond v. Croson, 488 U.S. 469, 552–53 (1989) (Marshall, J., dissenting) (“[A] majority of this Court signals that it regards racial discrimination as largely a phenomenon of the past, and that government bodies need no longer preoccupy themselves with rectifying racial justice.”).

\textsuperscript{169} A stumbling block in the fight against discriminatory spaces is that no government has fashioned a statute or code section that explicitly bans landscapes that ingrain ideas about racial difference. Thus, any attempt to attack the built environment must find cover under some broader piece of civil rights legislation.


\textsuperscript{171} Id.
Some observers will surely argue that the Equal Protection Clause of the Fourteenth Amendment could disrupt at least some of the public landscapes that ingrain archaic ideas about race. The text of the amendment, which provides that no state may “deny to any person within its jurisdiction the equal protection of the laws,” has consistently been interpreted to bar invidious discrimination based upon race. Indeed, a series of landmark cases during the last fifty years has demonstrated the power of equal protection jurisprudence to remove the stigma of racial difference and undo the structures of white supremacy. The decisions in Brown v. Board of Education, Loving v. Virginia, and NAACP v. Button all applied the lever of the Fourteenth Amendment to upend unjust power relationship between white majorities and black citizens.

Despite its illustrious history in the federal courts, the Equal Protection Clause offers little hope for those seeking to midwife an era of racially balanced landscapes. In the Supreme Court’s recent jurisprudence, the justices have shackled the expansive potential of the constitutional text to mend social wrongs. Current case law demands that a litigant challenging a facially neutral...
government action under the Equal Protection Clause satisfy a two-pronged test to
maintain a claim. The litigant must first establish that the government’s
undertaking produces disproportionate effects along racial lines, and then
demonstrate that racial discrimination was a substantial or motivating factor behind
the act. 179

These requirements, according to scholars, have sidelined many attempts to
remedy contemporary racial disparities. 180 In particular, academics worry that
divining the existence of discriminatory intent among government decisionmakers
has become “virtually impossible,” 181 “almost impossible,” 182 “nearly
impossible,” 183 “difficult if not impossible,” 184 and “prohibitively difficult.” 185
Consider the Lexington native who argues that the Confederate memorials
emplaced on the courthouse are not only a personal affront, but also a serious
violation of the rights conferred by federal law. Could the litigant demonstrate that
racial discrimination was a motivating factor in the decisions to erect the
Breckenridge and Morgan statues? Although observers agree that white
southerners erected Confederate monuments as part of larger campaigns to rewrite
the meaning of the Civil War and stake a message of “absolute racial difference”
upon the land, 186 in Lexington at least, it appears that city fathers avoided making

---

180 See, e.g., David A. Strauss, Discriminatory Intent and the Taming of Brown, 56 U.
CHI. L. REV. 935, 951–54 (1989) (discussing how the need to show discriminatory intent
fails to address many race-based harms).
181 John J. Delaney, Addressing the Workforce Housing Crisis in Maryland and
Throughout the Nation: Do Land Use Regulations that Preclude Reasonable Housing
Opportunity Based upon Income Violate the Individual Liberties Protected by State
182 Amanda K. Franzen, The Time Is Now for Environmental Justice: Congress Must
Take Action by Codifying Executive Order 12898, 17 PENN ST. ENVTL. L. REV. 379, 403
(2009).
183 Kenya Hart, Defending Against a “Death by English”: English-Only, Spanish-
Only, and a Gringa’s Suggestions for Community Support of Language Rights, 14
BERKELEY L. RAZA L.J. 177, 190 n.91 (2003).
184 Scott R. Hechinger, Juvenile Life Without Parole: An Antidote to Congress’s One-
185 D. Taylor Tipton, Note, The Dunkin’ Donuts Gap: Rethinking the Exclusionary
Rule as a Remedy in Constitutional Criminal Procedure, 47 AM. CRIM. L. REV. 1341, 1354
(2010).
186 Hoelscher, supra note 45, at 663, 671; see also SAVAGE, supra note 45, at 130–39
(examining the racial politics behind the memorialization of Robert E. Lee); Leib, supra
note 50, at 190 (“[M]onuments . . . were . . . constructed at the height of the Lost Cause era,
when Southern whites were (re)writing the history of the Confederate cause . . . .”).
any overtly race-based statements to the local media.\textsuperscript{187} Thus, even though the landscape has a disparate impact on black citizens, which prevents all community members from making full use of vital public spaces,\textsuperscript{188} the Equal Protection Clause offers no relief.\textsuperscript{189}

Moreover, even if African Americans could present direct evidence that racial discrimination motivated the design of the landscape, southern judges have proved stubbornly resistant to removing racialized symbols from state-owned property. Take, for example, the struggle over the Confederate flag. There can be little argument that racial animus motivated the decision of southern states to revive the Confederate battle flag during the 1950s and '60s.\textsuperscript{190} The events preceding the flag’s reemergence in South Carolina,\textsuperscript{191} Alabama,\textsuperscript{192} and Georgia\textsuperscript{193} all confirm that the symbol was raised, not as a historical artifact, but as a badge of defiance to court-ordered integration in the wake of \textit{Brown}.\textsuperscript{194} Despite the factual record and historical background, courts have routinely rejected claims that the Equal

\textsuperscript{187} I have looked through newspaper archives at the University of Kentucky and have yet to find any direct statements that the purpose of the statues was to oppress African Americans.

\textsuperscript{188} \textit{See supra} Part I.

\textsuperscript{189} \textit{But see} Hunter v. Underwood, 471 U.S. 222, 232 (1985) (striking down a 1901 provision of the Alabama Constitution that denied the franchise to people convicted of crimes of moral turpitude after finding that the law was clearly motivated by racial animus).

\textsuperscript{190} \textit{See Tesis}, \textit{supra} note 50, at 601–07.


\textsuperscript{194} James MacKay, a member of the 1956 Georgia legislature that incorporated the Confederate battle flag into the state flag, testified that the Confederate battle flag was adopted as a symbol of resistance to integration. \textit{Coleman}, 912 F. Supp. at 528. Dan Carter, a professor of southern history at Emory University has testified that “[b]y the mid-1950s, the Confederate battle flag had become the single most important symbol of white supremacy and defiant opposition to federally mandated laws on non-discrimination.” \textit{Coleman}, 117 F.3d at 529 n.5; \textit{see also} J. Michael Martinez, \textit{Traditionalist Perspectives on Confederate Symbols, in Confederate Symbols in the Contemporary South} 243, 254–55 (J. Michael Martinez et al. eds., 2000) (discussing the origins of the Ku Klux Klan’s use of the Confederate battle flag). Before World War II, display of the flag “was considered disrespectful unless it was displayed at a reunion or for another important purpose designed to honor memories of veterans who served in the Confederate forces during the war.” \textit{Id.} After the war, the Klan began displaying the flag “as part of a conscious effort to identify its message of intolerance and fear of other races with Confederate symbols.” \textit{Id.} at 255.
Protection Clause mandates the removal of the battle flag from state property.\(^\text{195}\) \textit{NAACP v. Hunt}\(^\text{196}\) illustrates the principle in full. In \textit{Hunt}, the Eleventh Circuit determined that the source of the plaintiffs’ grievance stemmed not from the shame and anger caused by viewing the Confederate flag, but an inability to control their “own emotions.”\(^\text{197}\) Other courts have reached similar conclusions,\(^\text{198}\) demonstrating that the Fourteenth Amendment offers little shelter to those seeking landscape fairness. Even when plaintiffs have summoned evidence against the discriminatory symbols that mar the built environment, courts have hesitated to impose the guarantee of equal protection.\(^\text{199}\)

2. \textit{The Weakness of a First Amendment Approach}

While the Equal Protection Clause may be the first stop in the search for a method to ban racialized spaces, it is not the only plot of constitutional ground worth exploring. Activists may also be tempted to argue that the First

\(^{195}\) Coleman, 117 F.3d at 528 (challenging Georgia state flag); NAACP v. Hunt, 891 F.2d 1555, 1559 (11th Cir. 1990) (challenging “the flying of the flag atop the Alabama capital dome”); Holmes v. Wallace, 407 F. Supp. 493, 494 (M.D. Ala. 1976), aff’d, 540 F.2d 1083 (unpublished table decision) (challenging Alabama’s display of the Confederate flag over the state capitol).

\(^{196}\) 891 F.2d 1555.

\(^{197}\) Id. at 1565.

\(^{198}\) See, e.g., Coleman, 117 F.3d at 530–31 (holding that the incorporation of the Confederate flag within the Georgia state flag does not violate citizens’ equal protection rights); Holmes, 407 F. Supp. at 497 (holding that flying the Confederate flag on the dome of the state capital does not constitute the “deprivation of any rights, privileges, or immunities secured by the Constitution”).

\(^{199}\) Not only does the equal protection doctrine fail to unmake racially divisive geographies, an aggressive reading of the Supreme Court’s Fourteenth Amendment jurisprudence may actually \textit{prevent} localities from enacting other programs designed to integrate existing spaces. The Supreme Court’s affirmative action cases, in particular, raise a thicket of problems for the notion of landscape fairness. In the affirmative action cases, the Court has been increasingly skeptical of arguments that a particular program is justified because it would combat discriminatory acts that occurred in the past. The position is summed up by Justice Scalia’s statement in \textit{Adarand} that “there can be no such thing as either a creditor or a debtor race. . . . In the eyes of government, we are just one race here. It is American.” Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 239 (1995) (Scalia, J., concurring in part). If the Court’s view holds—that all racial classifications are equally suspect—then attempts to disrupt existing racialized territory by introducing landscape forms that celebrate nonwhite groups may face a stiff challenge. See id. at 224 (majority opinion) (discussing equal protection scrutiny of racial classifications). There is, however, a key difference between affirmative action issues and the problem of landscape fairness. The wrongs imposed by the landscape are not \textit{just} historical. They continue to emit messages of racial inferiority today. In fact, landscapes bolster the argument that defenders of affirmative action often try to make: “The past is never dead. It’s not even past.” \textit{William Faulkner, Requiem for a Nun} 92 (1950).
Amendment’s guarantee of free speech has the power to invalidate racialized monuments and rebalance the landscape. The argument is easy enough to follow. The guiding light of free speech case law is that the government cannot discriminate against a speaker based upon distaste for his viewpoint or message. Yet when governments construct expressive landscape features like monuments and memorial parks, they invariably make content-based decisions about who has the right to speak in public places.

Examples abound. State officials in Georgia recently agreed to hang a portrait of statesman Sonny Perdue outside of the office of the governor, while simultaneously rejecting a petition to position a portrait of Martin Luther King, Jr. in the same hallway. In choosing to honor Perdue and exclude King, the government validated one group’s speech while displacing the expression of a rival faction—the very kind of viewpoint discrimination traditionally barred by the First Amendment. Similarly, Georgia’s decision to erect statues of Confederate icons like John Gordon and Benjamin Hill on the statehouse lawn, rather than using the space to glorify hometown heroes like W.E.B. DuBois, has carved certain

200 See, e.g., Owen M. Fiss, State Activism and State Censorship, 100 Yale L.J. 2087, 2100 (1991) (“The state must act as a high-minded parliamentarian, making certain that all viewpoints are fully and fairly heard.”).
201 See, e.g., Turner Broad. Sys., Inc. v. FCC, 512 U.S. 622, 641 (1994) (arguing the heart of the First Amendment is the prevention of viewpoint discrimination); Police Dep’t v. Mosley, 408 U.S. 92, 95 (1972) (“[T]he government has no power to restrict expression because of its message, its ideas, its subject matter, or its content.”).
203 Joseph Blocher, Viewpoint Neutrality and Government Speech, 52 B.C. L. REV. 695, 726 (2011) (“[A] principled method of distinguishing between government speech and viewpoint discrimination seems impossible to articulate.”); see also LEVINSON, supra note 18, at 83 (“In the age of the activist state, governmental speech is a pervasive method of regulation . . . .”)
206 During the war, Gordon was one of Lee’s most trusted advisors. After the war, Gordon became governor of the state. There is evidence that he also served as the head the state’s Ku Klux Klan. See RALPH LOWELL ECKERT, JOHN BROWN GORDON: SOLDIER, SOUTHERNER, AMERICAN 145 (1989). Hill served in the Confederate Senate and, after the war, became an outspoken critic of Reconstruction. ROBERT PRESTON BROOKS, HISTORY OF GEORGIA 309 n.1 (1913).
(white) voices into the state’s most sacred civic space at the expense of other (blacker) worldviews.

Despite the obvious tension between viewpoint neutrality and landscape design, attacking racialized spaces with the cudgel of the First Amendment will almost certainly fail to transform the built environment. Why? A kink in recent Supreme Court jurisprudence makes it clear that the government’s own expression “is exempt from First Amendment scrutiny.”207 That is, when the government speaks, it does not need to maintain even a scintilla of viewpoint neutrality—it can freely choose sides and condemn those who hold beliefs it finds repugnant.208 A recent case demonstrates that this principle applies squarely to landscape issues. In Pleasant Grove City v. Summum, a religious community known as the Summum sought to place a monument in a public park on the outskirts of Pleasant Grove, Utah.209 Although the park already hosted an assortment of privately donated sculptures,210 Pleasant Grove refused the monument, arguing that the piece was inconsistent with the city’s purported message of showcasing local history.211 The Summum then brought suit under the First Amendment, insisting that the city had engaged in viewpoint discrimination by excluding the group from a public forum intended for the display of permanent monuments.212

In a unanimous opinion, the Court ruled against the Summum. The decision noted, as an initial matter, that monuments displayed on public property certainly constitute a form of government speech: “A monument, by definition, is a structure that is designed as a means of expression. When a government entity arranges for

207 Johanns v. Livestock Mktg. Ass’n, 544 U.S. 550, 553 (2005). It should be noted, however, that the government speech doctrine is not limitless. See Pleasant Grove City v. Summum, 555 U.S. 460, 482 (2009) (Stevens, J., concurring) (“For even if the Free Speech Clause neither restricts nor protects government speech, government speakers are bound by the Constitution’s other proscriptions, including those supplied by the Establishment and Equal Protection Clauses.”); Daniel W. Park, Government Speech and the Public Forum: A Clash Between Democratic and Egalitarian Values, 45 GONZ. L. REV. 113, 145 (2010) (“The only clear limit on government speech is the Establishment Clause of the U.S. Constitution.”).

208 See Blocher, supra note 203, at 708–09 (arguing that if Democrats controlled the Department of Education, they could, in theory, “decide that all federally funded schoolchildren should be taught that Republicans are vile racists”); Developments in the Law: State Action and the Public/Private Distinction, 123 HARV. L. REV. 1248, 1293 (2010) (“Expansion of the government speech doctrine . . . threatens to erode constitutional protections by allowing the government to discriminate based on viewpoint in an increasingly wide range of circumstances.”). See generally MARK G. YUDOF, WHEN GOVERNMENT SPEAKS: POLITICS, LAW, AND GOVERNMENT EXPRESSION IN AMERICA (1983).

209 Summum, 555 U.S. at 465.

210 Id. (including a display of the Ten Commandments, a historic granary, a wishing well, and a September 11 monument).

211 Id.

212 Id. at 466.
the construction of a monument, it does so because it wishes to convey some thought or instill some feeling in those who see the structure.”\textsuperscript{213} The opinion by Justice Alito then affirmed that “government speech is not restricted by the Free Speech Clause.”\textsuperscript{214} The Court justified the loophole in its commitment to viewpoint neutrality by explaining that the state must have the ability to privilege certain facts and opinions for government to function properly. “If every citizen were to have a right to insist that no one paid by public funds express a view with which he disagreed,” Alito writes, “debate over issues of great concern to the public would be limited to those in the private sector, and the process of government as we know it radically transformed.”\textsuperscript{215}

With government speech doctrine standing as good law, it seems near impossible that activists can tame racialized landscapes with the Free Speech Clause. Just as the Summum could not force Pleasant Grove to accept their monument, black Georgians cannot compel the state government in Atlanta to hang a painting of King or install a statue of LeRoy Johnson—the state legislator who desegregated the Georgia General Assembly and pushed to revise the literacy test that had kept many African Americans from voting.\textsuperscript{216} In the landscape arena, First Amendment jurisprudence is not a sword to constrain government, but a shield to protect it.

To review, activists intent on unmaking racialized landscapes should think twice before pursuing any strategy that rests on the whims of judicial decisionmaking. Although the text of the Constitution seems to provide lawyers with room to attack discriminatory places, courts have turned an increasingly deaf ear toward civil rights cases based on racial bias. Current legal rules, like the intent standard in Fourteenth Amendment cases and the government speech doctrine in First Amendment disputes, so narrow the opportunity for courtroom victories that memorial activists should shift their efforts toward other, more direct, forms of engagement.

\textbf{B. The Conceptual Problems with Pursuing a Ban on Racialized Spaces}

The doctrinal narrowness of modern civil rights law is not the only difficulty confronting those who seek to outlaw racialized landscapes through judicial action. The approach also suffers from inherent conceptual weaknesses. The problem is twofold. First, there are reasons to be skeptical that courts have the institutional competence to make judgments about the meaning and implications of the built environment. Second, it is not always true that all monuments amount to government speech. A close reading reveals that the scope of the Summum decision remains open for debate.\textsuperscript{217}

\begin{footnotes}{\footnotesize
\item[213] Id. at 470. It is worth noting that the scope of the \textit{Summum} decision remains open for debate. A close reading reveals that it may not \textit{always} be true that \textit{all} monuments amount to government speech.
\item[214] Id. at 469.
\item[215] Id. at 468 (citing Keller v. State Bar of Cal., 496 U.S. 1, 12–13 (1990)).
\item[216] For more on Johnson, see \textsc{Timothy J. Crimmins} \& \textsc{Anne H. Farrissee}, \textsc{Democracy Restored: A History of the Georgia State Capitol} 139–40 (2007).
\end{footnotes}
environment. Even well-trained readers of the land sometimes struggle to tease out a single, coherent meaning from contested landscapes. Second, aggressive judicial intervention in disputes over cultural ideals often triggers harmful waves of public backlash that undermine the goals of the original ruling. As Professor NeJaime has argued, court decisions that compel change “disrupt the natural evolution of social change, thereby provoking backlash” that puts “substantial obstacles in a social movement’s path.” In light of these dynamics, even if courts demonstrated a willingness to take a more aggressive stance against racialized spaces, litigants should hesitate before asking them to intervene.

1. A Question of Institutional Competence

As an initial matter, those hoping to vest judges with the power to make legally privileged readings of the landscape should pause and consider whether courts have the competencies required to complete the job in an analytic and objective fashion. I worry that they do not. Unlike analyzing a car accident, which draws on a common experience of the everyday person, or applying well-defined legal principles to questions about the admissibility of evidence, assessing the meaning of the built environment would compel judges to engage in sociological and cultural analysis that far exceeds their institutional expertise. Few members of the judicial class have any training in the disciplines most concerned with unraveling the meaning of the built environment—geography, architecture, and city planning.

---

217 See Michael J. Klarman, Brown and Lawrence (and Goodridge), 104 Mich. L. Rev. 431, 473 (2005) (arguing that court rulings can cause political backlash when “they alter the order in which social change would otherwise have occurred”); see also William N. Eskridge, Jr., Channeling: Identity-Based Social Movements and Public Law, 150 U. Pa. L. Rev. 419, 519 (2001) (“The most serious criticism of the Court would be that it has meddled so much in the political process as to have ‘corrupted’ it.”).


219 Many scholars worry that judges lack the competence to interpret legislative history, a subject much closer to their area of expertise than landscape analysis. See, e.g., Adrian Vermeule, Legislative History and the Limits of Judicial Competence: The Untold Story of Holy Trinity Church, 50 Stan. L. Rev. 1833, 1833–34, 1858–63 (1998) (discussing the argument against judges interpreting legislative history).

220 See Thomas Healy, Stigmatic Harm and Standing, 92 Iowa L. Rev. 417, 471 (2007) (“Judges may be able to evaluate whether a plaintiff has lost money, faces increased competition, or has been deprived of the ability to enjoy a particular forest. But they are simply not competent to evaluate the meaning of social phenomenon.”).

221 See W. Michael Schuster, Claim Construction and Technical Training: An Empirical Study of the Reversal Rates of Technically Trained Judges in Patent Claim Construction Cases, 29 Quinnipiac L. Rev. 887 (2011). Mr. Schuster analyzed the undergraduate majors of U.S. district court judges, finding that few have degrees in the
Of course, as the previous sections of this Article suggested, a host of places across the country remain so unquestionably infected with archaic ideas about race that even a layman could properly decode them. The downtown of Fort Mill, South Carolina, is home to a monument dedicated to “faithful slaves,” 222 an intersection in central New Orleans commemorates a paramilitary group that worked to disenfranchise African Americans,223 and sites across the United States paint Native Americans as unenlightened savages.224

However, the difficulty with greater judicial oversight of the landscape is that other places offer a more complex set of messages that threaten to stretch courts’ know-how and institutional capacities. Imagine, for example, a statue of Confederate general James Longstreet sitting at the border of a traditionally black neighborhood. During the Civil War, Longstreet loyally served the cause of the South, winning key military victories at Fredericksburg, Second Bull Run, and Chickamauga.225 Yet, after the struggle, Longstreet defected to the Republican Party,226 enjoyed a long and successful career working for the U.S. government,227 embraced equal political rights for African Americans,228 and led black troops in defense of the Reconstruction-era government of New Orleans.229 Anyone exposed to even the faintest hint of postmodern theory can spot the dilemma awaiting a court in this case. Given Longstreet’s history, is the statue a spatial primer on white domination, or does it symbolize the possibility of racial reconciliation among formerly antagonistic groups? If reasonable minds could disagree about the interpretation, how would a judge decide which reading to privilege?

social sciences (other than political science and economics). Id. at 923–24. To be fair, these same kinds of problems—actually, these exact problems—regularly pop up in First Amendment cases, especially those involving art. See Christine Haight Farley, Judging Art, 79 TUL. L. REV. 805 (2005) (describing how judges have struggled to define what constitutes art).


226 Id. at ix.

227 Id. at 167 (detailing Longstreet’s work as United States Commissioner of Railroads).

228 See JEFFRY D. WERT, GENERAL JAMES LONGSTREET: THE CONFEDERACY’S MOST CONTROVERSIAL SOLDIER 410–11 (1993) (detailing Longstreet’s argument that Southerners must accept the terms offered by the North, including black suffrage).

229 PISTON, supra note 225, at 123 (“Longstreet led the state’s largely black troops against the insurgents, many of whom were Confederate veterans.”).
Defenders of substantive bans on racialized landscapes might contend that creative legal thinking can unwind the knot. A court could, for example, establish a presumption as a starting point for analysis. The doctrine could state that if some identifiable group of people reasonably takes offense at a government-sponsored message of second-class citizenship, then the challenged landscape would be razed or, perhaps, subjected to some form of heightened scrutiny. Although such an approach would screen out claims by the isolated curmudgeon who sees a racial insult at every turn, it would also raise other seemingly insoluble problems for judges to unlock. For example, would “identifiable groups” of all sizes have the power to raise a claim? Of equal import, a presumption fails to offer any systematic or reliable method for judges to measure what constitutes a “reasonable” response to culturally contested symbols. The consequences of this uncertainty could have far-reaching effects. If courts are handed the authority to demand the destruction of racialized spaces without an objective set of standards to ground their decisionmaking, innocent landscapes might be destroyed, racist places might be preserved, and inconsistent decisions would likely spring up between jurisdictions like mushrooms after a rain.

In sum, there are many reasons to doubt that a panel of judges can competently weigh and evaluate the claims of those who must experience the landscape’s close, visceral presence in the course of their daily spatial routines. At the most fundamental level, judges lack the training to read the messages stitched into the built environment. Moreover, many spaces resist single, clear interpretations, thereby increasing the difficulty of articulating coherent legal tests and, ultimately, enmeshing judges in the midst of disputes that they lack the tools to solve.

2. A Question of Political Backlash

In addition to fears that the entire interpretive project is plagued with ambiguities, a court-centered strategy may also trigger damaging political backlash against landscape fairness issues. An emerging chorus of “backlash scholars” has demonstrated that when courts announce decisions that diverge too sharply from prevailing norms, opponents will endeavor to delegitimize the transformative aspects of the ruling and contest the ethical vision that underpins it. More
specifically, scholars have shown that backlash sparks efforts to restrict the class of individuals protected by the legal change, attacks on the moral worth of the new regime’s beneficiaries, and efforts to catalogue how the ruling results in “unfair, absurd, or otherwise normatively undesirable outcomes.” \footnote{Krieger, supra note 231, at 493.} The danger of these effects is clear: they ultimately impede the realization of the goals of the original lawsuit.

Would a judicial ruling that mandates widespread landscape reform trigger such potentially unproductive results? The answer, almost certainly, is yes. Any decision that orders a dramatic change to the built environment would likely ruffle the feathers of important stakeholders within a community. Take, for example, the landscapes that celebrate the achievements of the Confederacy. These spaces remain the most viscerally racialized places in the South and would be the obvious first targets of a judicial campaign against discriminatory environments.\footnote{See Jonathan I. Leib, Separate Times, Shared Spaces: Arthur Ashe, Monument Avenue and the Politics of Richmond, Virginia’s Symbolic Landscape, 9 CULTURAL GEOGRAPHIES 286 (2002) (outlining the racial politics that underlie Richmond’s monuments to Confederate heroes).


David Weigel, How Many Mississippi Voters Wish the South Had Won the Civil War?, SLATE (Apr. 25, 2011, 2:03 PM), \url{http://www.slate.com/blogs/weigel/2011/04/25/how_many_mississippi_voters_wish_the_south_had_won_the_civil_war.html} (reporting that only 21% of Mississippi Republicans said they were glad the North won the Civil War).}

Nonetheless, large swaths of the region’s citizenry still profess affection for the Lost Cause and continue to value the memorials that transmit its values and creeds. Public surveys reveal that a majority of southerners find it appropriate for public officials to acclaim Confederate leaders,\footnote{Jon Cohen, Views on Confederates, Circa 2001, WASH. POST (Apr. 8, 2010, 10:57 AM), \url{http://voices.washingtonpost.com/behind-the-numbers/2010/04/views_on_confererates_circa_2001.html}.} and in certain places a majority of citizens still wish the South had emerged victorious from the war.\footnote{David Weigel, How Many Mississippi Voters Wish the South Had Won the Civil War?, SLATE (Apr. 25, 2011, 2:03 PM), \url{http://www.slate.com/blogs/weigel/2011/04/25/how_many_mississippi_voters_wish_the_south_had_won_the_civil_war.html} (reporting that only 21% of Mississippi Republicans said they were glad the North won the Civil War).} Consider, too, the recent tumult in Richmond, Virginia, over the fate of Monument Avenue, a cherished row of Confederate memorials. In the mid-1990s, the city council approved a proposal to erect a statue of tennis great Arthur Ashe amidst the statues change would otherwise have occurred”); Linda Hamilton Krieger, Afterword: Socio-Legal Backlash, 21 BERKELEY J. EMP. & LAB. L. 476, 477 (2000) (“[B]acklash tends to emerge when the application of a transformative legal regime generates outcomes that diverge too sharply from entrenched norms and institutions to which influential segments of the relevant population retain strong, conscious allegiance.”); Robert Post & Reva Siegel, Roe Rage: Democratic Constitutionalism and Backlash, 42 HARV. C.R.-C.L. L. REV. 373, 388–90 (2007) (summarizing the history of the “Backlash Thesis” and the work of prominent backlash scholars); Michelle A. Travis, Lashing Back at the ADA Backlash: How the Americans with Disabilities Act Benefits Americans Without Disabilities, 76 TENN. L. REV. 311, 311–12 (2009) (documenting the backlash against the Americans with Disabilities Act (ADA) and the individuals covered by the ADA).}
of Lee, J.E.B. Stuart, and Stonewall Jackson. Although the council never contemplated moving or destroying any of the Confederate statues, tensions between whites and blacks still flared dramatically. One caller to a Richmond radio station commented, “We need to protect our heritage . . . . We don’t need blacks on Monument Avenue . . . . They’ve taken over our city; they’ve tried to take over our government. If you’ve got daughters like I’ve got daughters, they’re trying to take them over, too.” Given these sociological facts, it seems appropriate to worry that a judicial attack on Confederate landscapes would harden racial attitudes and complicate the implementation of further reforms to the built environment.

Professor Michael Klarman has demonstrated how this dynamic gripped the South in the aftermath of Brown v. Board of Education. He argues, persuasively, that the Brown decision sparked a wave of anger that hampered the enforcement of desegregation orders for nearly a decade. Klarman writes, “By propelling southern politics dramatically to the right on racial issues, Brown created a political climate conducive to the brutal suppression of civil rights demonstrations.” Supreme Court rulings on affirmative action programs have also engendered widespread anger and resentment amongst whites. Many academics believe that the Grutter v. Bollinger and Gratz v. Bollinger

---

237 Leib, supra note 233, at 300.
238 See Eskridge, supra note 217, at 519 (“The most serious criticism of the Court would be that it has meddled so much in the political process as to have ‘corrupted’ it.”); Klarman, supra note 217, at 473; (arguing that judicial opinions cause backlash because “they alter the order in which social change would otherwise have occurred”); NeJaime, supra note 218, at 952. Not every academic would agree, however. A handful of scholars contend that fears about backlash have been greatly exaggerated. See Randall Kennedy, Persuasion and Distrust: A Comment on the Affirmative Action Debate, 99 HARV. L. REV. 1327, 1330 (1986) (“Given the apparent inevitability of white resistance and the uncertain efficacy of containment, proponents of racial justice should be wary of allowing fear of white backlash to limit the range of reforms pursued.”). As Professor Schragger has argued, there are reasons to be skeptical of activists shaping their “behavior to avoid backlash or to seek to ameliorate conflict in this way. Theories of political backlash are often based on small, historical samples over relatively short timeframes . . . and they tend to overestimate the impact of Court decisions.” Richard C. Schragger, The Relative Irrelevance of the Establishment Clause, 89 TEX. L. REV. 583, 637 (2011). While this observation, as a general matter, may contain a dose of truth, it flounders when confronted with the specific history of disputes over racial justice. Judicial decisions that have attempted to disrupt entrenched racial hierarchies have consistently ignited counterproductive backlashes.
240 Id. at 11.
242 539 U.S. 244 (2003).
decisions have undermined long-term support for race-conscious diversity programs; tastemakers and opinion leaders have rallied against the injustice of “reverse discrimination,” the mere suggestion that the government unfairly disadvantages whites carries “a powerful political punch,” and policies against affirmative action have become law in Washington, California, and Texas. In addition to these fights, disputes over access to integrated neighborhoods, criminal justice reform, and school busing have spawned further recoil from dominant groups that undermines the cause of racial justice.

According to scholars who study backlash, the lesson to be drawn from previous clashes over racial meaning is that judges should not prematurely constitutionalize issues that enflame primordial passions when the nation has yet to reach consensus. “Force of law,” after all, “is often little better than force of arms” in settling matters when the country is intensely divided. Taking this advice to heart, those interested in changing norms about public space should consider a more incremental approach to achieving landscape fairness than a substantive ban. As Professor Dan Kahan puts it, activists are often better off proceeding with “gentle nudges” rather than with “hard shoves.” A slower, step-by-step approach seems poised to ward off complaints that outside agitators have unfairly confiscated prized social endowments. Although not as satisfying to radicals, it may better serve the goal of landscape fairness over the long term.

A close examination of American legal precedents and cultural norms helps explain why racialized landscapes maintain their hold on dozens of cities and towns across the country. First and foremost, racialized spaces often have the firm backing of dominant political groups who seek to inscribe their own versions of history onto the land. And second, direct legal attacks on places that ingrain ideas about racial difference seem both quixotic and conceptually troublesome. Along

---

246 See ANDREW HACKER, TWO NATIONS: BLACK AND WHITE, SEPARATE, HOSTILE, UNEQUAL 44 (1992) (writing about the resistance of whites to integrated neighborhoods).
249 Levinson, supra note 223, at 1107.
these lines, any successful attempt to improve the balance of the landscape should both empower subaltern communities and navigate the political economy of dominant group decisionmaking at the local level. Otherwise, meaningful change is unlikely to be forthcoming.

IV. LANDSCAPE IMPACT ASSESSMENTS: AN ALTERNATIVE APPROACH

As earlier sections of this Article detailed, the problem of unbalanced, exclusionary landscapes imposes a disparate set of costs on African American populations. In principle, this form of discrimination is a solvable social dilemma. As the success of Kelly Ingram Park demonstrates, municipalities have the capacity to design, build, and maintain spaces that honor African Americans. Again and again, however, local governments have used their discretion to engrave the stories of dominant groups upon the landscape while ignoring the history and perspectives of subaltern communities. Can anything reverse the trend?

The pages that follow argue that what is required to rebalance the landscape is no less than a comprehensive procedural strategy that would integrate consideration of the built environment into the fiber of municipal decisionmaking. The following section begins with a forward-looking proposal. To prevent the production of new discriminatory spaces, a scheme should be implemented that would require municipalities to prepare, publish, and review a detailed impact assessment before approving any proposed construction projects. The document, in a nutshell, would describe the racial impacts of actions that reshape built environment. This Article refers to this initiative as the Landscape Impact Assessment program and argues that it would force local governments to consider the landscape outcomes of their decisionmaking and encourage them to mitigate the inequities caused by new building projects.

The Article then concludes with an attempt to dislodge established sites of discrimination. The regulatory regime envisioned would compel jurisdictions to reassess the divisive landscapes presently embedded within their borders. Specifically, the scheme would enforce a sunset provision on existing monuments and honorary spaces; each memorial in a municipality would face destruction unless the relevant local government deliberated over the meaning of the space and voted to reaffirm its value to the landscape.

This double-barreled approach has several firm advantages. The procedural mandates of both schemes would have the democracy-enhancing virtue of

---

251 See supra Part I.
252 See supra Part II.
253 See supra Part I.
254 Alexander Bickel, among others, made the case that procedural remedies are particularly appropriate in cases where parties find it difficult to settle on acceptable substantive rules. See ALEXANDER M. BICKEL, THE LEAST DANGEROUS BRANCH: THE SUPREME COURT AT THE BAR OF POLITICS 150–53, 177–83 (1962).
exposing the policy process to greater public participation, thus generating better-informed and more racially sensitive decisions. In addition, focusing on procedure does not require any sweeping overhaul of the machinery of local government. Many jurisdictions are already familiar with writing impact statements to comply with state-level environmental regulations, while the evaluations of older memorials could take place during already-scheduled city council meetings. Finally, this Article puts forth that unlike attempts to outlaw racialized spaces, a procedural approach to landscape fairness will unite the interests of all parties concerned with the fate of the built environment: subaltern groups, dominant political parties, and local governments.

A. Impact Assessments and the Future of the Landscape

1. The Architecture of the Landscape Impact Assessment Program

The goal of the Landscape Impact Assessment (LIA) proposal is to stop the production of discriminatory spaces by encouraging municipalities to consider the racial effects of actions that significantly affect the landscape. Before examining the inherent strengths of this approach, it may be helpful to pause and briefly explain the nuts and bolts of the program. What, exactly, would such a scheme ask of cities?

Under the LIA program, a local jurisdiction would need to complete three core tasks before approving any decision that affects the built environment. More specifically, I envision that the LIA requirement would apply to the full spectrum of local government decisions that affect the built environment: infrastructure planning, the approval of development agreements, zoning judgments, historic preservation standards, street renamings, discretionary building permit decisions, and subdivision regulations.

---


257 More specifically, I envision that the LIA requirement would apply to the full spectrum of local government decisions that affect the built environment: infrastructure planning, the approval of development agreements, zoning judgments, historic preservation standards, street renamings, discretionary building permit decisions, and subdivision regulations.
there, and explain whether the proposed construction efforts would negatively impact minority populations or ingrain ideas about racial power.\textsuperscript{258}

The draft LIA would not end with a description of the land. The scheme would also call upon local governments to inform citizen-voters of reasonable alternatives to any proposal that is found to negatively impact minority communities. The scope of this analysis should include discussion of other feasible sites, designs, signage, and technologies that have the potential to mitigate encoded messages of racial power. For example, in a case where a municipality proposed building a Confederate monument in the town square, several alternatives could be imagined—the local government could place the artwork in a less visible location, erect prominent signage questioning the underlying valor of the Confederate cause, or choose to honor a hero from a different age. Following the analysis of the alternate courses of action, the LIA would conclude by providing a nonarbitrary reason for why the government chose to reject those options. The sweep of the LIA would, therefore, reveal the costs and benefits associated with altering the landscape and outline strategies to ameliorate unintended racial consequences.

After completing the draft, the municipality would then need to complete a second critical task: making the report available for review and comment by the citizenry. Under LIA, localities would be required to accept both written comments from citizens and information offered at public hearings. The purpose of the hearings is both to inform local citizens about the possible landscape effects of proposed construction projects and to elicit their views about projected changes.\textsuperscript{259} Importantly, such interventions rarely happen under the current landscape regime. In Lexington, historic preservation officer Bettie Kerr reported that the city held no public hearings to discuss the recent relocation of the statue dedicated to

\textsuperscript{258} A proper analysis of the land would also closely examine the relationship between the proposed action and the spectrum of past, present, and reasonably foreseeable future changes to the built environment. A consideration of these cumulative effects would better capture the landscape’s broad meanings and prevent jurisdictions from segmenting a project into smaller, less burdensome actions. The LIA system could borrow a definition of “cumulative impact” from the environmental context. The Council on Environmental Quality defines “cumulative impact” as “the impact on the environment which results from the incremental impact of the action when added to other past, present, and reasonably foreseeable future actions . . . from individually minor but collectively significant actions taking place over a period of time.” 40 C.F.R. § 1508.7 (2012).

\textsuperscript{259} Of course, not everyone remains enthusiastic about the inherent value of public hearings. Critics often complain that community forums offer few useful insights, fail to draw a representative sample of the public, and do little more than legitimate decisions that have already been made. See, e.g., Barry Checkoway, The Politics of Public Hearings, 17 J. APPLIED BEHAV. SCI. 566, 566–72 (1981) (“[P]ublic hearings are used to achieve agency ends rather than to make effective use of citizen participation.”); Daniel J. Fiorino, Citizen Participation and Environmental Risk: A Survey of Institutional Mechanisms, 15 SCI. TECH. & HUM. VALUES 226, 230–31 (1990) (noting that hearings do not draw a representative sample of the public).
Confederate General John Breckinridge.\textsuperscript{260} Moreover, neither municipal records nor media reports reveal that the city provided any official opportunity for public input into the redesign of the courthouse plaza.\textsuperscript{261}

LIAs make such evasions impossible. The system would ensure the robust participation of private citizens by requiring governments to publish the report on the Internet;\textsuperscript{262} welcome local advocacy groups into the decisionmaking process;\textsuperscript{263} and advertise the meeting through news outlets, in local community centers, and with signs at the site of the proposed project.\textsuperscript{264} Additionally, all LIA hearings would take place at a location within the affected community (not in a faraway municipal office building) and after regular work hours, as studies have repeatedly demonstrated that attendance at public hearings spikes if the meetings occur in the evening.\textsuperscript{265} These tweaks to the traditional public hearing model will, hopefully, provide minority communities with a meaningful opportunity to comment on the scope of the LIA and on the analysis found in the document.\textsuperscript{266}

The final phase of the process would require the LIA drafters to convey to the relevant decisionmakers a final report that reflects on the information gleaned during the period of public intervention. The document should include a response to all substantive written comments and oral remarks received at the public hearing. This requirement is critical since it forces decisionmakers to grapple with criticisms and concerns they might otherwise have overlooked.\textsuperscript{267} Although some

\footnotesize{
\textsuperscript{260} Interview with Bettie Kerr, Historic Pres. Officer, Lexington-Fayette Urban Cnty. Gov’t (Feb. 11, 2011).
\textsuperscript{261} The city, it appears, only revealed finalized plans to the public. In its defense, the city notes that the public was free to talk with the designer and elected officials at the unveiling of the plans. See Beverly Fortune, Public Gets View of Downtown Streetscape Improvement Plans, LEXINGTON HERALD-LEADER, July 29, 2009, at D3.
\textsuperscript{263} See L. ELLIS WALTON, JR. & JEROME R. SAROFF, CITIZEN PARTICIPATION IN PUBLIC HEARINGS IN VIRGINIA 35 (1971) (advocating for the inclusion of interest groups in the hearing process).
\textsuperscript{264} See Checkoway, supra note 259, at 567 (explaining the ineffectiveness of advertising only in the legal section of newspapers).
\textsuperscript{265} Id. (explaining how the location of public hearings affects turnout); WALTON & SAROFF, supra note 263, at 35 (“An analysis of daytime versus evening hearings indicated that attendance is significantly higher at evening hearings.”).
\textsuperscript{266} For more on the special challenges of eliciting participation in minority communities see, John C. Duncan, Jr., Multicultural Participation in the Public Hearing Process: Some Theoretical, Pragmatical, and Analectic Considerations, 24 COLUM. J. ENVTL. L. 169, 188–202 (1999).
}
may raise alarms at the burdens such a condition threatens to impose, note that some environmental regulations enforce a similar obligation, which have not proved overly taxing. Armed with the completed LIA, the final decisionmaker would then have four choices. It could: (1) deny the project’s application because it deems the landscape impacts unacceptable; (2) approve a revised project that includes mitigation strategies designed to reduce landscape discrimination; (3) approve the original project while adopting a finding of “No Observable Effect on the Landscape” (NOEL); or (4) approve the original project and promulgate a “Declaration of Other Concerns” conveying its view that, while the proposed project will have significant adverse impacts on minority communities, other specified considerations justify approval.

2. Why the LIA Scheme Works

The thrust of the LIA proposal needs reemphasis here: its mandates are solely procedural. Local governments remain free to approve any and all proposed changes to the built environment, no matter how racially insensitive. Nonetheless, the inner workings of the program have the potential to remake the landscape. The well of the LIA’s power draws from three distinct streams. First, it empowers black communities in their clashes with city hall over the narratives being inscribed on the land. Second, it pushes government officials to consider more sophisticated (and more race-conscious) information when making decisions about the landscape. And third, LIAs mitigate the psychic harms that discriminatory places impose on African Americans. The procedural requirements, in other words, create substantive outcomes.

influence decisionmaking through public participation in the decisionmaking process); William Krueger, In the Navy: The Future Strength of Preliminary Injunctions Under NEPA in Light of NRDC v. Winter, 10 N.C. J. L. & TECH. 423, 425 n.14 (2009) (noting that in the environmental context an agency must respond to public comments to ensure that the government has taken environmental issues into account when making final decisions).

NEPA’s implementation regulations require that an agency must respond in a final impact statement to “any responsible opposing view which was not adequately discussed in the draft statement.” 40 C.F.R. § 1502.9(b) (2012).

NEPA has a similar, but more extensive, basic format. NEPA requires that an environmental assessment (EA) be prepared by a federal agency to determine if an impact statement is necessary. The EA is a concise summary of the proposed action and the possible environmental effects. If an agency expects no significant environmental impacts it can issue a finding of no significant impact (FONSI) and refrain from authoring the more detailed environmental impact statement (EIS). See Uma Outka, NEPA and Environmental Justice: Integration, Implementation, and Judicial Review, 33 B.C. ENVTL. AFF. L. REV. 601, 603–05 (2006) (explaining NEPA’s basic procedures); Emily M. Slaten, “We Don’t Fish in Their Oil Wells, and They Shouldn’t Drill in Our Rivers”: Considering Public Opposition Under NEPA and the Highly Controversial Regulatory Factor, 43 IND. L. REV. 1319, 1325–28 (2010).
(a) **Empowering Black Communities**

The LIA scheme upends disputes about the landscape by handing black communities *two* new weapons in the struggle to return balance to the built environment. First, the structured participation requirements would become an effective tool for citizens to pressure government officials about landscape issues. Currently, parties concerned about the meaning of the built environment must spend considerable organizational and political capital just to establish a beachhead for their voices and interests.\(^{270}\) To cite one example, in the early 2000s a group of landscape activists in Kentucky struggled to find a space to express their view that the grounds of the state capital excluded African Americans.\(^{271}\) They careened between appealing to unsympathetic legislators, uncaring agency administrators, and a bemused press corps.\(^{272}\) Like an echo in the woods, the activists’ arguments quickly faded from public consciousness and found no traction with government decisionmakers.\(^{273}\)

The LIA program, in contrast, would grant citizens automatic access to the seat of local power. The scheme’s public hearing requirement purposively channels and amplifies criticism of the built environment toward those with ultimate authority over the land. This result substantially reduces the transaction costs of dissent.\(^{274}\) Citizens who once struggled to organize and obtain information about local construction projects would now know exactly where to voice their frustrations—and government officials would be compelled to listen and

---


\(^{274}\) The transaction costs of participating in the democratic process can lead to suboptimal policy outcomes. See David A. Super, Laboratories of Destitution: Democratic Experimentalism and the Failure of Antipoverty Law, 157 U. PA. L. REV. 541, 565 (2008) (“An actor recognizing the value of an option that minimizes burdens on other players may nonetheless eschew cooperative, dialogic policymaking processes for several reasons. First, she may seek to avoid the transaction costs of such processes.”).
respond. Consequently, the number of individuals who voice complaints about
discriminatory landscapes and raise questions about the significance of the built
environment should swell. It is also worth noting that, as a result of the lower cost
of participation, the quality of the debate about the meanings embedded in the
landscape will improve. Under an LIA program, activists can expend more effort
crafting their arguments about the importance of the physical environment rather
than scrambling to find an audience for their complaints.

In addition to the public hearing process, the requirement that a locality
explain its actions in written records would also empower African American
communities in their attempts to ward off the bruises inflicted by discriminatory
places. Written accounts of the decisionmaking process would lead to a heightened
level of transparency and public scrutiny over landscape decisions, with
consequent political implications that elected officials disregard at their peril. Additionally, the final written LIA document could serve as the basis for lawsuits
against the government. Although the LIA has no substantive mandate, persons or
groups affected by the landscape decision could still challenge the adequacy of the
LIA procedures or the local jurisdiction’s decision to issue a NOEL finding. This
threat has already proven effective in the environmental context. Nonprofit
groups have routinely used the prospect of litigation under the National
Environmental Protection Act—another solely procedural statute—to influence the
outcome of local government decisions about the disposal of waste materials, the
construction of airports, and the siting of highways. Thus, both theory and

---

275 For a rich discussion on the benefits of reducing the costs associated with engaging
in dissent, see STEVEN H. SHIFFRIN, DISSENT, INJUSTICE, AND THE MEANINGS OF AMERICA

276 Requiring the government to explain itself “results in better decisions, elimination
of (at least) some unnecessary impacts, and arguably fewer dumb projects.” James M.
McElfish, Jr., NEPA and Liberty, Now and Forever, 39 ENVT. L. REP. NEWS & ANALYSIS
10,629, 10,630 (2009). See also Bradley C. Karkkainen, Toward a Smarter NEPA:
Monitoring and Managing Government’s Environmental Performance, 102 COLUM. L.

(2004) (outlining how NEPA can be employed as a “negative weapon”).

278 See Robert Stephans, A Case for Rancher-Environmentalist Coalitions in Coal Bed
Methane Litigation: Preservation of Unique Values in an Evolving Landscape, 8 WYO. L.
REV. 449, 454–57 (2008) (discussing the effective use of NEPA to mitigate the discharge
of dredge associated with the development of coal bed methane).

279 See Timothy R. Wyatt, Balancing Airport Capacity Requirements with
Environmental Concerns: Legal Challenges to Airport Expansion, 76 J. AIR L. & COM.

280 ENVT. LAW INST., NEPA SUCCESS STORIES: CELEBRATING 40 YEARS OF
TRANSPARENCY AND OPEN GOVERNMENT 28–29 (2010); see also Ted Boling, Making the
Connection: NEPA Processes for National Environmental Policy, 32 WASH. U. J.L. &
POL’Y 313, 325 (2010) (finding that, on average, “129 lawsuits are filed challenging NEPA
compliance each year”); David S. Mattern, Reader-Friendly Environmental Documents:
practice reveal that a comparatively undemanding procedural mandate has the potential to substantially rebalance the dynamic between the white majorities and subaltern groups in their struggles over the fate of the built environment.

(b) Changing Government Behaviors

Beyond empowering black communities, the LIA process also compels government actors to make better-informed and more racially conscious decisions about the fate of the built environment. The scheme advances toward this goal along two fronts. To start, LIAs solve a market failure in the production of information. Under the current land use regime, the meaning of the built environment remains inaccessible to decisionmakers. The government is not required to compile information about the landscape, no other party has sufficient resources or incentives to produce it, nor is it obvious that officials have authority to spend public funds to acquire such information. Thus, even officials who want to prevent the incursion of landscape unfairness into black communities may not have the information necessary to effectuate that goal. LIAs solve this problem. The procedural mandates compel government officials to gather information about the landscape and then review, understand, and address the knowledge that they scrape together. Basic democratic theory suggests that compiling more and better information should enrich the decisionmaking process and produce better outcomes. Empirical studies have largely confirmed this insight—gathering relevant information from a diversity of sources changes the behavior of decisionmakers and, ultimately, generates better and more creative solutions to difficult problems.

Opportunity or Oxymoron?, 39 ENVTL. L. REP. NEWS & ANALYSIS 10,624, 10,625 (2009) (“[B]ecause of NEPA’s procedural nature, the threat of litigation is an obvious and important tool for project opponents.”).

281 The same argument was made about environmental data before the passage of NEPA. See Karkkainen, supra note 276, at 910.

282 As Professor Sunstein has argued, America strives to be a “deliberative democracy in which representatives, accountable to the people, . . . make decisions through a process of deliberation . . . . Without better information, neither deliberation nor democracy is possible. Legal reforms designed to remedy the situation are a precondition for democratic politics.” Cass R. Sunstein, Informing America: Risk, Disclosure, and the First Amendment, 20 FLA. ST. U. L. REV. 653, 658 (1993); see also Cary Coglianese et al., Seeking Truth for Power: Informational Strategy and Regulatory Policymaking, 89 MINN. L. REV. 277, 277 (2004); Matthew C. Stephenson, Information Acquisition and Institutional Design, 124 HARV. L. REV. 1422, 1423 (2011) (“[I]nformation is the lifeblood of effective governance.”); Thomas O. McGarity, Regulatory Reform in the Reagan Era, 45 MD. L. REV. 253, 259 (1986).

283 Deborah Gruenfeld et al., Group Composition and Decision Making: How Member Familiarity and Information Distribution Affect Process and Performance, 67 ORGANIZATIONAL BEHAV. & HUM. DECISION PROCESSES 1, 4 (1996) (arguing that the inclusion of a devil’s advocate “can lead groups to generate more arguments, apply more
The LIA system further promises to change legislator behavior by injecting a new policy norm into government: local officials should not sanction projects that engrave racial hierarchy into the landscape. Of course, working in the shadow of a procedural-based law will not immediately alter the behavior of decisionmakers. It is easy to envision bureaucrats occasionally resisting the LIA process by approving developments with obvious and unabated impacts on black communities. Proponents of norm activation theory, however, suggest that to consistently ignore the recognizable harms may not be sustainable. The repeated approval of projects that damage black people and black neighborhoods conflicts strongly with the widely (although not universally) shared taboo against overtly racist acts. Over time, the tension between the harms identified by the LIA and the decisionmakers’ (presumed) antiracist ideals will cause many officials to internalize or accommodate the new norm governing the landscape. They may acquiesce for several reasons (because they were previously sympathetic to the needs of black communities, they desire to act consistently with their internal norms, or they received pressure from peers) but, eventually, their behavior should change under the weight of new expectations.

(c) Undoing the Violence of Memory

The LIA regime offers one final advantage. As the first Part of this Article detailed, the landscape decisions of elected officials effectively erase African American histories from the built environment. Although the LIA scheme cannot...
guarantee the construction of new physical spaces that exalt the black experience, it can help African Americans reclaim territory in the underlying battle over communal memory. A sustained and inclusive deliberation about the meaning of the land would cement African Americans’ own stories within the official histories of cities, counties, and states. Indeed, the true strength of the LIA may rest in its ability to enshrine a mechanism that allows black citizens to bring hidden memories out of the shadows and have them recognized in an official government forum.

Take, for example, the contested history of lynching. Between 1890 and 1968 nearly 3,500 blacks were systematically and brutally killed in communities scattered across the United States. Given the horror, it should surprise no one that reverberations of these deaths can still be felt within African American neighborhoods. “[B]lack Americans,” scholars have observed, “share a kind of communal memory of lynching that is not bound by region or by time,” and continue to view the specter of these incidents as “a powerful source of deep distrust, disconnection, and cynicism . . . between blacks and whites.” Yet, despite the ongoing repercussions, most communities have actively buried this legacy. Rather than confront the history of racial violence, locales have simply ignored the haunting presence of lynching’s memory, denying African

289 Ifill, supra note 79, at xi.
290 Ifill, supra note 287, at 269; see also James Brooke, Memories of Lynching Divide a Town, N.Y. TIMES, Apr. 4, 1998, at A6 (describing the reverberations of a 1925 lynching on the white and black communities in a small Utah town).
291 See Ifill, supra note 79, at 134 (“Some whites simply do not regard these incidents as having continuing relevance . . . . For them, discussing lynching is merely an exercise in dredging up the past . . . .”)
292 See CYNTHIA CARR, OUR TOWN: A HEARTLAND LYNCHING, A HAUNTED TOWN, AND THE HIDDEN HISTORY OF WHITE AMERICA 34, 103 (2006) (reporting that white communities have generally tried to act like lynchings never happened); Ifill, supra note 79, at xiii (arguing that dominant political interests downplay the history and effects of lynching); Garland, supra note 287, at 795 (discussing the attempt to deny and forget lynching); Ifill, supra note 287, at 266 (noting that “few public efforts have been launched to address the devastating history of lynching in the United States”).
Americans the dignity of having their experiences understood and their stories acknowledged. The LIA process can help diffuse the raw emotion that haunts such unexamined memories. Lynchings, for example, enter the ambit of the LIA program because the killings had a heavily spatial component; mobs chose the site of their executions for darkly symbolic purposes. The offenders’ macabre sense of space often led them to kill their victims on public land within view of the local courthouse, thus reinforcing the message that black bodies stood outside of the protection of the legal system. The LIA program would force local communities to examine these episodes. Upon the remodeling or redevelopment of any space where a lynching occurred, a final LIA report (which must discuss race) and the community participation requirements would compel a retelling of previously one-sided stories. The process would drag submerged black perspectives onto firm ground and provide a vehicle to confront the anger and confusions that still unsettle black-white relationships. Moreover, conducting these conversations about the landscape in an official government forum conveys the elementary principle that people deserve to be consulted about decisions that impact their lives.

The transformative power of reexamining the past is not idle speculation. Humanities scholars who study Holocaust remembrance have demonstrated that the debates surrounding the design of a Holocaust memorial generate more sophisticated thought and introspection about the fate of Europe’s murdered Jews than any completed monument. As Professor Young observes, continuous

---

293 See CARR, supra note 292, at 369 (describing lynching as “a living secret” that people “never stopped having feelings about”); IFILL, supra note 79, at 132–53 (arguing that it is critical for victimized communities to have their stories heard); MADISON, supra note 288, at 2 (arguing that communities need “closure” on lynching). Note, too, that lynching also traumatized whites. See IFILL, supra note 79, at 140–41.

294 See WILLIAM FITZHUGH BRUNDAGE, LYNCHING IN THE NEW SOUTH, GEORGIA AND VIRGINIA, 1880–1930, at 41 (1993) (showing that “[m]ass mobs chose execution sites for explicitly symbolic reasons”); KARLA F.C. HOLLOWAY, PASSED ON: AFRICAN AMERICAN MOURNING STORIES 62–63 (arguing that “[l]ynch mobs had both a macabre and sick sense of occasion and space”); PATTERSON, supra note 287, at 205 (“The selection of the lynch site was a decision loaded with religiopolitical symbolism.”).

295 See IFILL, supra note 79, at 8–9.


297 Holocaust remembrance has struggled with these issues. See, e.g., JAMES E. YOUNG, AT MEMORY’S EDGE: AFTER-IMAGES OF THE HOLOCAUST IN CONTEMPORARY ART AND ARCHITECTURE 119 (2000).

298 See id. at 92; BRIAN LADD, THE GHOSTS OF BERLIN: CONFRONTING GERMAN HISTORY IN THE URBAN LANDSCAPE 11 (1997) (“The process of creating monuments, especially where it is openly contested, as in Berlin, shapes public memory and collective identity.”); James E. Young, The Counter-Monument: Memory Against Itself in Germany Today, 18 CRITICAL INQUIRY 267, 267–70 (1992) (“[T]he surest engagement with memory lies in its perpetual irresolution.”); see also Pierre Nora, Between Memory and History: Les
reflection can remake and renew the life of memory, while “the finished monument . . . completes memory itself, puts a cap on memory-work, and draws a bottom line underneath an era that must always” unsettle civilized peoples. Legal theorists, too, acknowledge that deliberation about the past in recognized public forums can reshape the contours of official memory and help heal the wounds of formerly subjected groups. For instance, in her work on truth and reconciliation committees, Professor Ifill shows that many communities have recreated themselves by engaging in difficult, systematic discussions about the traumas hidden within the coils of history. She notes that the opportunity to challenge official histories is a key element in the reconciliation process and goes “a significant way in balancing the harm done.”

Summing up, the LIA program has potential to reshape the ways that communities will construct the built environment. The ultimate triumph of the LIA process, however, may not hinge entirely on concrete changes to the landscape. A LIA regime can also succeed by enshrining a more inclusive set of memories into local history and by providing African Americans the opportunity to tell their own stories (in their own voices) to local powers. It is precisely because silence was used to coerce complicity and impose terror in black neighborhoods that truth-telling and public speaking remain so vital to the struggle to define what kind of memory to preserve and in whose name.

3. Answering Objections

(a) Political Economy

Of course, not everyone will agree that the LIA program mounts an effective assault on the problem of landscape fairness. Skeptics of the idea would have several bridges to defend their fortress. Most notably, this Article has yet to address whether locally dominant political groups would ever willingly adopt an LIA regime. Arguably, the same factions that have worked to construct discriminatory landscape features would conspire to oppose any law that helps black communities defend their interests. Admittedly, this criticism merits scrutiny. Yet, on close examination, the complaint does little to derail the case for landscape impact assessments.

To start, the underlying premise of the critique—that dominant political groups will resist the LIA idea—lacks a solid foundation. LIAs, as mentioned

Lieux de Memoire, 26 REPRESENTATIONS, Spring 1989, at 7, 12–13 (arguing that physical monuments supplant a community’s memory work, rather than embody it).

See Ifill, supra note 79, at 117–31 (looking at reconciliation and lynching in an international context).

Id. at 128 (quoting TERESA GODWIN PHELPS, SHATTERED VOICES: LANGUAGE, VIOLENCE, AND THE WORK OF TRUTH COMMISSIONS 55 (2006)).
above, do not command any particular substantive outcome. Instead, the proposal encourages municipalities to draw a diverse set of voices into the decisionmaking vortex and thoughtfully disclose the downstream consequences of its actions. The scheme, in other words, adopts—rather than resists—the political philosophy of groups most likely to oppose the LIA’s underlying aims. The focus on procedure, for example, preserves “individual choice while avoiding direct governmental interference.”

Moreover, the information-forcing component should appeal to those with a promarket political bent, as it addresses a market failure without otherwise clogging the free flow of ideas. Finally, the inviting rather than punitive design grants all citizens (not just African Americans) an opportunity to participate in shaping the landscape and, consequently, works to diffuse claims that any one group has received special protections or additional rights.

For those not yet convinced, note that sponsors of the impact assessment approach could also make the scheme more politically acceptable by mandating that the LIA documents consider the built environment’s effect on gender as well as race. Geographers have forcefully argued that the landscape plays a dramatic role in the construction of gender identities. Although a feminist critique of the landscape lies beyond the scope of this Article, strong evidence exists that the built environment is not just “scenery for the playing out of gender,” but rather a tool to “shape the ways gender identities and relations are played out, reinforced or modified.”

Women, too, have been denied equal access to important public


305 The American experience with environmental regulations adds weight to the argument that procedural laws often incite fewer political objections than purely substantive legal rules. The success and longevity of procedure-based statutes like NEPA, the Emergency Planning and Community Right to Know Act, the 1996 amendments to the Safe Drinking Water Act, and California’s Proposition 65, all demonstrate that it remains remarkably difficult to incite public anger or moral outrage against a statute that seeks only to ensure that the parties have contemplated all sides of an issue.

306 See generally A COMPANION TO FEMINIST GEOGRAPHY (Lise Nelson & Joni Seager eds., 2005) (discussing all aspects of the feminist geography including contexts, work, city, body, environment, state, and nation); GILLIAN ROSE, FEMINISM & GEOGRAPHY: THE LIMITS OF GEOGRAPHICAL KNOWLEDGE (1993) (discussing the intersection of feminist thought with the practice of human geography); FEMINISMS IN GEOGRAPHY: RETHINKING SPACE, PLACE, AND KNOWLEDGES (Pamela Moss & Karen Falconer Al-Hindi eds., 2008) (providing a unique, reflective approach to what feminist geography is and who feminist geographers are).

places and excluded from prominent roles in the nation’s memorial landscapes.\textsuperscript{308} Expanding, slightly, the ambit of the LIA report to tackle the intersection of place and gender should capture the additional political support necessary to enact the scheme into law.

\textit{(b) Cost}

Another criticism that opponents could make against the LIA proposal is that its costs would far outweigh any of the ultimate benefits. LIAs, like any government regulation, would impose a new set of expenses on local jurisdictions.\textsuperscript{309} The costs of the program include the hiring of staff to produce and disseminate the draft reports, organize public hearings, and respond to comments received from interested parties about landscape issues. The LIA process would inevitably delay some construction projects, increasing financial costs for developers and their allies. Additionally, legal fees could become a significant expense if activists aggressively challenge the adequacy of a jurisdiction’s compliance with the procedural mandates.

One may accept the full weight of this criticism yet still recognize that a system of LIAs produces valuable society-wide benefits. A reminder of the existing system may help explain. Recall that government agents, under pressure from locally dominant groups, currently use the land to record selective versions of the past—versions that valorize white heroes, fail to acknowledge black suffering, and disregard African American accomplishment.\textsuperscript{310} The distortions and omissions ultimately drive black citizens from important public spaces and discourage their full participation in the polity. Gerald Smith, a professor at the University of Kentucky, asks a question that encapsulates the problem: “Why would African Americans want to go anywhere there’s a confederate soldier? There’s a painful memory in that.”\textsuperscript{311} Arguably, the costs of this current reality—the exclusion of black people from vital civic spaces, the attendant loss of their voices, and the sting of demoralization—far outweigh the monetary expense imposed by an LIA regime. That is, if we put “on-screen” the full panoply of costs that accrue under both systems, we may conclude that the impact assessment approach actually comes with the lower price tag, even though it forces governments to absorb a set of expenses that show up on a municipal budget.

\textsuperscript{308} See \textsc{The Dictionary of Human Geography} 244–48 (Derek Gregory et al. eds., 5th ed. 2009).
\textsuperscript{309} Iglesias, \textit{supra} note 284, at 514.
\textsuperscript{310} See \textit{supra} Part I.
\textsuperscript{311} Interview with Gerald Smith, Professor of History, Univ. of Ky. (June 22, 2011).
B. Sunset Provisions and the Landscape’s Past

The LIA scheme would do much to prevent the construction of projects that threaten to carve racial fault lines into the built environment. Even at its best, however, the proposal does not fully untangle the knot of landscape unfairness. The core weakness is easily visible to the naked eye: LIAs would have little power to challenge the divisive landscapes that already mar the American scene. Put more bluntly, discriminatory places built before the passage of an LIA statute would linger on without rebuke, freely transmitting their messages of racial difference from one generation to the next. To patch this hole, this Article envisions a set of regulations that would require municipalities to confront the racialized landscapes presently ensconced within their borders. Specifically, the scheme would enforce a sunset provision on all existing monuments and honorary spaces. In use since the founding of the republic, sunset clauses provide that a government policy will cease to have legal effect on a predetermined date unless a legislature takes affirmative steps to extend the measure.\(^{312}\) Thus, under a sunset regime each memorial in a jurisdiction would face destruction unless the relevant local government deliberated over the meaning of the space and voted to reaffirm its value to the landscape.\(^{313}\)

It is an obvious point that threatening to demolish thousands of cherished monuments raises difficult questions about what kind of memory societies should preserve. Arguably, the destruction of memorial landscapes amounts to an attack on a community’s shared history and cultural understandings. As Professor Brophy has eloquently argued, “[R]emoval . . . threatens our memory of the past. It is not that I have any particular interest in honoring someone like Thomas Ruffin, who . . . wrote an opinion that released a man from criminal liability for abusing a slave in his custody.”\(^{314}\) “Yet,” Brophy writes, “I think we should keep his name on the dormitory on the University of North Carolina campus because it is part of

\(^{312}\) Black’s Law Dictionary defines a “sunset law” as a “statute under which a governmental agency or program automatically terminates at the end of a fixed period unless it is formally renewed.” BLACK’S LAW DICTIONARY 1574 (9th ed. 2009). Defined loosely, we can trace the intellectual lineage of sunsets as far back as Thomas Jefferson, who wrote, “Every constitution . . . and every law naturally expires at the end of 19 years.” Letter from Thomas Jefferson to James Madison, (Sept. 6, 1789), in 15 THE PAPERS OF THOMAS JEFFERSON 396 (Julian P. Boyd ed., 1958). Additionally, Alexander Hamilton, in The Federalist No. 26, argued in favor of limiting military appropriations for two-year periods in order to encourage periodic legislative oversight of expenditures. THE FEDERALIST NO. 26 (Alexander Hamilton).

\(^{313}\) But see Alfred L. Brophy, The Law and Morality of Building Renaming, 52 S. TEX. L. REV. 37, 46–51 (2010) (discussing complications that could arise from gifts to cities that are conditioned on the presence of memorials).

\(^{314}\) Id. at 66.
our history and because we should remember that there was a time when his ideas were triumphant."315

This argument against upsetting the established order carries with it a heavy theoretical and emotional punch. Ultimately, however, claims that removing monuments amounts to rewriting history are wrongheaded. The fundamental difficulty with Brophy’s argument is that the landscape has never been a neutral record of the past. As this Article has labored to explain, the landscape is a normative discourse fashioned by powerful groups and used to express messages about who belongs and who does not.316 Thus, changing the composition of a jurisdiction’s monuments does not erode any universal, objective truth in the name of political correctness; rather, it initiates a process of critically rethinking what values a community holds and who deserves the honor of being remembered in steel and stone.

Forcing locales to reimagine the landscape has two primary benefits for communities. First, a consensus has emerged in the academic literature that legislation with temporal limits produces information effects that ultimately improve decisionmaking.317 The logic of the argument is easy enough to follow. To extend the life of a policy encumbered with a sunset provision requires a reboot of the legislative process in later time periods.318 Most scholars agree that these subsequent stages of procedure provide an additional opportunity to integrate valuable information into the policy process and guard against the continuation of unwise decisions.319 In the landscape context, for example, a sunset provision would compel elected officials to reevaluate its memorial spaces—some of them hundreds of years old. As questions about the value of individual monuments came forward, municipalities would have an opportunity to incorporate fresh scholarly

---

315 Id. at 67.
316 See supra Part I.
318 See Gersen, supra note 317, at 251.
research into their assessments of the people, places, and events that they have honored.\textsuperscript{320} New historical understandings should ultimately push local governments to reappraise the value and meaning of existing landscapes. Some cities may choose to remove a small handful of monuments, but many more should tweak the built environment to promote the inclusion of minority groups and improve the historical accuracy of memorial spaces.

Second, changing the default rule on policy continuation would have the democracy-enhancing virtue of improving oversight and accountability of local politicians.\textsuperscript{321} Under the current legal regime, elected officials often choose to sidestep conflicts over controversial spaces rather than risk upsetting a deep-pocketed interest group—like the United Daughters of the Confederacy—that could recruit and fund a credible challenger. Like a lightning flash from a hot cloud, sunset provisions would shatter legislators’ ability to continue ignoring landscapes that marginalize subaltern groups. Faced with the looming destruction of a memorial space, the law would force an elected official to “deliberate” upon racialized places, “come to a new resolution” on their value, and cast a “vote in the face of [potentially angry] constituents.”\textsuperscript{322} The accountability that results from requiring legislators to record a position on landscape fairness would open up fresh avenues for localized democratic expression.\textsuperscript{323} For the first time, African Americans unsatisfied with the shape of the built environment would know who stands in sympathy with their ideals—and who does not. Legislators who continually vote to re-anchor racialized monuments and memorials would face the potential loss of political support from African Americans (and their allies), and an attendant drop in campaign contributions.

Importantly, these benefits could be achieved at an acceptable cost to municipalities. At first blush, however, observers concerned about the efficiency of local governments may balk at the scale of the undertaking. In New York City alone, over one thousand monuments lie scattered across the metropolis,\textsuperscript{324} and in Chicago, the park system hosts nearly two hundred memorial fountains and

\textsuperscript{320} See, for example, the recent case of Simkins Dorm at the University of Texas. As new scholarly information came to light about William Simkins, the university fell under more and more pressure to remove his name from a student dorm. Brophy \textit{supra} note 313, at 37.

\textsuperscript{321} \textit{See, e.g.,} Finn \textit{supra} note 319, at 447.

\textsuperscript{322} \textit{See} \textit{THE FEDERALIST NO. 26,} at 71 (Alexander Hamilton) (Clinton Rossiter ed., 1961).


sculptures. Requiring local jurisdictions to evaluate thousands of monuments could, conceivably, throw sand in the gears of municipal government and divert time and resources from more important undertakings. There are two defenses to this claim. First, over the long haul, sunset clauses would promote efficiency and good government by chipping away at the presence of exclusionary landscapes and better integrating minority communities into the social fabric of cities. Second, governments could easily structure sunset clauses to reduce the strain on individual legislators. One promising option would cabin the scope of the sunset program by requiring elected officials to evaluate spaces only after a private citizen filed an official “reassessment request.” Alternatively, the scheme could grant municipal governments a generous ten- or fifteen-year period to assess the monuments and memorial spaces within their borders. Dispensing the work over a larger span of time would allow elected officials to treat the landscape with the moral seriousness it deserves, without encroaching too far on the other business of government. It is also comforting that bigger cities—the places with the most memorial landscapes and the largest burdens under this scheme—have entire departments dedicated to maintaining and evaluating public art. In New York City, for example, both the Art and Antiquities Division and the Design Commission could provide the city with expertise on the most controversial spaces, further lightening the burden of the legislative body.

To sum up, the proposal to regulate monuments through a web of sunset clauses is a purely procedural mandate that does not require any destruction. As with the LIA regime, sunset provisions would not impose unwanted substantive changes on unsuspecting municipalities. Local jurisdictions would remain free to reanchor any landscape features under their control, as long as the legislative body recorded a public vote to reaffirm the threatened monument. Thus, rather than exact a widespread razing of the land, enacting a sunset clause would begin a public conversation about the shape of the built environment and make it more difficult for governments to protect the memorial landscapes that marginalize subaltern groups.

CONCLUSION

It is the spring of 2009 and I am standing under the rotunda of the capitol building in Frankfort, Kentucky. My wife has just taken the oath of office for the


Kentucky Bar and I am trying to find her among a throng of state officials and excited family members. As I wait, my eyes lock on the Jefferson Davis statue.

Looking at Davis, I feel a hot anger crawl under my skin. Why is he here? Why does a state that never joined the Confederacy—a state that sent thousands of its sons to die in Union colors—honor the founder of a republic raised solely to preserve black slavery? The moment suddenly gets worse. A father coaxes his young daughter, dressed in a Mickey Mouse t-shirt and cut-off jeans, to stand in front of the statue. As she looks up at Davis and smiles, her father snaps a picture. The caption on the statue’s pedestal where she rests her hand is the final dagger: “Jefferson Davis: Patriot • Hero • Statesman.”

A landscape like the capitol grounds in Frankfort, Kentucky, would have garnered few supporters in the early days of the American Republic. Many in the founders’ generation liked to claim that democratic governance rendered commemorative landscapes obsolete—a holdover from the days of mass illiteracy and superstition. True memory, it was thought, did not lay sealed in bronze or entombed in earthworks. Rather, it resided in the living hearts of the people.

Since the founders’ time, American opinion has dramatically changed course. Cities and towns now routinely embed histories and conduct meanings through the built environment. Commemorative landscapes have become utterly commonplace. Yet, despite its omnipresence, legal scholarship has not fully examined the role of the landscape in the cultural practices of the nation.

This Article has attempted to fill the scholarly void with three insights. First, the landscape plays a powerful role in shaping how communities think about race and racial power. All too often, a city’s parks, street names, monuments, and memorial spaces conspire to tell stories that praise white achievement, ignore or misrepresent the history of African Americans, and work to physically exclude blacks from important civic spaces. Second, current efforts to reform the landscape seem doomed to fail. The approaches suggested by academics in law and geography either turn a blind eye to the political economy of local decisionmaking or fail to consider entrenched legal precedent. Third, adopting a set of basic procedural requirements that encourage municipalities to consider the landscape’s absences and marginalizations would result in better-informed and more racially sensitive decisionmaking. Procedural rules would force government to internalize values it might otherwise ignore, allow citizen-critics to challenge dominant historical narratives and push communities to view the past (and express its hopes for the future) in much more diverse terms.

---

327 Reflecting on Congress’ unwillingness to fund a monument to George Washington, John Quincy Adams remarked that “democracy has no monuments.” 1 MEMOIRS OF JOHN QUINCY ADAMS, COMPRISING PORTIONS OF HIS DIARY FROM 1795 TO 1848, 433 (Charles Francis Adams ed., Phila., J.B. Lippincott & Co. 1874).
TURNOVER ACTIONS AND THE “FLOATING CHECK” CONTROVERSY

David R. Hague*

Abstract

When a debtor files for Chapter 7 bankruptcy, a Chapter 7 trustee is appointed and is charged with collecting and reducing to money the property of the bankruptcy estate. One of the most basic collection methods a trustee possesses is its turnover power under § 542(a) of the Bankruptcy Code. Pursuant to § 542(a), an entity in possession, custody, or control, during the bankruptcy case, of property that the trustee may use, sell, or lease, must deliver to the trustee, and account for, such property or the value of such property.

An interesting issue has arisen that is placing debtors in very problematic situations. Prior to filing for bankruptcy, debtors are writing and issuing checks, but the checks are not clearing until after the bankruptcy case is filed. Armed with the § 542(a) collection power, trustees are demanding that the debtor replenish the bankruptcy estate and turn over the account balance that existed on the date the debtor filed for bankruptcy. But debtors are refusing to comply with this demand because the funds represented by the checks are no longer in the account. So who is responsible for replenishing the estate for the transferred funds? Is the onus on the debtor to turn over the funds, even if those funds have been transferred from the estate to the payees? Or does the trustee bear the burden to seek the postpetition payments from the payees of the checks through avoidance actions?

This Article examines the “floating check” controversy and the language of § 542(a) of the Bankruptcy Code. It also examines a Chapter 7 trustee’s duties to maximize the bankruptcy estate for the benefit of creditors. This Article then reviews one of the leading cases on the floating check controversy, which holds that a debtor is not liable to the bankruptcy estate for the value of the funds if she lacks current possession or control of the actual funds at the time the trustee makes the demand for turnover. Several courts have followed this decision. These courts rely on pre-Bankruptcy Code practice and hold that turnover is permissible only when the entity has possession or control of the property at the time the turnover action is filed. Some of these courts also justify their decisions with policy-based arguments, analyzing who is in the best position to prevent transfers by postpetition check and remedy the damages to the bankruptcy estate.

* © 2013 David R. Hague. Assistant Professor of Law, South Texas College of Law. I would like to thank Professor Chad Pomeroy and Professor Stephen Ware for their helpful comments.
After examining these arguments, this Article uses the relevant provisions of the Uniform Commercial Code governing the status of funds represented by an issued check to argue that a payee of a check only obtains possession and control of those funds represented by the check once the funds are available to the payee. This Article then argues that the enactment of the Bankruptcy Code altered the pre-Code “current possession or control” requirement because § 542(a) expressly permits a trustee to recover “the value” of the property, in addition to the property itself, from one who possessed the property at any time “during the case.” As such, this Article concludes that if a debtor writes checks against funds prepetition, but the checks do not clear the debtor’s account until after she files for bankruptcy, the trustee is entitled to a money judgment against the debtor for the value of the funds.

Finally, while this Article addresses the policy concerns, it raises a new approach that courts have failed to consider. Instead of analyzing who is in the better position to prevent transfers by postpetition checks or which party is in the best position to remedy the damages to the bankruptcy estate, this Article poses a simple question: which approach for recovering the funds is in the best interest of the estate and its creditors? This Article concludes that a Chapter 7 trustee has several nonexclusive remedies and, in the exercise of her business judgment, may choose whatever recovery method is in the best interest of the estate. At times, recovering from the debtor might make the most sense because such remedy allows the trustee to recover the value of all the prepetition checks from one source and without having to commence a lawsuit. Sometimes, however, recovering the funds from the payees of the checks provide the greatest return. Not only does this approach comport with the trustee’s duties under the Bankruptcy Code to maximize a return to creditors, but it is what § 542(a) and the Bankruptcy Code allow.

I. INTRODUCTION

“Timing is everything in bankruptcy.”1 Shortly before filing for Chapter 7 bankruptcy, a debtor writes several checks from her checking account, but before the payees of the checks present them to the bank for payment, the debtor files for bankruptcy. A few days later—after the bankruptcy filing—the checks clear, leaving only a few dollars in the debtor’s bank account.

When a debtor files for Chapter 7 bankruptcy, a new legal person—the estate—is automatically created. A Chapter 7 trustee is appointed to act on behalf of the estate and is charged with collecting and reducing to money the property of

the estate.\textsuperscript{2} It is ultimately the trustee’s duty to collect property of the estate “as expeditiously as is compatible with the best interests of the parties in interest.”\textsuperscript{3} One of the most basic collection methods a trustee possesses is her turnover power under § 542(a) of the Bankruptcy Code. Section 542(a) requires any entity that is in possession, custody, or control of property\textsuperscript{4} that the trustee may use, sell, or lease to turn that property over to the trustee and account for such property or its value.\textsuperscript{5}

In administering the bankruptcy estate, the trustee typically reviews the debtor’s bank account statements to determine if there were funds in her account on the date the debtor filed for bankruptcy. In the example above, because the checks did not clear the debtor’s bank account until after the filing date—postpetition—there were clearly funds in her bank account when she filed for bankruptcy. As a result, the trustee files a turnover action against the debtor, pursuant to § 542(a), demanding the debtor to turn over to the estate the account balance that existed on the date the debtor filed for bankruptcy. Because the funds represented by the checks are no longer in the account, however, the debtor refuses to (or simply cannot) comply with the trustee’s demands.

Are the funds that the bank used to honor the checks property of the debtor’s bankruptcy estate and, therefore, subject to turnover? If so, who is responsible for replenishing the estate for the transferred funds? Is the debtor responsible even though she no longer has control or possession of the funds in the account and even if it means she will have to pay the funds twice—once to the payees to whom the checks were originally payable and the second time to the trustee? Is the trustee’s only remedy against the payees through separate avoidance actions? Or does the trustee have the choice whether to recover from the debtor or the payees?

Nearly all bankruptcy court decisions addressing this floating check controversy agree that if the funds are still in the debtor’s bank account on the petition date, then such funds are property of the estate. But § 542(a)’s second requirement of “possession, custody, or control during the case” is not clear-cut, and courts have split over its interpretation. Some hold that the trustee bears the burden to seek the postpetition payments from the payees of the checks since the debtor no longer has current possession or control of the funds, while other courts put the onus on the debtor to turn over the funds, even if those funds have been transferred by the debtor’s bank to payees.

This Article argues that the latter courts are correct. If a debtor writes checks against funds prepetition, but the checks do not clear the debtor’s account until after she files for bankruptcy, the trustee—if she so chooses—is entitled to a


\textsuperscript{3} Id. § 704(a)(1).

\textsuperscript{4} The “property” referred to in § 542(a) is essentially “property of the estate.” See id. § 541(a)(1) (defining property of the debtor’s estate as “all legal or equitable interests of the debtor in property as of the commencement of the case”).

\textsuperscript{5} Id. § 542(a).
money judgment against the debtor for the value of the funds. Not only does this view comport with the plain language of § 542(a), but it also may be the most practical and efficient way to maximize the value of the bankruptcy estate. Some courts that have considered the floating check controversy fail to consider the relevant provisions of the *Uniform Commercial Code* (U.C.C.) governing the status of funds represented by an issued check. And several courts support their holdings on pre-Bankruptcy Code practice without accounting for the text of the current Bankruptcy Code. These courts hold that § 542(a) permits a trustee to compel turnover only from entities that have control of property of the estate at the *time* of the turnover demand. But what is absent in these courts’ decisions is appropriate deference to the Bankruptcy Code’s language giving the trustee the right to compel the delivery of “the value of such property”6 instead of the property itself. Similarly, several courts take the position that the trustee is in a better position than the debtor to recover funds from payees and, therefore, this somehow obligates the trustee to pursue these payees instead of the debtor. Such reasoning is flawed and inconsistent with the Bankruptcy Code and the trustee’s duties to maximize the estate as “expeditiously as is compatible with the best interests of parties in interest.”7

Part II of this Article provides an overview of Chapter 7 bankruptcy. It begins by discussing a debtor’s duties to cooperate as necessary to enable a Chapter 7 trustee to perform the duties given it under the Bankruptcy Code. One of those duties is to use its ability to compel the turnover of property of the estate to the trustee pursuant to § 542(a) of the Bankruptcy Code. This power is what entitles a trustee to recover a money judgment against the debtor for the value of funds withdrawn postpetition, even if the debtor no longer has possession or control of such funds.

Part III of this Article examines the language of § 542(a) and the elements a trustee would need to establish to prevail on a § 542(a) turnover claim. In addition, Part III examines one of the leading cases siding with the debtor on the floating check controversy—*Brown v. Pyatt (In re Pyatt)*.8 This case holds that a debtor is not liable to the bankruptcy estate for the value of the funds if she lacks current possession of the actual funds at the time the trustee makes the demand for turnover.9 Several courts have followed the *Pyatt* decision. This Article argues that *Pyatt*’s reasoning is without merit and should not be followed.

Part IV argues that for a trustee to maintain a turnover action against the debtor, the trustee need only prove that the funds in the debtor’s bank account are property of the estate and that the debtor had control over the funds at some point during the bankruptcy case. To satisfy these elements, Part IV uses the relevant provisions of the U.C.C. and further compares the pre-Code “present possession” requirement against the plain language of the current Bankruptcy Code.

---

6 Id.
7 Id. § 704(a)(1).
8 486 F.3d 423 (8th Cir. 2007).
9 Id. at 429–30.
Finally, while Part IV addresses the policy concerns that have been raised by several courts regarding the floating check controversy, it also raises a new approach that has not been addressed by the courts. Specifically, Part IV poses the following question: Which approach for recovering the funds is in the best interest of the estate and its creditors? Several courts that analyze the floating check controversy err in concluding that the crucial question is who is in the better position to prevent transfers by postpetition checks or which party is in the best position to remedy the damages. Rather than focusing on these issues, this Article argues that the proper analysis is not necessarily which party is in a better position, but which method of recovering the funds provides the greatest return to the bankruptcy estate. The Bankruptcy Code requires the trustee to collect property of the estate as expeditiously as is compatible with the best interests of the parties in interest. As set forth in Part IV, a trustee has a choice of nonexclusive remedies. In some cases, recovering the funds from the debtor will provide the greatest return to creditors. But in other cases, recovering the funds from the payees might be more efficient and in the best interests of the estate. The point is that a trustee has options under the Bankruptcy Code, and the business judgment rule governs her choices. This approach—and more specifically the trustee’s right to choose her remedy—is supported by the plain language of § 542(a) and comports with the trustee’s duties under the Bankruptcy Code.

II. OVERVIEW OF CHAPTER 7 BANKRUPTCY

A. Bankruptcy—In General

To fully understand the floating check controversy, one must understand general bankruptcy law. Bankruptcy law is federal law. “The substantive provisions of the Bankruptcy Reform Act of 1978 are found in Title 11 of the United States Code and are referred to as the ‘Bankruptcy Code’ or simply the ‘Code.’”

There are two overall forms of bankruptcy relief: (1) liquidation, and (2) rehabilitation or reorganization. The Bankruptcy Code divides the substantive law of bankruptcy into the following five chapters: (1) Chapter 7 cases, (2) Chapter 9 cases, (3) Chapter 11 cases, (4) Chapter 12 cases, and (5) Chapter 13 cases.

---

10 Thomas J. Salerno et al., Advanced Chapter 11 Bankruptcy Practice § 1.4 (2013). The jurisdiction, venue, and administrative provisions appear in Title 28 of the United States Code. 28 U.S.C. § 1334 (jurisdiction); id. §§ 1408–1410 (venue); id. § 1411 (jury trial); id. § 1452 (removal of cases from state courts); id. §§ 151–155 (judgeship provisions); id. § 156 (administrative/staff); id. § 157 (referral of cases from district court to bankruptcy courts); id. § 158 (appellate procedures, creation of bankruptcy appellate panels); id. § 1930 (filing fees); id. § 602 (creation of private panel of trustees); id. §§ 581–589 (U.S. trustees); id. at § 959 (capacity of trustees to be sued).


12 Chapter 9 is available only to a municipality and only by means of a voluntary petition. Id. § 109(c).
The availability of each of these chapters, however, is based on characteristics of the debtor at the time of filing.16

“Bankruptcy serves to mitigate the effects of financial failure.”17 For debtors who are individuals, the Bankruptcy Code affords the possibility of a fresh start through the bankruptcy discharge, as well as the ability to restructure their debts in certain circumstances. The Supreme Court of the United States has long stated that “[t]he principal purpose of the Bankruptcy Code is to grant a ‘fresh start’ to the ‘honest but unfortunate debtor.’”18

While state law focuses on individual action by a particular creditor and puts a premium on quick action by that creditor (e.g., the first creditor to execute on the property wins), bankruptcy, on the other hand, compels more of a collective creditor action and emphasizes equality of treatment, rather than a sprint to the courthouse and to the debtor’s assets.19 Indeed, after a debtor files for bankruptcy, a creditor cannot improve its position vis-à-vis other creditors by seizing assets of the debtor or taking further action against the debtor or the property of the estate to

13 Chapters 11 and 13 deal generally with debtor rehabilitation or reorganization, not liquidation of the debtor’s assets. DAVID G. EPSTEIN, BANKRUPTCY AND RELATED LAW IN A NUTSHELL 21 (8th ed. 2013). Typically, in a Chapter 11 or 13 case creditors look to future earnings of the debtor, not the property of the debtor at the time of the bankruptcy petition, to satisfy their claims. The debtor usually retains its assets and makes payments to creditors pursuant to a court-approved plan. Chapter 11, like 7, is available to all forms of debtors—individuals, partnerships, and corporations. Chapter 13, on the other hand, can only be used by individuals with regular income who have unsecured, noncontingent, and liquidated debts of less than $360,475 (adjusted periodically) and secured debts of less than $1,081,400 (adjusted periodically). 11 U.S.C. § 109(e); id. § 104 note (Supp. 2011) (adjustment of dollar amounts).

14 Chapter 12 was added in 1986 because of what was perceived to be a major economic crisis in the farming community, particularly with respect to farms that had been held in the family over a long period of time. It was difficult or impossible for the farmer to use Chapter 11 because the absolute priority rule would require the farmer to give up all equity in the property to the mortgagee when loan payments went into default. J. David Aiken, Chapter 12 Family Farmer Bankruptcy, 66 NEB. L. REV. 632, 632 (1987).


16 The focus of this Article is on Chapter 7 bankruptcy. Turnover actions under § 542(a) of the Bankruptcy Code are predominantly commenced by trustees in Chapter 7 bankruptcy, and nearly every court to address the floating check controversy has done so in Chapter 7 cases. In Chapter 11 and 13 cases, the debtor is typically allowed to keep her property through a reorganization. Accordingly, the issue presented by this Article typically does not arise in Chapter 11 and 13 cases.

17 1 COLLIER ON BANKRUPTCY ¶ 1.01[1], at 1-4 (Alan N. Resnick & Henry J. Sommer eds., 16th ed. 2012).


19 See EPSTEIN, supra note 13, at 44.
collect its claims. To that end, the filing of a bankruptcy petition automatically stays, that is, it “restrains, creditors from taking further action against the debtor, the property of the debtor, or the property of the estate to collect their claims.”

The automatic stay is one of the most important protections in the Bankruptcy Code. Obtaining the protections of the automatic stay is often the primary reason for filing a bankruptcy petition. Congress, in enacting the Bankruptcy Code, was quite clear as to the purpose of the stay:

The automatic stay is one of the fundamental debtor protections provided by the bankruptcy laws. It gives the debtor a breathing spell from his creditors. It stops all collection efforts, all harassment, and all foreclosure actions. It permits the debtor to attempt a repayment or reorganization plan, or simply to be relieved of the financial pressures that drove him into bankruptcy.

The stay is also fundamental to other policies underlying the Bankruptcy Code: equal distribution to creditors of equal priority and orderly administration of the estate. It ends the state law policy of a “race to the court house” or “first come, first serve” and replaces it with the Bankruptcy Code policy of equal treatment to similarly situated creditors.

---

20 Id.
21 Id.
22 The automatic stay is so named because it becomes effective automatically upon the filing of a bankruptcy petition. The debtor does not have to do anything to make it effective. The creditor has the burden of moving to get the stay lifted. It is no excuse that the creditor did not have notice of the filing before the action was taken. The action will still be set aside as void. See, e.g., Ellis v. Consol. Diesel Elec. Corp., 894 F.2d 371, 372–73 (10th Cir. 1990) (holding summary judgment void where entered prior to lifting automatic stay).
23 H.R. REP. NO. 95-595, at 340 (1977); see also S. REP. NO. 95-989, at 49–51 (1978) (describing how the automatic stay “also provides creditor protection” by “provid[ing] an orderly liquidation procedure under which all creditors are treated equally,” and further defining automatic stay).
24 See, e.g., In re Curtis, 40 B.R. 795, 798–99 (Bankr. D. Utah 1984) (“The stay insures that the debtor’s affairs will be centralized, initially, in a single forum in order to prevent conflicting judgments from different courts and in order to harmonize all of the creditors’ interests with one another.” (internal quotation marks omitted)).
25 Generally, the automatic stay remains in effect until the particular property is no longer property of the bankruptcy estate, the court enters an order granting relief from the stay, the case is closed or dismissed, or a discharge is granted or denied. 11 U.S.C. § 362(c)–(d) (2006). Lenders have attempted to circumvent the automatic stay by including provisions within loan documents that protect them from the automatic stay. Courts have generally been loath to accept such provisions. See, e.g., In re Ames Dep’t Stores, Inc., 115 B.R. 34, 38 (Bankr. S.D.N.Y. 1990). Section 362 sets forth the categories of actions that are stayed by the filing of a bankruptcy petition. 11 U.S.C. § 362(a).
The automatic stay is particularly significant in Chapter 7 bankruptcy cases and in relation to the floating check controversy. Indeed, if a debtor files a Chapter 7 petition, the appointed Chapter 7 trustee needs time to collect the property of the estate (e.g., funds in the debtor’s bank account) and make pro-rata distributions to creditors. Continued creditor actions against the debtor and property of the estate would inevitably interfere with the trustee’s orderly bankruptcy administration.

### B. Chapter 7 Bankruptcy

Chapter 7 of the Bankruptcy Code is entitled “Liquidation” or “Straight Bankruptcy” and is the most common chapter used by debtors. Its purpose is to provide debtors with a fresh start. In a Chapter 7 case, a trustee is appointed and her duty is to collect the nonexempt property of the debtor, convert that property to cash, and distribute that cash to creditors in accordance with the distribution scheme of the Bankruptcy Code.26 Essentially, the debtor gives up all nonexempt property she owns at the time of the filing of the bankruptcy petition in exchange for a discharge of all her debts.27

In a very simplified overview, the basic stages of a Chapter 7 case are: (1) the Chapter 7 petition is filed, (2) the filing of the petition results in the stay of creditor collection activity28 and the appointment of a trustee to administer the case,29 (3) the debtor exercises exemption rights with respect to her property,30 (4) the trustee collects and sells or liquidates any property available for distribution to the creditors,31 (5) the proceeds from the sale of the property are used to pay administration expenses and the claims of the creditors according to the Bankruptcy Code’s priority scheme,32 and (6) the debtor may be discharged from any remaining prepetition debts33 that are not exempted from discharge.34

---

27 The right to a discharge is not absolute, and some types of debt are not discharged. For example, if the debtor has committed some bad act enumerated in § 727 of the Bankruptcy Code, the debtor may not be entitled to any discharge. 11 U.S.C. § 727(a). Moreover, certain enumerated debts set forth in § 523(a) are not dischargeable. Id. § 523(a). Additionally, bankruptcy discharge does not extinguish consensual liens on the debtor’s property. Young v. Wells Fargo Bank, N.A. (In re Young), No. 04-32102, 2007 WL 1159952, at *2 (Bankr. S.D. Tex. Apr. 16, 2007).
29 Id. § 701(a).
30 Id. § 522(b).
31 Id. § 704(a).
32 Id. § 726.
33 Id. § 727.
34 Id. § 523(a).
1. The Chapter 7 Petition

Debtors voluntarily file the majority of Chapter 7 cases. The process for filing is fairly simple. The debtor files the petition, statement of financial affairs, and schedule of assets and liabilities, all upon prescribed forms. Consumer debtors, but not other debtors selecting Chapter 7, are required to pass an income threshold referred to as the “means test.” The means test is a formula designed to keep bankruptcy filers with higher incomes—or those that could feasibly fund a Chapter 11 or 13 plan—from filing for Chapter 7.

The official forms and schedules list information required of the debtor about present and past financial condition. “Much additional information is required, some certified under oath by the debtor.” These include, among other things:

- Identity documents;
- Prepetition financial instruction by an approved service;
- Any budget payment plan worked out;
- Copies of “payment advices” such as paycheck stubs for the preceding 60 days;
- Itemized net monthly income;
- 12-month projection of income or expenses reasonably expected;
- Intentions for redemption or reaffirmation as to collateral held under a purchase money security agreement;
- Current tax returns; and
- Interest in any IRA.

A Chapter 7 debtor is also required to turn over bank records showing the debtor’s account balance on the day of the bankruptcy petition. Indeed, the debtor must “surrender to the trustee all property of the estate and any recorded...”
information, including books, documents, records, and papers, relating to property of the estate . . . .”

Failure by the debtor to comply with these debtor duties and supply this information within the appropriate time limits may result in dismissal of the bankruptcy case.

2. The Chapter 7 Trustee and His Duties

In every Chapter 7 case, a trustee is appointed. The Chapter 7 trustee is an impartial person representing the collective interest of the creditors. A Chapter 7 trustee is active, controlling and administering the bankruptcy estate as “the representative of the estate.” The trustee is essentially the successor of the debtor and is entitled to collect and reduce to money the property of the estate. The trustee also represents the creditors. In that capacity, he may exercise rights that would have belonged to creditors before the beginning of the bankruptcy case. Some of the trustee’s responsibilities include, among others, the duty to:

1. Collect and reduce to money the property of the estate and attempt to close the estate as soon as possible;
2. Account for all property of the estate;
3. Investigate the financial affairs of the debtor;
4. If necessary, examine proofs of claim and object to the allowance of claims;
5. If advisable, oppose the discharge of the debtor;
6. Furnish information concerning the estate and the administration of the estate;
7. File periodic operating reports of the operation of the debtor’s business; and
8. Submit a final report and file with the court a final accounting of the administration of the estate.

Another duty of the trustee is to conduct the first meeting of creditors under § 341 of the United States Code, often referred to as the “§ 341 meeting.” The debtor is required to appear at the § 341 meeting and submit to examination under oath. The scope of this examination is very broad. Indeed, “[t]he purpose of the examination is, generally, to ascertain the status of the debtor’s assets and liabilities; to determine whether any avoidable or improper transfers have occurred

---

41 Id.
42 Id. § 707.
43 Id. § 323.
44 Id. § 704.
45 Id. § 343.
or whether assets are being concealed, or whether there are grounds for objection to exemptions or discharge."46

With the opening of the bankruptcy estate, the trustee should begin to promptly assemble the property belonging to the estate. This may include taking possession of the debtor’s known bank accounts and books and records, seizing vehicles and other property, changing locks at the debtor’s premises, and reviewing the debtor’s schedules for nonexempt assets and potential avoidance claims. In some cases, the trustee may need to collect property from the hands of third persons, or perhaps noncompliant debtors in possession of estate property. One of the most basic collection methods is the trustee’s turnover power under § 542(a) of the Bankruptcy Code.

III. SECTION 542(A) TURNOVER ACTIONS AND CASE LAW ADDRESSING THE FLOATING CHECK CONTROVERSY

To understand the floating check issue, it is important to understand the inception of turnover actions. Prior to the enactment of the Bankruptcy Code, turnover procedures were not prescribed by statute, but rather were considered “judicial innovation[s]” derived from the notion that the courts were empowered under the Bankruptcy Act (which governed from 189847 to 1978) to “cause the estates of the bankrupts to be collected.”48 But the jurisdiction of the bankruptcy court’s ability to recover property of the estate was limited. Indeed, “a court’s authority to compel turnover under the former Act was limited to only its own summary jurisdiction over the estate’s actual property.”49 The enactment of the Bankruptcy Code, however, expanded the court’s authority to compel turnover and recover property of the estate.50

In 1978, with the enactment of the Bankruptcy Code, a substantive cause of action for turnover of property of the estate was created by § 542. This section allows a trustee to recover property of the estate in possession, custody, or control of any entity. The statute provides, in relevant part, as follows: “[A]n entity . . . in possession, custody, or control, during the case, of property [of the estate] . . . shall deliver to the trustee . . . such property or the value of such property . . . .”51

The purpose of this section is “to expand the trustee’s power to ‘bring into the estate property . . . at the time the bankruptcy proceedings commenced,’” ensuring that a broad range of property is included in the estate in order to promote the congressional goal of encouraging reorganizations.”52 This turnover power even “reaches property in the hands of secured creditors.”53 Indeed, a turnover action by a trustee “invokes the court’s most basic equitable powers to gather and manage property of the estate.”54

A turnover proceeding may be brought as an adversary proceeding,55 by service of complaint,56 or by motion as a contested-matter proceeding if it is against the debtor.57 Although the Bankruptcy Code contains no express time limitation for the commencement of a turnover proceeding, such an action may nevertheless be time barred.58 An order granting or denying a trustee’s turnover

---

53 5 COLLIER ON BANKRUPTCY, supra note 17, ¶ 542.01, at 542-3.
54 Braunstein, 571 F.3d at 122.
55 Rule 7001(1) of the Federal Rules of Bankruptcy Procedure includes in the list of adversary proceedings “a proceeding to recover money or property, other than a proceeding to compel the debtor to deliver property to the trustee.” FED. R. BANKR. P. 7001(1). Thus, a trustee is entitled to proceed against the debtor via motion since the rule states that a proceeding to recover money or property is an adversary proceeding unless it is a proceeding to compel the debtor to turn over property of the estate to the trustee.
56 See id. 7003 (“Rule 3 F. R. Civ. P. applies in adversary proceedings.”).
58 See, e.g., In re Fleming, 424 B.R. 795, 805 (Bankr. W.D. Mich. 2010) (denying trustee from compelling turnover of tax refund under § 542(a) because trustee waited three years before seeking to compel). As the Fleming court observed:

Bankruptcy courts are frequently described as courts of equity and laches is certainly an equitable defense. But § 105(a) itself empowers bankruptcy courts
request is a final order within the meaning of section 158, title 28 of the United States Code; therefore, the district court has jurisdiction to hear any appeal from an order granting or denying such a request. If a party, including a debtor, fails to comply with a turnover order, that party is guilty of civil contempt and may be punished. Furthermore, a debtor’s failure to turn over property and comply with the trustee’s demand can also result in the revocation of her discharge.

In analyzing a motion for turnover pursuant to § 542(a), “[t]he first question is what precisely constitute[s] property of the bankruptcy estate on the date of to “issue any order, process, or judgment that is necessary or appropriate to carry out the provisions of this title.” Moreover, the Supreme Court and the Sixth Circuit have held that the Bankruptcy Code takes precedent over equity in determining the bankruptcy court’s authority to act . . . . Therefore, while laches may seem like a reasonable approach, the court concludes that it is better to address Trustee’s delay with the tools already provided by Section 105(a).

Id. (citations omitted).

See, e.g., Prof’l Ins. Mgmt. v. Ohio Cas. Grp. of Ins. Cos. (In re Prof’l Ins. Mgmt.), 285 F.3d 268, 281 (3d Cir. 2002) (stating that a turnover order “is widely regarded as a final order for purposes of appeal”).

See, e.g., Maggio v. Zeitz, 333 U.S. 56, 59 (1948) (discussing contempt for failure to comply with a turnover order); In re Shore, 193 B.R. 598, 603 (Bankr. S.D. Fla. 1996) (affirming contempt order for failure to comply with turnover order). The bankruptcy court “may issue any order, process, or judgment that is necessary or appropriate to carry out the provisions of [Title 11].” 11 U.S.C. § 105(a) (2006). Section 105(a) “authorizes bankruptcy courts to impose sanctions for civil contempt to compel compliance with a court order or compensate parties for losses caused by noncompliance.” Scrivner v. Mashburn (In re Scrivner), 535 F.3d 1258, 1265 n.3 (10th Cir. 2008) (citing Mountain Am. Credit Union v. Skinner (In re Skinner), 917 F.2d 444, 447 (10th Cir. 1990)). “The standard for finding a party in civil contempt is well settled: The moving party has the burden of showing by clear and convincing evidence that the contemnors violated a specific and definite order of the court.” Knupfer v. Lindblade (In re Dyer), 322 F.3d 1178, 1190–91 (9th Cir. 2003) (citation omitted). In Hansbrough v. Birdsell (In re Hercules Enters., Inc.), 387 F.3d 1024 (9th Cir. 2004), the court awarded civil sanctions, including attorneys’ fees, when the debtor’s principal failed to turn over property of the estate. See id. at 1026–27. The defendant in Hercules operated a gymnasium with several pieces of exercise equipment. Id. at 1026. When the defendant removed the equipment from the gym, the trustee moved the court to compel its return. Id. The court granted the motion and ordered the defendant to turn over the equipment to the trustee. Id. When the defendant still failed to turn over the equipment, the court ordered that he be incarcerated as a coercive sanction. Id. at 1028. The court further ordered the defendant to pay to the trustee “an amount approximating the fees and costs incurred by him as the result of [the defendant’s] misconduct.” Id. at 1027.

filing.”62 If the property subject to turnover is not property of the debtor’s bankruptcy estate, the trustee’s § 542(a) turnover action will undoubtedly fail.63

A. Property of the Estate

“Section 541 embodies the essence of the Bankruptcy Code. It creates the bankruptcy estate, which consists of all of the property that will be subject to the jurisdiction of the bankruptcy court.”64 “It is from this central core of estate property that the creditors will be paid.”65 Congress’s intent was to make the bankruptcy estate as “inclusive as possible”66 and to construe it generously.67 Indeed, the property of the estate includes, but is not limited to, “‘all legal or equitable interests of the debtor in property as of the commencement of the [bankruptcy] case,’ wherever located and by whomever held.”68 “It would be hard to imagine language that would be more encompassing.”69 Section 541(b) enumerates those items that are specifically excluded from being property of the estate. These exclusions are narrow.70

“Even though section 541 provides the framework for determining the scope of the debtor’s estate and what property will be included in the estate, it does not

---

63 In seeking the entry of a turnover order, the burden is on the trustee to show that the property or proceeds are part of the bankruptcy estate. See Maggio, 333 U.S. at 64; see also Rish Equip. Co. v. Joe Necessary & Son, Inc. (In re Joe Necessary & Son, Inc.), 475 F. Supp. 610, 613 (Bankr. W.D. Va. 1979). The general rule is that “[t]he trustee succeeds only to such rights as the bankrupt possessed; and the trustee is subject to all claims and defenses which might have been asserted against the bankrupt but for the filing of the petition.” Bank of Marin v. England, 385 U.S. 99, 101 (1966).
64 5 COLLIER ON BANKRUPTCY, supra note 17, ¶ 541.01, at 541-10.
65 Id.
66 Id.
67 See Baer v. Jones (In re Montgomery), 224 F.3d 1193, 1194 (10th Cir. 2000).
69 5 COLLIER ON BANKRUPTCY, supra note 17, ¶ 541.01, at 541-10.
70 Id. ¶ 541.01, at 541-10 to -11 (“Enumerated exclusions consist of powers that the debtor may exercise solely for another’s benefit, any interest of the debtor as lessee under a lease of nonresidential real property after the expiration of the lease term, the debtor’s eligibility to participate in Higher Education Act programs or accreditation or licensure status as an educational institution, certain interests in liquid or gaseous hydrocarbons, funds placed in a Coverdell Education Savings Account, the debtor’s interest in funds used to purchase certain tuition benefits, the debtor’s interest in amounts withheld from wages or contributed by the debtor to certain employee benefit or deferred compensation plans, the debtor’s interest in property subject to possessory pledges to certain licensed lenders, such as pawnbrokers, and certain cash or equivalent proceeds from the sale of a money order made within fourteen days prior to the commencement of the case if the proceeds are required to be segregated from the debtor’s other property.”).
provide any rules for determining whether the debtor has an interest in property in the first place.\textsuperscript{71} The Supreme Court of the United States has stated,

> Congress has generally left the determination of property rights in the assets of the bankruptcy’s estate to state law. . . . Unless some federal interest requires a different result, there is no reason why such interests should be analyzed differently simply because an interested party is involved in a bankruptcy proceeding.\textsuperscript{72}

Section 541(a) also includes any property recovered by the trustee using the turnover powers conferred by § 542. Indeed, Congress intended to include a broad range of “property in the estate,” subject to turnover.\textsuperscript{73} The United States Supreme Court has stated: “Although these statutes [(§§ 542(a), 363(a) and (b), and 541(a)(1))] could be read to limit the estate to those ‘interests of the debtor in property’ at the time of the filing of the petition, we view them as a definition of what is included in the estate, rather than as a limitation.”\textsuperscript{74}

With respect to funds in a debtor’s bank account on the petition date, nearly all courts that have addressed the floating check issue—even those siding with the debtor—agree that if the funds are present in the debtor’s account, such funds are property of the debtor’s bankruptcy estate.\textsuperscript{75} A small minority of courts, however, hold that the funds themselves are not property of the estate, but rather it is the debtor’s right to collect the funds from the bank that constitutes property of the estate.\textsuperscript{76} As stated by one court, the debtor

> did not own the money in the account, but was merely a creditor of Merrill Lynch . . . . Property of the debtor is defined to include “all legal or equitable interests of the debtor,” and obviously includes the interest that a depositor has in the money in his account, more precisely the money owed him by the bank by virtue of the account.\textsuperscript{77}

\textsuperscript{71} Id. ¶ 541.03, at 541-15.
\textsuperscript{74} Id. at 203.
\textsuperscript{75} See cases cited infra note 95.
\textsuperscript{76} E.g., In re Ruiz, 440 B.R. 197, 201 (Bankr. D. Utah 2010) (“Zions Bank was in possession, custody and control of the funds on deposit and the funds belonged to Zions Bank on the petition date. The Debtors were in possession, custody, and control of a promise to pay from the bank to the Debtors.”), rev’d, 455 B.R. 745 (B.A.P. 10th Cir. 2011).
This minority view is flawed because it attempts to narrow the scope of § 541 (i.e., property of the estate). In at least one decision addressing this issue, the court’s narrow holding that the estate’s interest in the checking account amounted to nothing more than a beneficial interest in the bank’s promise to pay the funds held in the account (as opposed to the money in the account), was derived from language contained in the United States Supreme Court’s opinion in Citizens Bank of Maryland v. Strumpf.

In Strumpf, a debtor had both a checking account and a delinquent loan with a creditor bank on the petition date. When the debtor filed the bankruptcy petition, the bank placed an administrative hold on that part of the funds contained in the checking account required to offset any prepetition debt that the debtor owed the bank on the loan. The debtor brought an action against the bank, alleging that the bank had violated the automatic stay because the administrative hold was actually an improper setoff of the debtor’s funds in violation of § 362(a)(7) of the Bankruptcy Code. The Supreme Court held that the bank’s actions did not constitute a setoff, and thus it had not violated the automatic stay. Following that holding, the Supreme Court also briefly dismissed the debtor’s contentions that the bank had violated the automatic stay, noting that the bank did not actually take possession of any of the debtor’s property or exercise control over the debtor’s property. Rather, the Court held that the bank merely failed to perform its promise to pay the debtor the funds held in the account.

From the Strumpf holding, the United States Bankruptcy Court for the District of Utah in In re Ruiz took a very narrow view, holding:

[Strumpf] makes clear that funds held on deposit in a debtor’s bank account do not belong to the debtor and do not belong to the bankruptcy estate. Funds held by a bank consists of nothing more or less than a promise to pay, from the bank to the depositor. [Debtors’ bank] was in possession, custody and control of the funds on deposit and the funds belonged to [Debtors’ bank]. The Debtors were in possession, custody, and control of a promise to pay from the bank to the Debtors.

---

78 In re Ruiz, 440 B.R. at 197.
80 Id. at 17.
81 Id. at 17–18.
82 The automatic stay prevents “the setoff of any debt owing to the debtor that arose before the commencement of the case under this title against any claim against the debtor.” 11 U.S.C. § 362(a)(7).
83 Strumpf, 516 U.S. at 17–18.
84 Id. at 19–20.
85 Id.
86 Id. at 21 (noting that a bank account “consists of nothing more or less than a promise to pay, from the bank to the depositor”).
87 In re Ruiz, 440 B.R. 197, 201 (Bankr. D. Utah 2010), rev’d, 455 B.R. 745 (B.A.P. 10th Cir. 2011) (citation omitted).
In rejecting this holding (and disagreeing with other similar holdings), the Bankruptcy Appellate Panel (BAP) of the Tenth Circuit stated that although the language contained in *Strumpf* does “facially support Debtors’ position, the context of that case is entirely different from the case currently before this Court.”\(^{88}\) The Tenth Circuit BAP’s reason for rejecting the *Strumpf* holding in the context of the floating check issue was that *Strumpf* “solely involved the automatic stay and the relationship between the bank and the debtor. The issue of what constituted property of the estate under § 541 was neither argued nor decided.”\(^{89}\) For that reason, the Tenth Circuit BAP stated that the “language in *Strumpf* is not dispositive under the facts, or the [floating check] issue presented, in this case.”\(^{90}\)

The Tenth Circuit BAP further held that the relationship between the bank and the debtors was “considerably different than the typical debtor-creditor relationship that existed in *Strumpf*.”\(^{91}\) Indeed, the “[d]ebtors maintained the right to withdraw the funds in their account at any time, to direct [their bank] to deliver the funds to any third party, or to leave the funds on deposit.”\(^{92}\) The court further concluded that “although [the bank] did make a promise to pay the funds in the account to the debtors, the checking account constituted much more than that promise and Debtors’ rights to those funds exceeded those of a typical creditor.”\(^{93}\)

Finally, the Tenth Circuit BAP recognized that the scope of § 541 (i.e., property of the estate) is broad “and should be generously construed” and that “the bankruptcy court’s attempts to narrow the scope of § 541 in relation to funds on deposit in a checking account does not satisfy the standard the Court must apply when considering § 541.”\(^{94}\) With that, the Tenth Circuit BAP adopted the “prevailing view of nearly every court to consider this issue by holding that the funds in [a Debtor’s bank] account, rather than merely the promise to pay over those funds, constitute[] property of the bankruptcy estate.”\(^{95}\)

---

\(^{88}\) *In re Ruiz*, 455 B.R. 745, 749 (B.A.P. 10th Cir. 2011).

\(^{89}\) *Id.*

\(^{90}\) *Id.*

\(^{91}\) *Id.*

\(^{92}\) *Id.*

\(^{93}\) *Id.*

\(^{94}\) *Id.* (quoting Williamson v. Jones (*In re Montgomery*), 224 F.3d 1193, 1194 (10th Cir. 2000)).

\(^{95}\) *Id.* (emphasis added); *see also* Brown v. Pyatt (*In re Pyatt*), 486 F.3d 423, 427 (8th Cir. 2007) (“[T]he funds transferred by the [prepetition] checks are property of the estate.”); Yoon v. Minter-Higgins, 399 B.R. 34, 42–44 (N.D. Ind. 2008) (“Funds on deposit, in a debtor’s checking account on the petition date that are not otherwise exempt are property of the bankruptcy estate.”); Maurer v. Hedback (*In re Maurer*), 140 B.R. 744, 746 (D. Minn. 1992) (holding that prepetition checks remain property of the estate until honored); Shapiro v. Henson (*In re Henson*), 449 B.R. 109, 112 (Bankr. D. Nev. 2011) (“The bankruptcy court held that the checks written pre-petition by Debtor became property of the estate because they had not been honored when Debtor filed for bankruptcy.”); *In re Brubaker*, 426 B.R. 902, 905 (Bankr. M.D. Fla. 2010) (“[B]oth schools [of thought] agree that the funds are property of the estate.”), aff’d, 443 B.R. 176 (M.D.
While the issue of what constitutes “property of the estate” is rather clear-cut, the second element of a § 542(a) turnover claim—“possession, custody, or control” of that property—presents a thornier issue and one that has been the subject of much debate.

B. Possession, Custody, or Control of Property of the Estate

After a court determines that the subject property is “property of the estate,” the next question in analyzing a trustee’s turnover request is whether the party subject to the turnover action had “possession, custody, or control” of that property “during the case,” such that she can be required to turn over the property, or its value, to the trustee pursuant to § 542(a). 96 “While it is true that the phrase ‘possession, custody, or control’ is not defined in the Code, reading the phrase in the context with the language of § 541 helps clarify its meaning and purpose.” 97 Section 541 provides, in relevant part, that “[t]he commencement of a case . . . creates an estate. Such estate is comprised of all the following property, wherever located and by whomever held: (1) . . . all legal or equitable interests of the debtor in property as of the commencement of the case.” 98 “That language . . . indicates the very broad scope of a bankruptcy estate.” 99 Indeed, “Congress drafted the sweeping language of §§ 541 and 542 deliberately to grant broad powers to an estate and its trustee.” 100

99 Carlton, 196 B.R. at 808.
100 Id. at 808–09.
A handful of courts have held that funds in the debtor’s account as of the bankruptcy filing, which were dissipated postpetition, are not subject to turnover by the debtor pursuant to § 542(a), even though the funds are property of the debtor’s bankruptcy estate. These courts hold that an entity cannot be compelled to turn over property that is no longer within its “possession, custody, or control.”

The leading case on this issue is Brown v. Pyatt (In re Pyatt). There, when the debtor filed for Chapter 7 relief, there was $1,938.76 in his checking account. Postpetition, and before the trustee was appointed and could even make a turnover demand, the debtor subtracted the amount of several prepetition checks written to creditors when he reported the balance of his checking account apparently leaving all but $300 of the $1,938.76 in the debtor’s account. The debtor’s bankruptcy schedules reflected this asset in the sum of $300, which the debtor believed would be the amount left in his account after all the checks cleared. At the § 341 meeting, however, the trustee learned of the true amount still in the account as of the petition date, not considering the checks that were honored subsequent to filing, and demanded turnover pursuant to § 542(a). The bankruptcy court concluded that because the assets represented by the checks were still in the debtor’s account as of the date of his filing, the trustee was allowed to compel turnover under § 542(a).

The debtor appealed to the Eighth Circuit BAP, which reversed. The panel majority concluded that the trustee was in a better position to recover the funds paid out by a bank to third parties after the debtor’s filing because the trustee was authorized to avoid postpetition transfers pursuant to § 549 of the Bankruptcy Code. The panel concluded that “[i]f the trustee were to recover the transferred funds, it is the only forum in which the estate could be restored to the balance as of the day of the filing.”

---


101 Pyatt, 486 F.3d at 425.

102 Id. at 426.

103 Id.

104 Id.

105 Id.

106 Id.

107 Id.

108 Id. Section 549 of the Bankruptcy Code allows a trustee to avoid unauthorized postpetition transfers, subject to certain exceptions. 11 U.S.C. § 549 (2006). The purpose of § 549 is to allow the trustee to avoid specified postpetition transfers that have depleted the bankruptcy estate. 5 COLLIER ON BANKRUPTC Y, supra note 17, ¶ 549.01, at 549-3.

Examples of postpetition transfers not authorized by the Bankruptcy Code or the bankruptcy court that are recoverable by the trustee are payments to prepetition creditors, including payments of penalties to a state agency, the placement of judgment liens upon the debtor’s property, a deduction made from a debtor’s wages to pay a credit union debt, a payment of a postpetition loan and a bank’s setoff of funds to apply to prepetition indebtedness.
funds, the claims paid by the checks could be reinstated and the recovered funds could be distributed equally among all creditors.”109 “The concurring opinion disagreed that the trustee [was] in a better position to collect” the funds, but argued that “section 542(a) does not authorize any procedure used by the [trustee] because the debtor no longer had control of the funds” and, therefore, could not be required to turn them over.110

The Eighth Circuit affirmed, noting that § 542(a) of the Bankruptcy Code requires any entity that “is in possession, custody, or control, during the case,” of property of the estate to turn it over to the trustee.111 The Eighth Circuit in Pyatt did not dispute that the funds transferred by the checks were property of the estate.112 However, the Eighth Circuit concluded that the words “during the case” fails to acknowledge the other language of § 542(a).113 While § 542(a) imposes an obligation on any entity who comes into “possession, custody, or control” of property of the estate after the bankruptcy petition is filed to deliver it to the trustee, it says nothing about whether that obligation continues after possession, custody or control cease, and does not specify whether an entity that lacks control may properly be subject to a motion to compel turnover.114 The Pyatt court also relied on pre-Code practice and the 1948 Supreme Court decision in Maggio v. Zeitz,115 which held that possession or control of the property by a party at the time of the turnover proceeding is required to compel turnover.116 Several courts have followed, and continue to follow, the Pyatt holding.117 Pyatt’s reasoning, however,
is flawed. The enactment of the Bankruptcy Code altered the pre-Code “current possession or control” requirement because § 542(a) expressly permits a trustee to recover “the value” of the property, in addition to the property itself, from one who possessed the property at any time “during the case.”

IV. THE TRUSTEE’S ABILITY TO RECOVER ACCOUNT FUNDS PAID POSTPETITION

Notwithstanding all that has been decided, it remains unsettled whether the trustee can obtain the transferred funds through a motion for turnover against the debtor, or whether he must seek postpetition payments from the payees of the checks. Nearly all bankruptcy court decisions addressing the floating check controversy agree that if the funds are in the debtor’s bank account on the petition date, such funds are property of the estate. As the concurrence in Pyatt pointed out, “there is no doubt that the funds on deposit in the debtor’s account when the case was filed was property of the estate. This is true whether one thinks of that property as cash, a credit of some sort, or a debt owed by the bank to the debtor.” However, the second requirement of a § 542(a) claim (i.e., “possession, custody, or control during the [bankruptcy] case”) is not clear-cut and the courts that have considered this matter have arrived at completely opposite conclusions.

to the time that a turnover proceeding is instituted. In this case, Debtor does not possess the property or the proceeds of the property.” (citations omitted); In re Anderson, 410 B.R. 289 (Bankr. W.D. Mo. 2009) (“The Trustee concedes that under the Eighth Circuit’s decision in Pyatt, a debtor is not required to turn over money represented by checks which had been written prior to the bankruptcy filing, but had not yet cleared until after the bankruptcy case was filed.” (citations omitted)).

118 Pyatt v. Brown (In re Pyatt), 348 B.R. 783, 787 (8th Cir. B.A.P. 2006) (Kressel, C.J., concurring), aff’d, 486 F.3d 423 (8th Cir. 2007). Since the controversy surrounding the floating check issue does not necessarily involve whether the funds in question are “property of the estate,” this Article focuses on the dispute with respect to the second element of a § 542(a) turnover action: “possession, custody, or control” during the bankruptcy case. 11 U.S.C. § 542 (2006). It is important to point out, however, that while the majority of courts that have considered the floating check issue have analyzed “property of the estate” and “possession, custody, or control” separately, this is not necessary and, perhaps, is a flawed method. When most bankruptcy courts analyze the issue of “property of the estate” in other contexts, they focus heavily on the concept of dominion and control. Indeed, the Tenth Circuit has held that dominion and control over property are the hallmarks for determining whether property is property of a bankruptcy estate. See Parks v. FIA Card Servs., N.A. (In re Marshall), 550 F.3d 1251, 1255 (10th Cir. 2008) (“[A] transfer of property will be a transfer of ‘an interest of the debtor in property’ if the debtor exercised dominion or control over the transferred property.” (citation omitted)); In re Amdura Corp., 75 F.3d 1447, 1451 (10th Cir. 1996); Amdura Nat’l Distrib. v. Amdura Corp., (In re Paige), 413 B.R. 882, 909 (Bankr. D. Utah 2009) (“The Tenth Circuit has interpreted [section 541] to mean that property that is titled in the name of the debtor and that is under the debtor’s ‘dominion or control’ is presumptively property of the estate.”). Accordingly, one might argue that the courts are putting the cart before the horse, so to speak, since a determination of what constitutes “property of the estate” can oftentimes be made by first analyzing the issue of dominion and control.
Some hold that the trustee bears the burden to seek the postpetition payments from the payees of the checks through avoidance actions, while others put the onus on the debtor to turn over the funds, even if those funds have been transferred from the estate to payees.

As set forth below, § 542(a) not only entitles the trustee to recover from the debtor, but it may also be the best method of maximizing the estate and providing the greatest return to creditors.

A. It Is Not Until the Checks Have Been Presented for Payment to the Debtor’s Bank that the Debtor Ceases to Have Possession, Custody, or Control of the Funds Represented by the Checks

The phrase “possession, custody, or control” is not defined in the Bankruptcy Code. When construing a term (where not defined in a statute), it is proper for a court to look to its commonly approved usage, an inquiry that is enhanced by the examination of dictionary definitions. Black’s Law Dictionary defines possession as “having or holding property in one’s power; the exercise of dominion over property . . . . Something that a person owns or controls.”119 Custody is defined as “[t]he care and control of a thing or person for inspection, preservation, or security.”120 And control is defined as “the power or authority to manage, direct, or oversee.”121 The phrase “possession, custody, or control” is disjunctive and only one of the enumerated requirements need be met. Thus, “actual possession” is not required.122

In the case of transferred funds from a bank account, given that a check is a negotiable instrument, the most relevant provisions governing the status of funds represented by an issued check, specifically as to who is in “possession, custody, or control,” are those found in the U.C.C. Nearly all states have adopted the U.C.C. by statute.123

---

119 BLACK’S LAW DICTIONARY 1281 (9th ed. 2009).
120 Id. at 441.
121 Id. at 378.
122 See, e.g., Carlton, Fields, Ward, Emmanuel, Smith & Cutler, P.A. v. Boyer (In re U.S.A. Diversified Prods., Inc.), 196 B.R. 801, 809 (N.D. Ind. 1996) (noting that Congress wrote the phrase “possession, custody or control” in the disjunctive rather than the conjunctive); id. (“Nothing in the phrase indicates that an entity must have any kind of ‘dominion’ over the property in question . . . . So, while the law firm may not have had ‘control’ over the money since it could only transfer it pursuant to the directives of its client, that did not prevent the firm from having ‘possession’ or ‘custody,’ which is enough to make it subject to a turnover action under § 542. To interpret the phrase ‘possession, custody or control’ as restrictively as appellant urges would fly in the face of the broad intent and purpose of §§ 541 and 542.”).
Under the U.C.C., the payee of an ordinary, uncertified check only obtains possession and control of those funds represented by the check once the funds are physically present in the payee’s account.\textsuperscript{124} But simply because a check has been issued, does not result in the drawer losing possession and control of the funds.

Article 3 of the U.C.C. applies to negotiable instruments.\textsuperscript{125} Section 3-408 provides that a “check or other draft does not of itself operate as an assignment of funds in the hands of the drawee\textsuperscript{126} available for its payment, and the drawee is not liable on the instrument until the drawee accepts it.”\textsuperscript{127} Thus, upon receipt of a check by the payee, the drawer has not effectively assigned the funds in the drawer’s bank account pending payment of the check.

Moreover, regarding whether the funds remain in the drawer’s account and hence in the drawer’s possession or control up until presentment, § 3-501 defines presentment as:

\begin{quote}
[A] demand made by or on behalf of a person entitled to enforce an instrument: (i) to pay the instrument made to the drawee or a party obliged to pay the instrument or, in the case of a note or accepted draft payable at a bank, to the bank, or (ii) to accept a draft made to the drawee.\textsuperscript{128}
\end{quote}

\textsuperscript{124}See U.C.C. §§ 3-408, 3-310(b)(1) (2002).

\textsuperscript{125}See id. § 3-102(a).

\textsuperscript{126}Id. § 3-104(a).

\textsuperscript{127}Id. § 3-408. “Drawee” means a person ordered in a draft to make payment. Id. § 3-103(a)(4).

\textsuperscript{128}Id. § 3-501(a).
Presentment has been construed as the point in time where the payee’s bank presents the check to the drawer’s bank for payment. 129 Significantly, after the drawer issues the check to the payee, but before presentment of the check to the drawer’s bank, the funds represented by the check undoubtedly remain in the drawer’s account and under the drawer’s control. Indeed, it is not until presentment that issuance of a check constitutes full and absolute payment. 130

Other indications of “possession, custody or control” are found in Article 4 of the U.C.C., which applies to bank deposits and collections. 131 Specifically, section 4-403 deals with the “Customer’s Right to Stop Payment” and provides in part:

A customer or any person authorized to draw on the account if there is more than one person may stop payment of any item drawn on the customer’s account or close the account by an order to the bank describing the item or account with reasonable certainty received at a time and in a manner that affords the bank a reasonable opportunity to act on it before any action by the bank with respect to the item described in section 4-303. If the signature of more than one person is required to draw on an account, any of these persons may stop payment or close the account. 132

Thus, a drawer of a check may issue a stop payment order at any time after the check has been written in an attempt to prevent the check from being paid and the funds from being transferred to the payee. This right to stop payment is clearly indicative of the “possession, custody, or control” a debtor has over funds in a bank account. A court should be hard pressed to deny that a debtor is not in control over funds in her account if she can actually take steps to stop payment of the check. 133

Furthermore, the United States Supreme Court has confirmed that under the Bankruptcy Code, a transfer of a check occurs when the drawee bank honors the

---


130 See HENRY J. BAILEY & RICHARD B. HAGEDORN, BRADY ON BANK CHECKS 4-16, 4-19 (7th ed. 1992) (“Since delivery of an ordinary, uncertified check is only conditional payment, dependent on the check being honored upon presentment, the check vests no title or interest in the payee as to funds on deposit in the drawer’s bank account, and the check is deemed revocable by the drawer until it is paid.” (emphasis added)).

131 See U.C.C. § 4-101 (2010) (“This Article may be cited as Uniform Commercial Code—Bank Deposits and Collections.”)

132 Id. § 4-403.

133 In re Schoonover, No. 05-43662-7, 2006 WL 3093649, at *3 (Bankr. D. Kan. Oct. 30, 2006) (“Debtors also do not, and really cannot, dispute, that they had management of, or control over, the funds deposited in the accounts.”)
check, not when the payee receives it. Accordingly, any time before a bank honors a check, the funds represented by that check are within the “possession, custody, or control” of the drawer (i.e., the debtor in this case).

There are several other factors that demonstrate a debtor’s control over the funds in her account. These include, among other things, the ability to (1) manage the account, (2) determine when funds will be disbursed and to whom, (3) close the bank account, (4) contact the bank and provide notice of the bankruptcy, (5) withdraw funds from the bank, (6) use a debit card to purchase goods, and (7) transfer funds from one account to another. Thus, there can really be no question that if, on the day a debtor files for bankruptcy, the issued checks have not been honored, the debtor has complete “possession, custody, or control” of the funds represented by those checks. But this still leaves the question of whether that “possession, custody, or control,” which may last for only a day or a few hours

134 Barnhill v. Johnson, 503 U.S. 393, 400–01 (1992). The issue of when a transfer is deemed to have occurred is regularly discussed in the context of preferences under title 11, section 547 of the United States Code. A trustee may avoid any transfer of an interest of the debtor in property if the transfer (1) was made to a creditor, (2) was for payment of debt, (3) was made while the debtor was insolvent, (4) was within ninety days before the filing of the bankruptcy (one year if the creditor was an “insider” of the debtor), and (5) allowed to the creditor to recover more than he would recover in a Chapter 7 liquidation. 11 U.S.C. § 547(b) (2006). The date the transfer occurs can provide a significant defense to a creditor when litigating whether the payment was made within ninety days. Under Barnhill, the “date of honor” rule, not of delivery, applies to preferential transfers under § 547(b).

135 See In re Parker, No. 05-17912, 2008 WL 906570, at *4 (Bankr. N.D.N.Y. Apr. 3, 2008) (“The recipient of a check has no right to funds in an account until the check is presented for payment.”).

136 It is important to note that the floating check controversy and the § 542(a) analysis are not questions of what a debtor should do, but what a debtor can do. The issue is not the debtor’s duty; the issue is the debtor’s control of the funds in the bank account. The fact that a debtor can perform all of these tasks, which no courts dispute, demonstrates that a debtor has complete control over funds in her bank account until the moment of honor. It is the debtor’s control that is the key for purposes of § 542(a), not what otherwise might be her duty. It is important to note, however, that pursuant to 11 U.S.C. § 521(a)(4), debtors have an explicit duty to “surrender to the trustee all property of the estate.”
after the bankruptcy case is filed, constitutes “possession, custody, or control, during the case,” which is required by § 542(a).

B. A Debtor’s Control of Her Bank Account Funds on the Date of Filing Requires the Debtor to Deliver to the Trustee Such Funds or the Value of Such Funds

As set forth above, several courts make the argument that funds in the debtor’s account as of the bankruptcy filing, which were dissipated postpetition, are not subject to turnover by the debtor pursuant to § 542(a).137 These courts base their decisions upon flawed interpretations of the Bankruptcy Code and outdated common practice under the former Bankruptcy Act. Regarding the former, these courts conclude that § 542 has to be interpreted as permitting recovery against only a person who is currently in possession or control of the subject property at the time of the turnover action—otherwise, there could be an unintended double recovery. In Pyatt, for example, the Eighth Circuit made the following conclusion:

Here, both the debtor and the debtor’s payees had “possession, custody, or control” of the funds at some point after the bankruptcy petition was filed. Under the trustee’s reading of the provision, the trustee could proceed both against the debtor and against the payees and obtain double satisfaction. The code’s drafters apparently did not think it necessary to prevent the trustee from obtaining double satisfaction under § 542(a). Cf. 11 U.S.C. § 550(d) (prohibiting double satisfaction in avoidances under §§ 544, 545, 547–549, 553(b), and 724(a); no mention of § 542(a)). The absence of such a prohibition suggests that the drafters did not intend to authorize a trustee to proceed under § 542(a) against everyone who may have had control over property of the estate at some point after the petition was filed.138

This argument raised in Pyatt and the meaning of § 550 of the Bankruptcy Code needs further explanation. Pyatt (and other cases) cites to § 550 of the Bankruptcy Code for support that an unintended double recovery might occur if a

---

137 See, e.g., Lovald v. Falzerano (In re Falzerano), 454 B.R. 81, 86 (B.A.P. 8th Cir. 2011) (holding that § 542(a) permits a trustee to compel turnover only from entities that have control of property of the estate or its proceeds at the time of turnover demand); Shapiro v. Henson (In re Henson), 449 B.R. 109, 112–13 (Bankr. D. Nev. 2011) (holding that because debtor no longer had possession of the funds, trustee could not compel turnover); In re Anderson, 410 B.R. 289 (Bankr. W.D. Mo. 2009) (“The Trustee concedes that under the Eighth Circuit’s decision in Pyatt, a debtor is not required to turn over money represented by checks which had been written prior to the bankruptcy filing, but had not yet cleared until after the bankruptcy case was filed.”); In re Taylor, 332 B.R. 609 (Bankr. W.D. Mo. 2005) (holding that the trustee may proceed against the payees of the checks and not the debtor).

138 Pyatt, 486 F.3d at 427–28.
trustee is allowed to pursue a debtor for the value of the funds transferred postpetition. This reasoning is faulty.

Section 550 of the Bankruptcy Code permits a trustee, after avoidance of a transfer under the trustee’s avoiding powers, to recover the property transferred or the value of the property transferred.\textsuperscript{139} Section 550 applies only to transfers avoided against subsequent judicial lien creditors and bona fide purchasers or against actual creditors,\textsuperscript{140} avoided statutory liens,\textsuperscript{141} preferences,\textsuperscript{142} fraudulent transfers and obligations,\textsuperscript{143} postpetition transfers,\textsuperscript{144} setoffs within the ninety-day prepetition period,\textsuperscript{145} and liens secured by penalties.\textsuperscript{146} Under § 550, the trustee may recover the property itself or, on court order, the value of the property. Furthermore, the trustee may recover the property from the initial transferee or a subsequent transferee. But the trustee may not obtain a windfall for the estate by recovering from multiple transferees so that the recovery is in excess of the value of the property transferred.\textsuperscript{147} “Subsection (d) recognizes the possibility that more than one entity may be liable, but that the trustee’s remedy is limited to the recovery of property or its value, and not damages.”\textsuperscript{148} For example, if “the trustee could recover under subsection (a)(1) from either the initial transferee or from the entity for whose benefit the transfer was made, the trustee may recover from both, but only insofar as the recovery does not exceed the value of property or the property itself.”\textsuperscript{149}

Section 550, of course, does not mention § 542(a) because § 542(a) is not an avoidance provision. So Pyatt and others\textsuperscript{150} make the argument that because §

\textsuperscript{139} 11 U.S.C. § 550.
\textsuperscript{140} Id. § 544.
\textsuperscript{141} Id. § 545.
\textsuperscript{142} Id. § 547.
\textsuperscript{143} Id. § 548.
\textsuperscript{144} Id. § 549.
\textsuperscript{145} Id. § 553.
\textsuperscript{146} Id. § 724.
\textsuperscript{147} Id. § 550(d). Section 550(d) provides that “[t]he trustee is entitled to only a single satisfaction under subsection (a).” Id.
\textsuperscript{148} 5 COLLIER ON BANKRUPTCY, supra note 17, at ¶ 550.05, 550-27.
\textsuperscript{149} Id.
\textsuperscript{150} For example, in the recent case of Shapiro v. Henson (In re Henson), 449 B.R. 109 (Bankr. D. Nev. 2011), the court followed a nearly identical analysis as Pyatt, stating:

Finally, we note that under Trustee’s interpretation of sec. 542(a), Trustee might have obtained double satisfaction by proceeding against the debtor through a motion for turnover, and against the creditors through motions to avoid post-petition transfers of property of the estate. If possession is not required, nothing in sec. 542(a) or the provision governing double satisfaction would prevent Trustee from doing so. Double satisfaction under sections 544, 545, 547, 548, 549, 553(b), or 724(a) is expressly prohibited under 11 U.S.C. § 550(d). That provision does not include any reference to sec. 542(a), and the absence of such a prohibition lends credence to our interpretation that sec.
542(a) is not mentioned, there is nothing that would prohibit a trustee from obtaining a double recovery—once from the debtor and once from the payees to whom the checks were transferred. This argument borders on the frivolous.\textsuperscript{151}

It is ultimately the trustee’s duty to collect property of the estate “as expeditiously as is compatible with the best interests of [the] parties in interest.”\textsuperscript{152} Spending additional time and wasting the estate’s resources to collect funds, which the trustee has already obtained, would violate his duties under the Bankruptcy Code. Furthermore, if a trustee were to actually seek a double recovery, the party from whom the second recovery was sought could clearly raise equitable defenses against such actions. Most, if not all, states forbid double recovery. Indeed,

[i]t is generally recognized that there can be only one recovery of damages for one wrong or injury. Double recovery of damages is not permitted, and the law does not permit a double satisfaction for a single injury. A plaintiff may not recover damages twice for the same injury simply because he or she has two legal theories . . . .\textsuperscript{153}

Furthermore, if a trustee recovered from the debtor under \(\text{§} \ 542(a)\) and then attempted a double recovery against the payee under \(\text{§} \ 549\), what would prevent the payee from asserting a \(\text{§} \ 550(d)\) defense and what court would allow a trustee to obtain such a windfall?

Finally, Pyatt’s concern about double recovery is not remedied by the holding that an entity is required to have \textit{current} possession or control over estate property to be subject to a turnover action. As correctly stated by one court:

[\textit{I}n a case where funds remain in a checking account on the date a trustee seeks turnover, both the bank (which would currently be in \textit{possession} of the funds) and the debtor (who would currently be in \textit{control} of the funds) could be the subject of the trustee’s turnover demand.\textsuperscript{154}]

\textsuperscript{151} See \textit{In re Ruiz}, 455 B.R. 745, 751–52 (B.A.P. 10th Cir. 2011) (“[I]t would be extremely unusual for \(\text{§} \ 542(a)\) to be referenced in \(\text{§} \ 550(d)\), as a matter of statutory construction. . . . Therefore, little, if anything, should be read into the failure to include \(\text{§} \ 542(a)\) in the provisions of \(\text{§} \ 550\.”).

\textsuperscript{152} 11 U.S.C. \(\text{§} \ 704(a)(1)\).

\textsuperscript{153} 25 C.J.S. \textit{Damages} \(\text{§} \ 5\) (2002).

\textsuperscript{154} \textit{In re Ruiz}, 455 B.R. at 752.
Accordingly, the double recovery that the Pyatt court and others fear could still occur.

Another problem with Pyatt’s conclusion—and possibly even a greater problem—is that it opens the door for serious abuse of the bankruptcy process. Pyatt focuses on the remote possibility of unintended double recovery, but completely ignores the consequences of its holding. If, as Pyatt concludes, § 542(a) actually permits a debtor to use or transfer property of the estate without consequence, does this not open the door for debtors to game the bankruptcy system? Indeed, under Pyatt’s reasoning, debtors may now be incentivized to dissipate assets of the estate under the assumption that as long as they are not in “possession” of the property, no penalties will follow.155 For example, in In re Anderson,156 a case which follows the Pyatt decision, the debtors argued, “Pyatt stands for the proposition that, even though funds in a [debtor’s] checking account as of the petition date are property of the estate,” a debtor is “free to write checks on such account post-petition, without consequence, up until the time that the trustee makes demand for turnover of such funds.”157 Stated differently, if a debtor is able to dissipate funds from her checking account before the trustee has a chance to review her bank statements and make a turnover demand for such funds, the debtor is off the hook—“the trustee must look elsewhere to recover those funds.”158 While the court in Anderson essentially rejected the debtors’ argument, it is further proof that Pyatt’s interpretation of § 542(a) is flawed and that debtors will use Pyatt as a good-faith justification to game the system. Undoubtedly, debtors and bankruptcy practitioners alike will read Pyatt to mean that a debtor is free to spend or transfer estate funds postpetition until such time as the trustee makes demand for such funds.159

155 This could, of course, jeopardize a debtor’s discharge under 11 U.S.C. § 727. But actual intent to hinder, delay or default must be established and as Pyatt itself illustrates, it is certainly possible for a debtor to benefit from postpetition transfers made without also [possessing] the requisite fraudulent intent. Moreover, [denying] the debtor [a] discharge will not restore to the estate property that otherwise could have been distributed to creditors.


157 Id. at 295 (emphasis added).

158 Id.

159 Creditors would also have an incentive to game the system. See, e.g., Beaman v. Vandeventer Black, LLP (In re Shearin), 224 F.3d 353, 357 (4th Cir. 2000) (stating that to read § 542(a) as requiring current possession “would enable possessors of property of the estate to escape trustees’ demands ‘simply by transferring the property to someone else’” (citation omitted)); Yoon v. Minter-Higgins, 399 B.R. 34, 43–44 (N.D. Ind. 2008) (“[I]t would be simple for debtors and aggressive creditors to game the system by writing checks immediately pre-petition to pay selected creditors while remaining secure in the knowledge that the trustees would be unlikely to pursue very small § 549 [unauthorized post-petition transfers] actions. This in turn would result in an unfair distribution to those creditors the
Thus, *Pyatt*’s argument regarding the potential of double recovery by the trustee to support the requirement of current possession of property in a turnover action under § 542(a) is without merit and should not be followed.\textsuperscript{160}

*Pyatt* and similar courts also base their decisions on the former Bankruptcy Act—despite the enactment of the Bankruptcy Code in 1978—as well as the pre-Code United States Supreme Court case of *Maggio v. Zeitz*.\textsuperscript{161} *Pyatt* made these observations regarding the floating check controversy and pre-Code practice:

Precode practice suggests that § 542(a) permits a trustee to compel turnover only from entities which have control of property of the estate or its proceeds at the time of the turnover demand. Precode practice is relevant in construing the bankruptcy code. It is especially instructive when interpretation of a “judicially created concept” is at issue, and turnover proceedings were an uncodified creation of the courts before enactment of the current code.

The leading case on pre 1978 turnover proceedings is *Maggio v. Zeitz*. There, the president of a bankrupt enterprise was ordered to turn over property which he did not have. He was jailed for contempt when he did not comply with the order. The Supreme Court held that the president was not a proper defendant in a turnover action, for turnover proceedings are permissible “only when the evidence satisfactorily establishes the existence of the property or its proceeds, and possession thereof by the defendant at the time of the proceeding.” The use of a turnover remedy was inappropriate “if, at the time it is instituted, the property and its proceeds have already been dissipated.” Precode practice thus required control of the property at the time the motion to compel turnover was brought.\textsuperscript{162}

\textsuperscript{160} It is worth noting that the debtor may have to pay twice under any scenario. If the trustee seizes the money from the debtor’s account, the result may be a bounced check. On the other hand, if the trustee pursues the payees directly, the debtor may need to make amends with the creditors depending on the nature of the debt. For example, if a debtor made a payment to her mortgagee prepetition and the trustee seized the funds representing the check to the mortgagee postpetition, the debtor would still have an obligation to pay the mortgagee. Since the mortgagee is secured by the debtor’s residence, if the debtor wants to remain in the property, she would need to stay current on her mortgage.

\textsuperscript{161} 333 U.S. 56 (1948).

\textsuperscript{162} Brown v. Pyatt (*In re Pyatt*), 486 F.3d 423, 428–29 (8th Cir. 2007) (citations omitted).
Pyatt, however, does not adequately appreciate the differences between the old Bankruptcy Act and the current Bankruptcy Code, which clearly altered the pre-Code possession requirements. Furthermore, pre-Code practice is applicable only if Congress, in enacting the Bankruptcy Code, has been ambiguous or the pertinent code section (i.e., § 542(a)) requires additional interpretation. But “[w]here the meaning of the Bankruptcy Code’s text is itself clear . . . its operation is unimpeded by contrary . . . prior practice.”163

Section 542(a) provides that “an entity . . . in possession, custody, or control, during the case, of property . . . shall deliver [it] to the trustee, and account for, such property or the value of such property.”164 The plain language of this statute provides that current possession is not a requirement under § 542(a).165 “During the case” refers to the entire bankruptcy case, not just the moment a turnover proceeding is commenced by the trustee.166 And what is further absent in Pyatt’s and other courts’ decisions is appropriate deference to the ability of the trustee to now compel, pursuant to § 542, the delivery of “the value of such property” instead of just the property itself.167 Pyatt acknowledges that § 542 had no counterpart under the former Bankruptcy Act, but rather turnover was a product of “judicial innovation” derived from the general concept that the courts were able to “cause the estates of bankruptcy to be collected.”168 But § 542(a) expanded the power to collect assets of the bankruptcy estate. Indeed, § 542(a) now provides a much

165 Hill v. Muniz (In re Muniz), 320 B.R. 697, 700 n.2 (Bankr. D. Colo. 2005) (“The fact that a trustee cannot demonstrate a debtor’s possession of estate property at the time the turnover action is filed merely means that his remedy becomes a money judgment for the value of the estate property rather than an order for turnover.”).
166 Beaman v. Vandeventer Black, LLP (In re Shearin), 224 F.3d 353, 356 (4th Cir. 2000) (“We construe the language ‘during the case’ to refer to the entire bankruptcy case, not just the adversary proceeding.” (citation omitted)); accord Boyer v. Carlton, Fields, Ward, Emmanuel, Smith & Cutler, P.A. (In re USA Diversified Prods., Inc.), 100 F.3d 53, 55 (7th Cir. 1996) (applying § 542(a) to “[o]ne who during a bankruptcy proceeding is in ‘possession, custody, or control’ of property” belonging to the debtor’s estate (emphasis added)); Rajala v. Majors (In re Majors), 330 B.R. 880, 2005 WL 2077497, at *4 (B.A.P. 10th Cir. 2005) (unpublished table decision) (“The obligation to turnover extends not just to property presently in someone’s possession, custody or control but to the property in ‘possession, custody or control during the case.’”); Redfield v. Peat, Marwick, Mitchell & Co., (In re Robertson), 105 B.R. 440, 457 (Bankr. N.D. Ill. 1989) (stating that “[t]he statute plainly applies to estate property that was possessed by anyone ‘during the case’, whether or not they still have it.”).
167 11 U.S.C. § 542(a) (emphasis added). Were that not true, the holding that debtors must repay to the estate the prepetition portion of any tax refunds received, even when the debtor has spent the actual refunds monies, would be in doubt. See Barowsky v. Serelson (In re Barowsky), 949 F.2d 1516, 1519 (10th Cir. 1991) (holding that tax refunds attributable to prepetition portion of taxable year was property of the estate).
broader remedy than solely turnover of property held at the time of the turnover proceeding, which likely occurs well after the filing of a bankruptcy petition. It contemplates the likelihood that a debtor or other entity has had and subsequently lost control of estate property.169

Furthermore, if the statute is read to require current possession of the property, as Pyatt and other courts do, then what does the language allowing the trustee to alternatively recover “the value of the property” even mean? Such an interpretation runs afool of the “longstanding canon of statutory construction that terms in a statute should not be construed so as to render any provision of that statute meaningless or superfluous.”170 Under Pyatt’s and similar courts’ reading of the statute, “the value of the property” has no meaning.171 As appropriately stated by one court:

[I]t is difficult for this court to read the trustee’s ability under [s]ection 542 to now recover the “value of such property” as an alternative to recovering the property itself as simply a reiteration of what Maggio recognized was a product of judicial necessity under the former Act. Rather, the common meaning of the phrase “value of such property,” when juxtaposed with the notion of the property itself being turned over, suggests to this court as well as others that Congress intended to expand upon Maggio, rather than to just codify it.172

169 See Boyer, 100 F.3d at 56 (“But by the time the trustee got around to demanding the money from the law firm, the law firm no longer had it, so how could it deliver it to the trustee? The statute, however, requires the delivery of the property or the value of the property. Otherwise, upon receiving a demand from the trustee, the possessor of property of the debtor could thwart the demand simply by transferring the property to someone else. That is not what the statute says, and can’t be what it means.” (emphasis added) (citation omitted)).
171 See Rajala, 330 B.R. 880, 2005 WL 2077497, at *4 (“[I]f a lack of present possession, combined with an explanation, constitutes sufficient compliance, little, if any, purpose would be served by the statutory alternative of requiring deliver of ‘the value of such property.’” (emphasis added) (citation omitted)).
172 In re Fleming, 424 B.R. 795, 804 (Bankr. W.D. Mich. 2010); see also Boyer v. Davis (In re USA Diversified Prods., Inc.), 193 B.R. 868, 879 (Bankr. N.D. Ind. 1995) (“Proceedings under § 542 effectively combine both the summary and the plenary remedies a trustee was required to pursue under the old Act, thus, allowing the pursuit in a single proceeding of relief that previously may have required two separate proceedings. Consequently, if the turnover defendant is still in possession of property of the estate, the trustee may recover that property. If the defendant no longer has possession of property of the bankruptcy estate, the court may inquire into the propriety of its disposition and, if appropriate, enter a money judgment in favor of the trustee for the value of such property. In effect, turnover proceedings have become what the Supreme Court noted they were not under the old Bankruptcy Act. They are not only the means by which the trustee can recover specific property of the estate and its identifiable proceeds, but also the means by which it can recover ‘damages for tortious conduct such as embezzlement,
The text of § 542(a) is plain. And “where, as here, the statute’s language is plain, ‘the sole function of the courts is to enforce it according to its terms.’”\footnote{173} Funds on deposit in a debtor’s bank account are undeniably “property of the bankruptcy estate.”\footnote{174} The fact that a debtor writes checks against the funds prepetition, but the checks do not clear the debtor’s account until postpetition, does not defeat the trustee’s right to recover the full amount transferred under § 542(a). If the checks do not clear the debtor’s account until after she files for bankruptcy, the debtor undeniably maintained control of the funds “during the case.”\footnote{175} Furthermore, the statute expressly allows the trustee to recover the “value” of such funds representing the checks. Thus, all of the elements required under § 542(a) to establish a claim against a debtor for turnover of funds that remained in the debtor’s account on the day she filed her bankruptcy petition are satisfied. Although one may be sympathetic to a debtor in this situation, the law simply provides the debtor no valid defense to the trustee’s turnover action under § 542(a).

C. The Policy Concerns and the Best Interest of Creditors

Nearly every court to analyze the floating check controversy discusses the policy concerns and the difficult position that the parties, particularly the debtors, face.\footnote{176} The reasoning relied on by courts to arrive at their differing outcomes


\footnote{174} In re Parsons, 57 Collier Bankr. Cas. 2d. (MB) 273, 273 (Bankr. S.D. Ind. Nov. 17, 2006).


\footnote{176} See, e.g., In re Pauls, No. 10-13887, 2011 WL 6096292, at *2 (Bankr. D. Kan. Dec. 5, 2011) (“The facts in this case are lamentable: an elderly grandmother’s attempt to assist her granddaughter and husband by co-signing and paying their debts results in the debtors being liable for funds they have repaid to her and that she, in turn, has paid to their creditor. But unfortunate facts do not soften the outcome that the law coldly compels. Randy and Ruth Ann Pauls had control of the $9,713.49 in their checking account on the date of their bankruptcy petition and must be ordered to account to the Trustee for it as § 542(a) requires.” (citation omitted)); In re Brubaker, 426 B.R. 902, 907 (Bankr. M.D. Fla. 2010) (“Although this Court has an enormous amount of sympathy for the pro se Debtors there is nothing in the record to reflect that the Debtors were acting in bad faith or with fraudulent intent. The Debtors simply seemed to be depositing funds, allowing debits from their account by merchants, and writing checks in the ordinary course.”); In re Sawyer, 324 B.R. 115, 123 (Bankr. D. Ariz. 2005) (discussing the “enormous amount of sympathy for the pro se Debtors in this case who apparently acted in good faith”); In re Dybalski, 316 B.R. 312, 316–17 (Bankr. S.D. Ind. 2004) (“The result in this case, at least on its face, seems rather unfortunate. . . . [T]he Code is primarily intended to give debtors a ‘fresh
pertains largely to the parties’ respective duties under the Bankruptcy Code. Courts also analyze which of the parties (i.e., the trustee or the debtor) is in the best position to prevent transfers by postpetition checks and remedy the damages to the estate caused by the transfers. This focus, however, misses the mark. Courts should not be focusing on which party is in the best position, but rather what the statute requires and which of the remedies available to the trustee provides the greatest return to creditors and is in the best interest of the estate.

In Pyatt, for example, the United States BAP of the Eighth Circuit, in holding that the responsibility for recovery of several unauthorized postpetition transfers was properly placed on the trustee, and ignoring what was in the best interest of the estate, held that the trustee was “in a better position to prevent transfers by postpetition check because the trustee can do so without the risk of criminal liability.” The court further held that “[a] trustee also is in a better position to remedy the damage to the estate caused by postpetition transfers because the trustee is the only party authorized by the Bankruptcy Code to avoid postpetition transfers, pursuant to 11 U.S.C. § 549.” In addition, the court found that

A debtor, on the other hand, runs the risk of being prosecuted for writing a bad check if he attempts to stop payment on an outstanding check on the eve of bankruptcy. Even though the debtor would likely prevail if he faced criminal charges for such conduct, presuming he acted without fraudulent intent, it is still inappropriate and unnecessary to place the debtor between the “rock” of possible criminal prosecution and the “hard place” of defending a turnover action by the trustee.

Id. (citation omitted).

Id. (emphasis added); see also Shapiro v. Henson (In re Henson), 449 B.R. 109, 113 (Bankr. D. Nev. 2011) (“Nor is Trustee left without an adequate remedy under the interpretation that a motion to compel turnover may only succeed when the entity has current possession of the property. Unlike the turnover provision, which governs the duty of an entity in possession of property of the estate during the case to turn over the property or the value of such property, sec. 549 expressly provides that ‘the trustee may avoid a transfer of property of the estate’ that occurs post-petition.”). The Pyatt court also observed:

The bankruptcy court alluded to the possibility that the Debtor might recover from the payees of the checks in the amount the court ordered the Debtor to turn over to the Trustee, but it did not specify . . . any Bankruptcy Code provision that authorizes a debtor to recover funds from postpetition transferees.

Pyatt, 348 B.R. at 786.
“because a trustee is the only party the Code authorizes to recover postpetition transfers, placing responsibility on the trustee for doing so under these circumstances is also the only option that advances the goal of equal distribution among creditors.”

Finally, the court stated the following:

Whether characterized as concern for fundamental fairness or practicality, it simply makes more sense to directly collect the postpetition transfers from the creditors who received the transfers rather than from the debtors who, presumably, innocently made the payments prepetition. In a perfect world, there would be a place on a debtor’s schedules or statement of financial affairs where outstanding checks could be readily listed, thereby alerting the trustee of the possible need to notify the bank to stop payment on those checks. In the absence of such perfection, however, debtors should be encouraged to disclose that information to their attorneys who, in turn, can communicate that to the trustee in some fashion. And the transfers that slip through the cracks could be avoided by the trustee.

But what “makes more sense” should not be the focus. There is no question that the floating check issue places a debtor in a problematic situation. The funds that were in the debtor’s checking account on the date the debtor files for bankruptcy are no longer available. Accordingly, this will likely require the debtor to pay the funds twice—once to the payees to whom the checks were originally payable, and a second time to the trustee. Admittedly, this result might go

---

179 Pyatt, 348 B.R. at 786. The court also explained:

If a trustee recovers from creditors who receive the postpetition transfers of the kind at issue here, those creditors’ claims could be reinstated to the extent of the recovery, and the trustee could then equally distribute that recovery among all of the unsecured creditors. In contrast, if a debtor is held accountable for checks cashed postpetition (and actually has the money to repay those funds), the trustee’s subsequent distribution of those funds would not be equal among creditors because the creditors who have received the unauthorized postpetition transfers will have already been paid 100 percent of what they were owed (to the extent of the transfers), whereas other creditors would most likely receive less than 100 percent.

180 Id. at 787 (emphasis added).

181 In most cases, the debtor is not purposefully cheating the bankruptcy estate. Rather, the debtor likely wrote valid checks in the ordinary course with the belief that the checks would clear prior to filing for bankruptcy. In In re Ruiz, for example, several of the checks were made as necessary business expenses, including hay and feed for livestock, which, by the time the trustee made demand for turnover, had been consumed by debtors’ livestock. In re Ruiz, 440 B.R. 197, 199 (Bankr. D. Utah 2010).
against the Bankruptcy Code’s policy of providing a debtor with a “fresh start” through bankruptcy. However, allowing the trustee to recover the value of the funds from the debtor may be in the best interest of the bankruptcy estate and comports with the trustee’s duties under § 704 of the Bankruptcy Code.

As set forth in Part II, a Chapter 7 trustee is “charged with marshaling the non-exempt assets of the estate for the benefit of creditors.” When prepetition checks dilute the bankruptcy estate postpetition, the Bankruptcy Code provides trustees with various nonexclusive remedies to reimburse the estate. Some of the remedies include the following:

[Trustees] can give and rely upon the Fed. R. Bankr. P. 2015(a)(4) notification to the banks if they have sufficient information to do so. They can seek to recover the transferred funds from the payees of the checks by pursuing an avoidance action under § 549. Or, they can seek recovery of the funds from the debtor without an adversary proceeding by invoking either the debtor’s duty to surrender under § 521(4) or his duty to turnover under § 542(a). It is ultimately the trustee’s duty to collect the property of the estate “as expeditiously as is compatible with the best interests of the parties in interest.”

Thus, a plain reading of the Bankruptcy Code indicates that a trustee can recover the funds from the debtor or the payees of the checks through several nonexclusive choices. As one court has explained, “A Chapter 7 trustee is given a substantial degree of discretion in deciding how to administer the bankruptcy estate and his or her actions are governed by a business judgment standard.” Accordingly, a trustee has significant discretion in deciding how to maximize the estate, and her decisions should not be disturbed if they are reasonable under the

---

183 Id. (“While the Debtor is correct in asserting that the Trustee could have sought to recover the funds from the payees of the checks by seeking to avoid the post-petition transfers, the Debtor does not point to any authority indicating that the Trustee must first exhaust his other remedies before he can seek turn over from the Debtor.” (citing 11 U.S.C. § 549 (2006))).
185 Debtors in several cases make the argument that in addition to § 549, recovery should come from the payees of the checks pursuant to § 547 (i.e., preferential transfers). Section 547 is not applicable. As set forth, supra note 134, a “transfer” occurs when the check is honored, but not received by the creditor. Thus, with respect to the floating check controversy, the checks, although issued prepetition, are not honored until after the bankruptcy is filed, postpetition. Accordingly, the preferential transfer theory asserted by some debtors is not an option for the trustee to pursue to recover the funds.
This means that the trustee is not required to prosecute the payees of the checks under § 549 (although she certainly can) if it makes more sense to pursue the debtor pursuant to § 542(a).

*Pyatt* and other courts have opposed this position and have failed to consider the best interests of the estate and the trustee’s “substantial degree of discretion.” These courts also disregard the plain language of § 542(a) and instead make policy-based arguments that the trustee is in the best position to prevent this situation by immediately notifying the debtor’s banks of the bankruptcy filing. Despite misinterpreting the Bankruptcy Code and ignoring the trustee’s duties, these holdings assume the trustee receives sufficient information to do so within days or even hours of the commencement of the case. More significantly, the *Pyatt* court and others assume that the debtor’s statements and schedules will be filed immediately and contain accurate information relating to the bank accounts, including (1) account numbers, (2) bank names, and (3) actual amounts in the accounts. Assuming all of that, courts argue that the trustee has all of the necessary information to take quick action. In reality, however, those assumptions are far from accurate. As often as not, the debtor’s statements and schedules will be

---

187 Id.

188 *See* Pyatt v. Brown (*In re* Pyatt), 348 B.R. 783, 785 (B.A.P. 8th Cir. 2006), *aff’d*, 486 F.3d 423 (8th Cir. 2007).

189 *See In re* Parker, No. 05-17912, 2008 WL 906570, at *6 (Bankr. N.D.N.Y. Apr. 3, 2008) (“[T]hese courts place the burden on the trustee, in reliance on the duty imposed under Federal Rule of Bankruptcy Procedure 2015(a)(4), to notify a debtor’s bank of the bankruptcy filing so that no further checks would be honored, or to stop payment on any outstanding checks. If a bank honors a check before receiving notice of the filing, they assert the trustee could seek to recover from the payee under § 549(a). The problem with this rationale is that often a debtor’s check may clear post-petition before the trustee is notified of his appointment, assuming the trustee has been provided with all the information needed to notify the debtor’s financial institution. More often than not, the debtor merely discloses a ‘checking account’ on schedule B, without any identifying account information. In addition, most debtors do not identify outstanding checks at the time of filing, and it is not until the first meeting of creditors that the trustee learns of these estate assets.”); *In re* Spencer, 362 B.R. 489, 493 (Bankr. D. Kan. 2006) (“In the present case (as in many, the Court suspects), the debtors did not identify which bank branches held their accounts or the applicable account numbers, making it difficult for the trustee to give the Rule 2015 notification. In addition, had the trustee made the notification, both banks would likely have frozen the accounts, exposing the debtors to the criminal and civil penalties that attend returned checks and thwarting their fresh start by rendering the debtors unable to use their accounts at all. Moreover, most of the checks written and honored post-petition were for a few hundred dollars each. Had the trustee chosen the § 549 route, he would be burdened with commencing a plethora of § 549 actions to recover two or three hundred dollars. This would be a poor economy to the estate, not to mention wasteful of the Court’s time and resources. One could hardly consider it ‘expeditious’ as § 704 requires.”).

190 *See, e.g., In re* Ruiz, 455 B.R. 745, 754 (B.A.P. 10th Cir. 2011) (“This case provides a good example why this is not practical [for the trustee to recover the funds from others]. This bankruptcy was filed on a Saturday. It would be a rare situation for a panel trustee to 1) be sitting in his or her office on a weekend, with staff, in the off-chance a
missing or will misstate a piece of crucial information. Indeed, in most cases, the debtor’s Statement of Financial Affairs will erroneously indicate the balance that they presumed would be in the account after the checks clear.\textsuperscript{191} Moreover, under the Bankruptcy Code, debtors are not required to submit their statements and schedules until fourteen days after filing the initial petition.\textsuperscript{192} This fourteen-day period alone eliminates the trustee’s ability to issue notice of the bankruptcy filing to the banks in sufficient time to prevent prepetition checks from clearing.

In all reality, the debtors are clearly in the best position to prevent prepetition checks from diluting the bankruptcy estate:

[D]ebtors (especially those represented by knowledgeable bankruptcy counsel) are in the best position to prevent this situation. First, and undoubtedly the most simple solution, is for debtors to wait until all outstanding checks have cleared the bank before filing their petition. Second, if immediate filing is required, because of a pending foreclosure or otherwise, they can stop payment on the outstanding checks. Third, debtors have the option of simply closing the bank accounts. Fourth, debtors can contact their banks and provide a notice of the filing of the bankruptcy petition, creating a duty on the part of the institutions to not pay the checks. It is true that the latter three options could theoretically cause debtors criminal problems on the back end, if someone suggested these steps were taken with the purpose to defraud. As the concurring opinion notes in \textit{In re Pyatt}, however, “there is clearly

\begin{itemize}
  \item bankruptcy would be filed;
  \item that he would necessarily be the one appointed as the panel trustee on that new case;
  \item that the schedules would show the existence of enough money in a checking account worth immediately acting on; and
  \item that the schedules would provide sufficient information to allow the trustee to contact any listed banking institutions.
\end{itemize}

In addition, since all four checks cleared within four days of the bankruptcy filing (and one the first business day after it was filed), even if the Trustee had mailed the letter the first business day the case was received, he could not have prevented the check from being honored.”); \textit{In re Parker}, 2008 WL 906570, at *5 (noting that “[t]he Debtor did not provide the Trustee with his account number or the branch where he does his banking. It was not until the Trustee reviewed the Debtor’s bank account statements that he was able to ascertain the actual account balance at filing and the Debtor’s account number. Thus, the Trustee was not in a position to put Trustco Bank on notice of the filing under Federal Rule of Bankruptcy Procedure 2015(a) prior to the checks at issue being presented for payment.”).

\textsuperscript{191} See, e.g., \textit{In re Ruiz}, 455 B.R. at 754 (“[I]f debtors are not properly counseled, they may inadvertently indicate in their schedules the balance they show in their check register, rather than the accurate amount actually still within their account. That is what appears to have happened here; Debtors listed the account with $10.02 balance instead of the actual amount of $3,764.99. Since panel trustees stand to only receive $60 in a no-asset case, the system is not set up to require those same trustees to spend their personal assets to seize a $10 bank account.”).

\textsuperscript{192} \textit{FED. R. BANKR. P. 1007(c)}. 
no purpose to defraud if a bankruptcy debtor stops payment on a check in fulfillment of the debtor’s duties under a federal statute.”

Although trustees may have other remedies available, the Bankruptcy Code clearly allows a trustee, pursuant to § 542(a), to recover property of the estate directly from the debtor. Trustees often do not have enough information in a short enough period of time to give the Rule 2015 notification. In addition, even if a trustee did have the information soon enough and issued the notification with sufficient dispatch, banks would likely freeze the account, exposing the debtor to the very criminal or civil penalties the Pyatt court refers to in arguing debtors have no duty to issue stop orders on prepetition checks. In other words, the “potential” criminal liability would appear to attach regardless of who prevented the checks from clearing. But a debtor’s compliance with her duty under a federal statute vitiates any purpose to defraud, and such an argument is nothing more than a red herring.

Moreover, it is often several, relatively insignificant prepetition checks that are honored postpetition. If the trustee were forced to pursue such transfers under § 549, as several courts suggest, she would be forced to commence several separate actions to recover each individual amount. As a practical consideration, requiring a Chapter 7 trustee to chase down multiple transferees who received postpetition transfers will often be impossible or extremely burdensome. In fact, the Bankruptcy Code discourages—possibly prohibits—the trustee from doing so. Section 704(a)(1) requires the trustee to collect property of the estate and “close

---

194 See Pyatt, 348 B.R. at 786–87.
195 As noted by one court, “[M]any, if not most, jurisdictions require the state to prove actual intent to defraud when prosecuting an individual for writing a bad check.” In re Ruiz, 455 B.R. at 753 n.26.
196 See, e.g., In re Schoonover, 2006 WL 3093649, at *3 (“It is true that the latter three options could theoretically cause debtors criminal problems on the back end, if someone suggested these steps were taken with the purpose to defraud. As the concurring opinion notes in Pyatt, however, ‘there is clearly no purpose to defraud if a bankruptcy debtor stops payment on a check in fulfillment of the debtor’s duties under a federal statute.’”).
197 As the In re Ruiz court noted:

One must also remember . . . that if the Trustee brought an action to avoid the post-petition transfer under § 549, and recovered that transfer for the estate under § 550, the holder of the note secured by the mortgage, for example, would be entitled to a claim under § 502(h) for the amount it was required to return to the Trustee. That claim would retain its secured status, leaving Debtors in the same position with regard to the amount of money they owed the note holder on the secured claim.

455 B.R. at 752 n.23.
such estate as expeditiously as is compatible with the best interests of parties in
interest.” This duty is also “required by implication from other provisions in the
Code and in the Federal Rules of Bankruptcy Procedure, especially Rule 1001,
which provides that the rules shall be construed to secure the just, speedy and
inexpensive determination of every case.”

In short, it is the trustee’s duty to the bankruptcy estate’s creditors to realize
from the estate all that is possible for distribution among creditors and to this end
the trustee’s best option—at times—may be to pursue the debtor for the value of
the funds instead of chasing down multiple creditors, most of whom have already
been harmed by the debtor’s bankruptcy filing. By recovering from the debtor, the
trustee can recover the value of all of the prepetition checks from one source
without initiating a single adversary proceeding. In several cases, such action
would be in the best interest of the bankruptcy estate and would comport with the
trustee’s duties under the Bankruptcy Code.

Furthermore, placing the onus on the debtor to reimburse the estate makes the
most sense because the debtor is in the best position to prevent prepetition checks
from clearing postpetition. Only the debtor has the ability to conduct the
appropriate prepetition planning to prevent this scenario from arising. Indeed, an
informed debtor is aware that by filing bankruptcy she is establishing the
bankruptcy estate. She has knowledge that any balance in her accounts on the
petition date becomes property of the estate. Thus, the simplest solution is for the
debtor to conduct the appropriate prepetition planning and wait until all
outstanding checks have cleared before filing her petition. In no way does this
prevent the debtor from paying ordinary bills prepetition.

Rather, it just requires that she take certain steps to ensure that payments clear
the bank prior to filing. This can be as simple as using cash or cashier’s checks in
the period before filing. Moreover, if immediate filing is necessary, the debtor is in
the best position to prevent any outstanding prepetition checks from clearing. For
instance, the debtor can (1) put stop payment orders on the checks, (2) give the
bank notice of the bankruptcy filing, or even (3) withdraw the funds and close the
bank account.

Ultimately, the policy-based arguments should not overshadow the plain text
of § 542(a) and courts should not be spending time determining which way the
various policy arguments cut. As noted by one court, “We are not free to impose
our wishes to fashion an exception to the sometimes harsh results of section
542(a).” Furthermore, it is not the “[c]ourt’s duty to create policy, but that of
Congress.” Thus, “while judges might crave the freedom to always decree what

199 6 COLLIER ON BANKRUPTCY, supra note 17, ¶ 704.02[3], at 704-9.
10th Cir. Aug. 29, 2005) (unpublished table decision); see also In re Pauls, No. 10-13887,
the outcome that the law coldly compels.”).
201 Redmond v. Miller (In re Miller), 378 B.R. 418, 2007 WL 2332391, at *3 (B.A.P.
is equitable and socially useful in the cases before [them] the Supreme Court says that [judges] do not possess it when a statute or rule provides *clear direction.*

This plain meaning rule has even greater force when applied to the text of the Bankruptcy Code:

> Initially, it is worth recalling that Congress worked on the formulation of the Code for nearly a decade. It was intended to modernize the bankruptcy laws, and as a result made significant changes in both the substantive and procedural laws of bankruptcy . . . . In such a substantial overhaul of the system, it is not appropriate or realistic to expect Congress to have explained with particularity each step it took. Rather, as long as the statutory scheme is coherent and consistent, there generally is no need for a court to inquire beyond the *plain language* of the statute. The task of resolving the dispute over the meaning [of a statute] begins where all such inquiries must begin: with the language of the statute itself.

The provisions of § 542(a) are clear, and allowing a trustee to recover the value of the funds from the debtor further comport with his duties under the Bankruptcy Code. And although a literal interpretation of § 542(a) may produce results that Congress may not have foreseen, legislation is the function of Congress, and Congress, not bankruptcy judges, must rewrite the Bankruptcy Code. The Supreme Court’s mandate for statutory construction clearly requires bankruptcy courts to enforce the plain meaning of a facially clear statute, in spite of policy implications or pre-Code practices to the contrary. Thus, under the analysis set forth above, even if the trustee is in a better position (which she is not) and has the option of recovering from others (which she does), in no way does that affect her right to recover from the debtor pursuant to § 542(a).

The question is not who is in a better position or whether there will be unfortunate results. The question is what the Bankruptcy Code requires and what is in the best interests of the estate. Bankruptcy courts cannot disregard the clear provisions of federal and state law, which provide, at a minimum, that (1) a debtor’s interest in her bank account becomes property of the estate when she files her petition, and (2) although a debtor may not have possession of the funds represented by the checks written to creditors, she undeniably has control if the bank has not authorized, as a payor bank, final payment on such checks. Once the payor bank authorizes payment and the debtor no longer has possession or control of the funds, § 542(a) clearly entitles the trustee to a money judgment for the “value” of the funds. This result not only follows the plain meaning of § 542(a), but it is the only result that comports with the trustee’s duty to realize from the

---


CONCLUSION

If a debtor writes checks against funds prepetition, but the checks do not clear the debtor’s account until after she files for bankruptcy, the trustee is entitled to a money judgment against the debtor for the value of the funds. Not only does this view comport with the clear text of § 542(a), it also may be the most practical and efficient way to maximize the value of the bankruptcy estate. The text of § 542(a) is clear. An entity in possession, custody, or control of property of the estate during the bankruptcy case “shall deliver to the trustee, and account for, such property or the value of such property.” Property of the estate is defined to include “all legal or equitable interests of the debtor.” Funds on deposit in a debtor’s bank account are undeniably property of the bankruptcy estate.

This is true even if the debtor writes checks against the funds prepetition, but the checks do not clear the debtor’s account until after she files for bankruptcy. Up until the time the debtor’s checks are presented for payment, the debtor retains control over the balance in her account. She can close the account, withdraw the funds, or stop payment on the checks, regardless of any outstanding checks. This undeniably amounts to “possession, custody, or control” within the meaning of § 542(a). Furthermore, the enactment of the Bankruptcy Code altered the pre-Code present possession or control requirement because § 542(a) expressly permits a trustee to recover “the value” of the property, in addition to the property itself, from one who possessed the property at any time “during the [bankruptcy] case.” Thus, present possession is no longer a prerequisite to turnover liability. Under such circumstances, a trustee can proceed directly against the debtor under § 542(a) and courts are constrained by the Bankruptcy Code to hold the debtor liable to the estate for the value of the checks written prepetition, but cashed postpetition.

While it is undoubtedly challenging for the debtor to pay twice and experience the “fresh start” intended by bankruptcy, the burden on the debtor is entirely irrelevant. It is well within the law and the trustee’s sound discretion to decide to recover the payments from the debtor instead of the payees under § 542(a). This result not only follows the plain meaning of § 542(a), but—at times—it is the only result that comports with the trustee’s duty to realize from the estate all that is possible for distribution among creditors and “close such estate as expeditiously as is compatible with the best interests of parties in interest.” While this may be an unfortunate result, it is entirely avoidable by prepetition planning on the debtor’s part. Debtors seeking to avoid the unpleasant result of

205 Id. § 542(a) (emphasis added).
206 Id. § 541(a)(1).
207 Id. § 704(a).
multiple payments should be certain that any checks they write prepetition have cleared before they file for bankruptcy. If immediate filing is necessary, the debtor can place stop payment orders on the checks, give the bank notice of the bankruptcy filing, or even withdraw the funds and close the bank account.
RECONCILING LIBERTY AND EQUALITY IN THE DEBATE OVER PREIMPLANTATION GENETIC DIAGNOSIS

Jessica Knouse*

Abstract

This Article draws on postmodern theory to develop a framework for analyzing situations in which liberty and equality appear to conflict. It uses the debate over nontherapeutic preimplantation genetic diagnosis as an example. Preimplantation genetic diagnosis is almost entirely unregulated within the United States, and there seems to be relative consensus that its use for therapeutic or medical trait selection—for example, selection against certain genetic and chromosomal disorders—should be permitted. There is substantial disagreement, however, as to whether its use for nontherapeutic or nonmedical trait selection—selection based on prospective parents’ preferences for a particular sex, eye color, hair color, skin color, or even “disability,” for example—should be permitted. This Article considers whether it would be constitutional to ban nontherapeutic trait selection. While some have argued that liberty requires giving prospective parents free reign to select their prospective children’s traits, others have argued that equality requires constraining such selections. Proponents of liberty typically worry that banning nontherapeutic trait selection would infringe on parental rights to make decisions about procreation and child rearing, while proponents of equality typically worry that allowing nontherapeutic trait selection will result in a proliferation of blond-haired, blue-eyed, light-skinned, nondisabled males and exacerbate existing identity hierarchies.

While these are the dominant arguments from liberty and equality, postmodern theory illustrates that they are not the only arguments. There are, in fact, liberty arguments in favor of banning nontherapeutic trait selection and equality arguments in favor of allowing it. One might, for example, argue that liberty requires banning nontherapeutic trait

* © 2013 Jessica Knouse. Associate Professor at the University of Toledo College of Law. She would like to thank everyone who offered comments on her presentation of this Article at the June 2012 Ohio Legal Scholarship Workshop, the September 2011 Saint Louis University School of Law Faculty Workshop, the October 2011 Constitutional Law Colloquium at Loyola University School of Law, and the October 2011 Central States Law School Association Conference. She would especially like to thank Professors Rebecca Zietlow, Shelley Cavalieri, Lee Strang, and Garrick Pursley. She would, as always, like to thank the University of Toledo College of Law for its generous financial support, and Mike Loegering for his enduring encouragement, wisdom, and love.
selection in the interest of enabling children to self-determine rather than being forced to live out the stereotypes associated with the traits their parents selected. Or one might argue that equality requires allowing nontherapeutic trait selection in the interest of treating therapeutic and nontherapeutic selection similarly, since it can in some cases be difficult to categorize a given selection as definitively therapeutic or nontherapeutic. Alternatively, one might argue that equality requires allowing nontherapeutic trait selection in the interest of treating selection among embryos the same as selection among sperm or ova donors, given that sperm and ova banks presently give prospective parents broad discretion to select among donors based on traits such as appearance and ability. There are, in other words, liberty and equality arguments on both sides of the debate. And, inasmuch as the Fourteenth Amendment protects both liberty and equality, this Article argues that we ought to avoid privileging one right at the expense of the other. It proposes that, when presented with multiple interpretations of liberty and equality, we ought to select those that are most likely to further our commitment to democracy, and that we can accomplish this by interpreting both rights with the goal of promoting diversity. Applying this principle to the debate over nontherapeutic trait selection, this Article concludes that, based on the current evidence, prospective parents should be given relatively broad latitude to make choices about their prospective children’s traits.

INTRODUCTION

Postmodern literary theory requires us to accept that textual meaning is never singular and never stable. Constitutional interpretation mindful of this insight allows us to understand that our commitments to liberty and equality have multiple and variable meanings. While, as a historical matter, liberty and equality have often been perceived as in tension and, at times, even mutually exclusive,1 as a constitutional matter, both receive protection from the same section of the same amendment.2 While history has, in other words, positioned the two rights as binary opposites, our Constitution positions them as partners. This Article therefore argues that the conventional liberty/equality binary is constitutionally problematic, and uses postmodern literary theory to deconstruct the binary and reconcile these two seemingly incompatible commitments. It posits that, once we understand that

1 See, e.g., Alexander Hamilton, Constitutional Convention: Remarks on the Term of Office for Members of the Second Branch of the Legislature (June 26, 1787), in 4 THE PAPERS OF ALEXANDER HAMILTON 218 (Harold C. Syrett & Jacob E. Cooke eds., 1962) (“[I]nequality [will] exist as long as liberty exist[s].”).

2 U.S. CONST. amend. XIV, § 1 (“No state shall . . . deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws.”).
liberty and equality both have multiple meanings, and that some of those meanings are in fact compatible, we can begin to interpret the two rights congruently.

While postmodern literary theory reveals multiple compatible interpretations of liberty and equality, postmodern political theory guides the normative project of selecting among those interpretations. Just as literary theory can inform the process of constitutional interpretation, political theory can inform the substance of constitutional interpretation. To the extent that postmodern political theory can be described as promoting any singular substantive value, it is the value of diversity. By challenging entrenched power, resisting the myth of objectivity, and expressing “incredulity toward metanarratives,” postmodern political theory uncovers a reality full of incoherence, subjectivity, and fragmentation—in other words, a reality full of diversity. It is because of this ability to uncover a diverse array of identities (and ideologies) that postmodern political theory can serve a normative function in interpreting the Constitution. Diverse identities are valuable to democracy because they promote robust debate, ensure accurate representation, and thereby legitimize the exercise of government power. This Article accordingly argues that, in selecting among multiple compatible interpretations of liberty and equality, we ought to prefer those that uncover the diverse identities that always already populate our public forum and, in turn, strengthen our democracy.

While the deep flaws of the liberty/equality binary are evident across a variety of policy debates, they are particularly evident in debates surrounding new, some might say postmodern, technologies. One debate that the liberty/equality binary is particularly ill equipped to address—but which it has, nevertheless, dominated—involves assisted reproductive technology (ART) and, specifically, the nontherapeutic use of preimplantation genetic diagnosis (PGD). PGD occurs within the context of in vitro fertilization: After ova and sperm have been combined to create several embryos, PGD can be used to determine the eye color, hair color, skin color, sex, and even “disabilities” associated with each embryo. Prospective parents can then select which embryos to transfer and carry to term. While a number of nations, including England, Germany, and Italy, have banned some nontherapeutic uses of PGD, such uses (and, indeed, most ARTs) remain virtually unregulated within the United States. There is thus a growing body of

3 JEAN-FRANÇOIS LYOTARD, THE POSTMODERN CONDITION: A REPORT ON KNOWLEDGE, at xxiv (Geoff Bennington & Brian Massumi trans., Univ. of Minn. Press 1984) (“I define postmodern as incredulity toward metanarratives.”).

4 See infra Part II.A.2.c (claiming that PGD can be used to determine eye, hair, and skin color based on the fact that the Fertility Institutes advertised these services in 2009). While the advertisement was withdrawn after public outcry, its existence suggests that the capability to make such selections had been acquired.

5 I have used quotation marks to emphasize the subjective nature of the label disability. Although they are not used throughout the Article, the label is always subjective.

scholarship dedicated to assessing the advisability and constitutionality of banning nontherapeutic uses of PGD. Much of this scholarship is driven by the liberty/equality binary: Some argue that liberty rights prevent the government from banning nontherapeutic uses of PGD, while others argue that equality considerations require the government to ban such uses. This Article critiques both arguments and suggests that, by positioning liberty and equality as binary opposites, where in any given case one must prevail and the other yield, neither fulfills the promise of the Fourteenth Amendment.

Part I introduces the theoretical constructs of postmodernism, liberty, and equality. It begins by arguing that, when we interpret liberty and equality as binary opposites, we create a constitutional conundrum: It is inappropriate to select interpretations in which the two rights are mutually exclusive, when both receive protection from the same section of the same amendment. Part I proceeds by illustrating how postmodern theory can resolve this conundrum by deconstructing the binary opposition and allowing us to interpret liberty and equality as congruent rather than conflicting. Part II examines how the liberty/equality binary has driven (and arguably distorted) the debate over nontherapeutic PGD. It illustrates that the common arguments privilege one right while subordinating the other, and thereby dishonor the Fourteenth Amendment. Part III applies the postmodern theory set forth in Part I to the practical debate presented in Part II. It deconstructs the liberty/equality binary as applied to the debate over nontherapeutic PGD by explaining that, while the dominant liberty and equality arguments may indeed be incompatible, there are other currently marginalized liberty and equality arguments that are fully compatible. The Article concludes that, in selecting among multiple compatible interpretations of liberty and equality, we ought to favor those that reveal rather than obscure the diverse array of identities necessary for robust debate, accurate representation, and successful democracy.

I. POSTMODERNISM, LIBERTY, AND EQUALITY

Postmodern literary theory, by revealing that liberty and equality both have multiple meanings—some of which are incompatible, but others of which are entirely compatible—gives us the option of interpreting the Constitution so as to simultaneously further both rights. Postmodern political theory, by revealing the diverse identities that exist within our society, gives us the ability to promote public forum debate, ensure accurate representation, and ultimately improve our democracy. Part I, in sum, argues that to fulfill our constitutional commitments and improve our democracy, we ought to interpret liberty and equality with the shared goal of uncovering diverse identities and ideologies. Part I.A introduces postmodern literary and political theory. Postmodern literary theory reveals the multiplicity of meanings that are always already present in texts, while postmodern political theory reveals the multiplicity of identities that are always already present in societies. Postmodern literary theory can be useful in interpreting our Constitution, and postmodern political theory can be similarly useful in improving our democracy. Part I.B introduces the liberty/equality binary and illustrates how
postmodern literary and political theory can be used to deconstruct that binary, thereby reconciling our dual commitments.

A. Postmodern Theory

Postmodernism is both an era and a philosophy. It is enigmatically indefinable in that “the act of defining postmodernism is anti-postmodern.” While it is somewhat disingenuous to offer a definition under such circumstances, it is equally disingenuous to entirely reject the possibility of definition. As Professor Adam Todd has aptly observed, “A complete rejection of modernism is itself a modernist act.” A (concededly imperfect) definition of postmodernism is, therefore, offered: As an era, postmodernism succeeds modernism. While the modern era began with the Enlightenment and continued at least through the Industrial Revolution, the postmodern era began in the latter half of the twentieth century and persists to the present. As a philosophy, postmodernism transcends modernism. While modern theorists reified coherence, objectivity, and metanarratives, postmodern theorists embrace the reality of incoherence, subjectivity, and micronarratives. The remainder of this section is dedicated to introducing the specifics of postmodern literary and political theory. Postmodern literary theory will be used infra, in Part III.A to reveal that our constitutional commitments to liberty and equality have multiple meanings. Postmodern political theory will be used infra, in Part III.B, to provide guidance in the normative project of selecting among those meanings.

---

7 J.M. Balkin, What Is a Postmodern Constitutionalism?, 90 Mich. L. Rev. 1966, 1967–71 (1992) (“First, postmodernism is the cultural era in which we live . . . . Second, postmodernism is also the set of cultural products created during the era of postmodernity. . . . [T]hird, . . . [p]ostmodernism is also a set of critical or theoretical claims . . . .”); Adam G. Todd, Painting a Moving Train: Adding “Postmodern” to the Taxonomy of Law, 40 U. Tol. L. Rev. 105, 105 (2008) (“[T]he word ‘postmodern’ is applied as a label to certain products emerging from both the period and philosophical movements of the postmodern era.”).

8 Adam Todd, Neither Dead nor Dangerous: Postmodernism and the Teaching of Legal Writing, 58 Baylor L. Rev. 893, 902 (2006) (explaining that “the act of defining postmodernism is anti-postmodern because it is a reductionist and modernist act”); see Todd, supra note 7, at 109 (“Postmodernism is characterized by the idea that it is difficult, if not impossible, to come up with a clear definition for any term, including the term postmodernism.”).

9 Todd, supra note 8, at 907 (explaining that “the antimodernist postmodernists who espouse a total rejection of modernism have arguably fallen into a modernist trap”).

10 Todd, supra note 7, at 109–10.

11 Balkin, supra note 7, at 1971–72 (explaining that postmodernism attacks “‘totalizing’ theories or ‘master narratives’ that seek to explain all or substantially all of society, history, knowledge, the nature of femininity, or virtually anything else within a comprehensive and articulate theory”).
1. Postmodern Literary Theory

Postmodern literary theory illustrates that all texts have multiple and variable meanings.12 While a text generally has one dominant meaning, it always has additional marginalized meanings.13 By revealing that a text’s dominant meaning in fact rests on highly subjective (yet often widely accepted) assumptions, postmodern literary theory can uncover a text’s marginalized meanings.14 Our constitutional commitment to liberty is, for example, dominantly understood to protect a broad range of decisions, including those related to childrearing.15 This dominant meaning can, however, be subverted by revealing that it rests on the highly subjective assumption that protecting such decisions unambiguously enhances liberty. While protecting decisions about childrearing may enhance parents’ liberty, it may diminish children’s liberty.16 Postmodern theory thus reminds us that liberty’s dominant meaning is simply one of its many possible meanings.

Yet subverting a text’s dominant meaning is only the first step in the postmodern project, the aim of which is not merely to invert the hierarchy—thereby rendering the initially marginalized meanings dominant—but to deconstruct the hierarchy and reveal that there is no authentic meaning.17 As Derrida explained, “deconstruction” accomplishes both “an overturning of the classical opposition and a general displacement of the system.”18 Professor Pauline

12 Stephen M. Feldman, Diagnosing Power: Postmodernism in Legal Scholarship and Judicial Practice (with an Emphasis on the Teague Rule Against New Rules in Habeas Corpus Cases), 88 NW. U. L. REV. 1046, 1048 (1994) (“According to postmodernists, the meaning of a text is never grounded or stable, and therefore one can always find multiple meanings or truths.”); Todd, supra note 8, at 907 (explaining that postmodernism “removes the necessity of privileging one position over another and can allow multiple positions and interpretations to exist simultaneously”).
13 Feldman, supra note 12, at 1048 (“[O]ne performs a postmodern flip by taking a segment of a text, event, or concept that apparently has been reduced to a static meaning or truth and suggesting the possible existence of another (often radically different) meaning or truth. This alternative meaning or truth often emerges after one uncovers and disturbs the usually tacit assumptions underlying the original meaning. The postmodern flip then is completed by exploring how this new meaning or truth of the segment of the text, event, or concept might reorient one’s understanding of the whole.”).
14 Id.
15 See infra Part II.B.1.
16 See infra Part III.A.1.
17 Joel F. Handler, Postmodernism, Protest, and the New Social Movements, 26 LAW & SOC’Y REV. 697, 698–99 (1992) (describing postmodern literary theory as ascribing to the idea that “there is no one authentic reading or meaning of a given text”).
Rosenau, elaborating on Derrida’s explanation, wrote, “Deconstruction tears a text apart, reveals its contradictions and assumptions; its intent, however, is not to improve, revise, or offer a better version of the text.”¹⁹ Its intent is, instead, to illustrate that whenever one textual interpretation is expressed, others are inevitably repressed.²⁰ Deconstruction is, as a technical matter, accomplished by repeatedly re-inverting the initial hierarchy, such that the balance of power between the initially dominant and initially marginalized components becomes so uncertain that the concepts of dominant and marginalized themselves become irrelevant. It is not until the categories that initially populated the hierarchy (e.g., liberty and equality) have become irrelevant that the postmodern project is complete. Part III will deploy this theory to illustrate that deconstructing our constitutional commitments to liberty and equality entails not merely identifying their currently dominant and marginalized meanings, but also accepting the impossibility of identifying a single or stable set of authentic meanings.

2. Postmodern Political Theory

While postmodern literary theory can be used to uncover previously marginalized textual meanings, postmodern political theory can be used to uncover previously marginalized identities—or, in other words, to uncover existing diversity. As Professor Joel Handler has explained, postmodernism illustrates that “every ‘identity’ necessarily suppresses an alternative identity.”²¹ The expression of one identity is, indeed, always tantamount to the repression of another.²² Postmodern political theory, paralleling postmodern literary theory, reveals that the dominance of certain identities (e.g., nondisabled white men) is, in fact, dependent on the marginalization of other identities (e.g., disabled minority women). By revealing such dependencies, postmodern political theory has the capacity to deconstruct identity-based hierarchies.²³ Once a given hierarchy (e.g., abled/disabled, white/nonwhite, man/woman) has been repeatedly inverted, the categories that initially populated it become politically irrelevant, and the diverse identities they once obscured are revealed.

Yet observable traits like race, sex, and (in some cases) ability are not sufficient descriptors of identity. Identity is the product of many influences,²⁴ some

---

²⁰ Handler, supra note 17, at 699 (“When we define something—anything—we are necessarily excluding or ‘repressing’ something else. Thus, all meaning has a ‘surplus,’ that which is repressed along with that which is articulated.”).
²¹ Id.
²² Id.
²³ Id. at 700 (explaining how postmodern theorists argue that “hegemony is never stable”).
²⁴ Id. at 706 (“Identity is constructed by race, ethnicity, class, community, [and] nation . . . .”).
of which are genetically determined and some of which are socially constructed. It is, therefore, dangerous to reify either genetic essentialism or social constructionism. Professor Belinda Bennett, addressing the debate over cloning, described the dangers of genetic essentialism—that is, of “see[ing] people purely as the product of their genes”—stating: “[W]hat is written in code is only the introduction. The rest of the story we write ourselves.”25 While skin color may be genetically determined, racial groups are socially constructed; while anatomical sex may be genetically determined, gender groups are socially constructed; and, while ability may be genetically determined, ability groups (e.g., “deaf culture”) are socially constructed. Yet social constructionism is equally dangerous, because it deprives individuals of the ability to express their genetic predilections. Furthermore, neither our races, sexes, and abilities, nor our racial, gender, and ability groups can adequately account for our ideologies. Members of a single racial group can espouse different ideologies—they may, for example, be Republicans, Democrats, socialists, Christians, Muslims, or atheists. Conversely, members of different racial groups can share common ideologies. Identity is, from a postmodern perspective, both genetic and social, both physical and ideological, and “both multiple and unstable.”26

Just as tangible hierarchies based on race, sex, or ability can be deconstructed to reveal the multiplicity of races, sexes, and abilities that always already exist within our society, intangible hierarchies based on ideology can be deconstructed to reveal the multiplicity of viewpoints that always already exist within our society. Professor Handler has noted that, from the postmodern theorist’s perspective, any form of “foundationalism or essentialism—whether liberal capitalism or Marxism—[is] a fundamental obstacle to the deepening and extension of democracy throughout civil society.”27 Privileging any one identity—whether based on race, sex, ability, ideology, or some other characteristic—is detrimental to democracy because it inevitably obscures other identities and, in doing so, obscures diversity. Recognizing the diverse identities that populate our public forum is beneficial to democracy because it ensures robust debate and

25 BELINDA BENNETT, HEALTH LAW’S KALEIDOSCOPE: HEALTH LAW RIGHTS IN A GLOBAL AGE 48 (2008). She later writes:

The idea that there are essential, immutable human characteristics has been challenged by scholars from postmodernism, feminism and critical race theory. Our understandings of the self and of the body have been cut free of their traditional anchors and have moved into a new arena where meanings are socially, culturally and historically constructed. While there is no doubt that in biological terms the genome itself has changed little in recent history, the same cannot be said of the genome’s social and cultural significance.

Id. at 61.
26 Handler, supra note 17, at 706.
27 Id. at 700.
B. Liberty and Equality

Part I.B applies postmodern theory to the Constitution. First, Part I.B.1 observes that the dominant meanings of liberty and equality often conflict and that this conflict leads to viewing the two rights as binary opposites. It should be noted at the outset that the liberty/equality binary is not a typical binary because one right is not always privileged, and the other is not always subordinated. In some cases, liberty is privileged and equality is subordinated; in others, equality is privileged and liberty is subordinated. There is, of course, a third set of cases in which both liberty and equality are privileged—but these (easier) cases are not this Article’s focus. This Article is dedicated to providing guidance in those cases where the dominant meanings of liberty and equality conflict, such that reaching a resolution seems to require that one be privileged and the other, subordinated.

Second, Part I.B.2 reveals that both liberty and equality in fact have additional marginalized meanings and that uncovering these meanings gives us the option of interpreting the two rights congruently—which seems appropriate in light of the fact that they are both protected by the same section of the same amendment. It suggests that, when faced with multiple compatible meanings of liberty and equality, we should select those that will promote diversity. We can, by doing so, use postmodern theory to deconstruct the liberty/equality binary and reconcile the two commitments.

I. The Liberty/Equality Binary

Since our nation’s inception, liberty and equality have been perceived as in tension and, at times, even mutually exclusive. At the Philadelphia Convention in 1787, Alexander Hamilton said, “Inequality will exist as long as liberty exists, and

---

28 The Supreme Court has recognized the importance of diverse viewpoints, saying:

“[I]t is only through free debate and free exchange of ideas that government remains responsive to the will of the people and peaceful change is effected. The right to speak freely and to promote diversity of ideas and programs is therefore one of the chief distinctions that sets us apart from totalitarian regimes.”

Terminiello v. Chicago, 337 U.S. 1, 4 (1949).

29 Consider, for example, Lawrence v. Texas, 539 U.S. 558 (2003), Loving v. Virginia, 388 U.S. 1 (1967), and other decisions that could have been justified on both due process and equal protection grounds.
it unavoidably results from that very liberty itself.”30 Early Americans did not, however, have to grapple directly with this perceived tension between liberty and equality—at least not as a matter of constitutional interpretation—because the pre-Civil War Constitution protected liberty but failed to mention equality.31 It was not until the Reconstruction era that Americans made formal constitutional commitments to both liberty and equality: through Section 1 of the Fourteenth Amendment, they barred the government not only from depriving any person of liberty without due process of law, but also from denying any person equal protection of the laws.32 Americans have, thus, only had to navigate the complex and often tense relationship between these two constitutional commitments since 1868.

Professor Peter Westen has cataloged some of the perceived tensions between liberty and equality. He wrote, in a 1982 article titled The Empty Idea of Equality:

Equality is commonly perceived to differ from rights and liberties. Rights are diverse; equality is singular. Rights are complicated; equality is simple. Rights are noncomparative in nature . . . ; equality is comparative . . . . Rights are concerned with absolute deprivation; equality is concerned with relative deprivation. Rights mean variety, creativity, differentiation; equality means uniformity. Rights are individualistic; equality is social.33

When liberty and equality are thus perceived as binary opposites, a presumption arises that, in any given case, the two must be balanced to determine which should prevail and which should acquiesce. Under such an approach, Professor Westen observed, “Equality is sometimes said to be flourishing at the expense of rights, [and] rights are sometimes said to be flourishing at the expense of equality.”34 This Article builds upon Professor Westen’s view that “this contrasting of rights and equality is fundamentally misconceived,”35 and argues that, when liberty and equality are balanced, both are diluted, and the promise of the Fourteenth Amendment remains unfulfilled.

31 U.S. CONST. amend. V.
32 U.S. CONST. amend. XIV, § 1.
34 Id. at 539.
35 Id. But see Kenneth L. Karst, Why Equality Matters, 17 GA. L. REV. 245, 249 (1983) (“[I]n American public life and constitutional law the idea of equality carries a meaning quite removed from the empty tautology that like cases should be treated alike. This meaning is not derived from dictionaries or deductive logic, but from centuries of American experience.”).
2. Deconstructing the Liberty/Equality Binary

Postmodern literary theory illustrates that both liberty and equality have multiple meanings. Although, according to Aristotle, liberty meant that “every man should live as he likes,” more recent philosophers have suggested that liberty’s meaning is actually indeterminate unless and until the terms of a given debate have been specified. And although, according to Aristotle, equality meant that “things that are alike should be treated alike, while things that are unalike should be treated unalike in proportion to their unlikeliness,” Professor Westen’s work illustrates that this traditional definition of equality is a tautology and that, actually, “[e]quality is an empty vessel with no substantive moral content of its own.” Professor Philip Kurland similarly recognized that the rhetoric of equality “is subject to use, if not capture, by anyone on any side of the question.” Both liberty and equality, then, must have a multitude of meanings beyond their dominant meanings.

Rather than accepting the conventional—and constitutionally problematic—perception that we protect liberty for one set of reasons and equality for another (conflicting) set of reasons, this Article argues that we ought to protect liberty and equality for the same reasons. One reason that we might protect liberty and equality is that both promote the expression of diverse identities because this will, in turn, promote our broader commitment to democracy. Liberty promotes the expression of diverse identities by empowering individuals to self-determine according to their own predilections, which are inherently diverse. Equality

---

36 1 ARISTOTLE, THE POLITICS OF ARISTOTLE, bk. VI ch. 2, at cix (B. Jowett trans. 1885) (c. 384 B.C.E.) (“[O]ne principle of liberty is that all should rule and be ruled in turn. . . . Another principle of liberty is non-interference—every man should live as he likes, if this is possible . . . .”).

37 Gerald C. MacCallum, Jr., Negative and Positive Freedom, 76 PHIL. REV. 312, 333 (1967) (“[D]iscussions of the freedom of agents can be fully intelligible and rationally assessed only after the specification of each term of this triadic relation has been made or at least understood.”). Professor Peter Westen characterizes MacCallum as saying that “freedom, too, is a formal concept with no substantive content apart from the variables inserted to give it meaning.” Westen, supra note 33, at 547 n.31.

38 Westen, supra note 33, at 543 n.18, citing ARISTOTLE, ETHICA NICOMACHEA bk. V, at 1131a–1131b (W.D. Ross trans., 1925) (c. 384 B.C.E.).

39 Id. at 547.


41 One can see an example of this phenomenon in the fact that, as social norms regarding the “ideal” family have weakened, a more diverse set of families have emerged. Just as self-determination promotes diversity, diversity, conversely, promotes self-determination. See Armin von Bogdandy, The European Union as Situation, Executive, and Promoter of the International Law of Cultural Diversity—Elements of a Beautiful Friendship, 19 EUR. J. INT’L L. 241, 247 (2008) (describing the “postmodernism-inspired
promotes the expression of diverse identities by assuring individuals that, when they do self-determine, their differences will not be grounds for illegitimate marginalization. Diverse identities are, indeed, beneficial to democracy because they improve public forum debate, ensure accurate representation and, in so doing, legitimize the exercise of government power. This Article accordingly argues that, when presented with multiple meanings of liberty and equality, we ought to select those that are most likely to promote diversity.

II. THE DEBATE OVER NONTHERAPEUTIC PGD

While the liberty/equality binary influences many debates, it is particularly influential in debates surrounding ART and, specifically, the use of technology to select the traits of one’s prospective children. Part II.A describes the current technology, as well as its use and limited regulation. Part II.B describes the debate surrounding this technology and identifies the dominant arguments—some of which rest on our constitutional commitment to liberty and suggest that nontherapeutic trait selection should be permitted, and some of which rest on our constitutional commitment to equality and suggest that nontherapeutic trait selection should be prohibited.

A. Current Technology, Practice, and Regulation

Throughout history, at least some prospective parents have wanted to select the attributes of their prospective children—sometimes out of a desire to maximize their social status and opportunity and sometimes out of a desire to reproduce themselves. Some prospective parents in patriarchal societies have wanted male opinion [that] cultural diversity is central for the individual identity, since cultural diversity opens up concrete options used by the individual to design its identity”).

42 Expression of identity will not, in other words, be grounds for exclusion from or diminished power within the public forum.

43 The Supreme Court has recognized the promotion of diversity as a compelling interest—at least within the context of higher education. Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 722 (2007) (recognizing “diversity in higher education” as a “compelling” interest for purposes of strict scrutiny).

44 See Sigal Klipstein, Preimplantation Genetic Diagnosis: Technological Promise and Ethical Perils, 83 FERTILITY & STERILITY 1347, 1347 (2005); see also Dov Fox, Racial Classification in Assisted Reproduction, 118 YALE L.J. 1844, 1846 (2009) (observing that “many of those who wish to become parents would prefer a child whose racial features resemble their own”). The impulse to engage in sex selection has existed throughout history. As early as the fifth century BCE, the Greek philosopher Anaxagoras suggested that men tie off their left testicle prior to intercourse because sperm from the right testicle was believed to produce males. Owen D. Jones, Sex Selection: Regulating Technology Enabling the Predetermination of a Child’s Gender, 6 HARV. J.L. & TECH. 1, 4 (1992), citing Ronald J. Levin, Human Sex Pre-Selection, 9 OXFORD REV. REPROD. BIOLOGY, 161, 162 (1987). This theory endured for millennia. Id. (“French noblemen, for
some in racially stratified societies have wanted blond-haired, blue-eyed, light-skinned children; and some in deaf communities have wanted deaf children. While prospective parents have always been able to make general choices about their children’s attributes through “eugenic dating” (and, since the 1960s, through the use of donor material), they could not, until the 1970s, make specific choices about their children’s attributes. And, even in the 1970s, such choices could only be made through chorionic villus sampling or amniocentesis, both of which are only available during ongoing pregnancies. Chorionic villus example, were still advised, more than 2200 years later, that removal of the left testicle guaranteed male heirs.”). Others promoted different methods of sex selection. Aristotle, for example, taught that vigorous intercourse would increase the chances of a male child. Id.

Cecilia L.W. Chan et al., Gender Selection in China: Its Meanings and Implications, 19 J. ASSISTED REPROD. & GENETICS 426, 426 (2002) (“Gender selection is welcomed by many societies with gender-specific preference, especially those patriarchal societies such as Chinese communities.”).

See Kari L. Karsjens, Boutique Egg Donations: A New Form of Racism and Patriarchy, 5 DEPAUL J. HEALTH CARE L. 57, 78 (2002) (“Wealthy couples, who utilize egg brokers or high profile advertisements, do not seek general traits. These couples are seeking a ‘perfect gene pool’ for their commodity—notice the highly sought after donor is a woman who has blonde hair [and] blue eyes, received a 1400 on her SAT, attends an Ivy League school, and who preferably has some additional talents . . . .” (citation omitted)). But see Leslie Bender, Genes, Parents, and Assisted Reproductive Technologies: ARTs, Mistakes, Sex, Race, & Law, 12 COLUM. J. GENDER & L. 1, 65 (2003) (discussing African American prospective parents who used ART, in part, because they “wanted to reproduce themselves as African-Americans”).

Richard Gray, Couples Could Win Right to Select Deaf Baby, TELEGRAPH (Apr. 13, 2008), http://www.telegraph.co.uk/news/uknews/1584948/Couples-could-win-right-to-select-deaf-baby.html (recounting the story of a deaf couple wishing to select for deafness). Professor Dena Davis has written on the issue of “disabled persons wishing to reproduce themselves in the form of a disabled child.” Dena S. Davis, Genetic Dilemmas and the Child’s Right to an Open Future, 28 RUTGERS L.J. 549, 569 (1997). While Professor Davis allows that deafness may be a culture rather than a disability, she maintains, “deliberately creating a deaf child counts as a moral harm.” Id. at 575.

Note, Regulating Eugenics, 121 HARV. L. REV. 1578, 1586 (2008) (“Eugenic dating is simply dating with the goal of finding a mate who will provide desirable genes for one’s offspring, whether such genes are for hair color or for intelligence.”).

Nan T. Ball, The Reemergence of Enlightenment Ideas in the 1994 French Bioethics Debates, 50 DUKE L.J. 545, 549 (2000) (“The first officially sanctioned human sperm banks began to appear around the world in 1963, first in the United States . . . .”). Sperm and ova banks, of course, enable only limited control over children’s attributes. Unless donor sperm or ova is used in conjunction with technologies such as chorionic villus sampling, amniocentesis, or PGD, prospective parents cannot be certain of their children’s attributes in advance of birth.

sampling is best performed at ten to fourteen weeks and amniocentesis, even later, meaning that the sole remedy for “unsatisfactory” test results is abortion. But, in 1990, new technology enabled prospective parents to make specific choices about their children’s attributes prior to the advent of pregnancy. Part II.A.1 describes this new technology, Part II.A.2 describes how it is currently being used in fertility clinics across the United States, and Part II.A.3 describes the virtually complete absence of legal regulation.

1. Current Technology

When the first successful use of PGD was reported in 1990, prospective parents using in vitro fertilization (IVF) acquired the ability to determine the genetic and chromosomal characteristics of their embryos prior to transfer and implantation. Ova, extracted using a transvaginal needle guided by ultrasound, were placed in vitro and inseminated. The first day after insemination, fertilization was confirmed; the third day after insemination, when the embryos had reached the six to eight cell (or “cleavage”) stage, one or more cells (or “blastomeres”) were removed from each embryo and tested to determine its genetic and chromosomal characteristics. Prospective parents were then able to select which embryos they wished to transfer and carry to term.

51 Thomas M. Jenkins & Ronald J. Wapner, First Trimester Prenatal Diagnosis: Chorionic Villus Sampling, 23 SEMINARS PERINATOLOGY 403, 403 (1999) (“When performed between 10 and 14 weeks’ gestation, [chorionic villus sampling] is both safe and effective in the diagnosis of fetal chromosomal, biochemical, and molecular disorders, with risks comparable to those of second trimester amniocentesis.”).
52 Jain, supra note 50, at 649.
54 Id. at 419.
55 Id. at 420.
56 Id. at 420 (“Fertilization is confirmed the day following insemination. On day 3 post insemination, the in vitro generated embryos should each have six to eight cells.”).
57 Id. at 420 (“Two types of diagnostic tests are most commonly performed in the biopsied cells: polymerase chain reaction (PCR) and fluorescent in situ hybridization (FISH). The former is used for the diagnosis of specific gene defects, and the latter for numerical or structural anomalies of the chromosomes.”); see also Susannah Baruch et al., Genetic Testing of Embryos: Practices and Perspectives of U.S. In Vitro Fertilization Clinics, 89 FERTILITY & STERILITY 1053, 1054 (2008) (“The genetic material for testing can be obtained from one or more cells (called blastomeres) removed from the early embryo or from polar body cells cast off by the egg as it matures and is fertilized.”).
Today, PGD can be used to test for a vast array of characteristics. Although it is difficult, if not impossible, to accurately categorize the various tests, some are considered therapeutic while others are considered nontherapeutic. Therapeutic tests are commonly used to detect single gene disorders such as cystic fibrosis, Tay-Sachs, and sickle cell anemia, as well as chromosomal abnormalities such as Down syndrome and Turner syndrome. Nontherapeutic tests can be used to select for a particular sex, and sometimes even for a particular disability, such as Down syndrome or deafness. Nontherapeutic tests can also theoretically be used to select for a particular eye color, hair color, or skin color. And they may, in the future, be used to select for less tangible traits like intelligence and personality. It should be noted that some tests—such as those used to detect sex and disabilities—have both therapeutic and nontherapeutic relevance. Testing for sex has therapeutic relevance in that it can prevent X-linked diseases such as hemophilia, as well as nontherapeutic (social) relevance for prospective parents who want a child of a given sex. Testing for disabilities has therapeutic relevance for prospective parents who want a child without a given disability, and nontherapeutic (social) relevance for prospective parents who want a child with a given disability. While therapeutic

---

58 Zhuang & Zhang, supra note 53, at 420 (“At present, PGD can be performed, in principle, for any genetic condition for which there is sufficient sequence information.”).

59 It should be noted that the terms therapeutic and nontherapeutic are socially constructed—and, in some cases, highly controversial. The therapeutic/nontherapeutic binary is discussed in detail, infra Part III.A.2.

60 Klipstein, supra note 44, at 1348; see also Zhuang & Zhang, supra note 53, at 420.

61 Gregory Katz and Stuart O. Schweitzer, reporting on PGD practices within the United Kingdom, write, “There has been a small number of cases in which deaf couples have used IVF and PGD to select embryos with the same genetic traits that they themselves have in order to share a common lifestyle with their offspring.” Gregory Katz & Stuart O. Schweitzer, Implications of Genetic Testing for Health Policy, 10 YALE J. HEALTH POL’Y, L. & ETHICS 90, 114 (2010). Professor Kirsten Smolensky writes, “[O]ne IVF doctor has reported that he ‘flatly refused a couple who asked him to identify an embryo with Down syndrome, so they could give their Down-affected child a similar sibling.’” Kirsten Rabe Smolensky, Creating Children with Disabilities: Parental Tort Liability for Preimplantation Genetic Interventions, 60 HASTINGS L.J. 299, 304 (2008) (quoting Melissa Healy, Fertility’s New Frontier: Advanced Genetic Screening Could Help Lead to the Birth of a Healthy Baby, L.A. TIMES, July 21, 2003, § F (Health), at 1).

62 Klipstein, supra note 44, at 1349 (“[The human genome project] may allow us to determine which genes lead to physical characteristics that we deem beautiful, whether these be facial structure, hair, or eye color.”).

63 Id. at 1348 (“PGD could theoretically allow for the selection of any trait whose genetic composition is known. These could range from sex, height, [athletic ability,] intelligence, and beauty, to even theoretically personality traits such as cheerful disposition.”).

64 Id. at 1349.
uses of PGD are relatively uncontroversial, nontherapeutic uses, which are the focus of this Article, are highly controversial.65

2. Current Practice

PGD is not yet common, but as its accessibility increases, its commonality is increasing. A 2006 study performed by Susannah Baruch of the Genetics and Public Policy Center at Johns Hopkins University reported that there are over 400 ART clinics in the United States and that, of the 186 that participated in the study, nearly three quarters (74%) offered PGD.66 Many of the clinics that did not offer PGD in 2006 predicted that they would offer it in the future.67 The study revealed that, in 2005, approximately 3,000 cycles of PGD were performed, and that (in all likelihood) between 4% and 6% of all IVF cycles included PGD.68 While PGD was most often performed for therapeutic reasons, 9% of the cycles performed in 2005 involved nontherapeutic sex selection.69 While Baruch’s study was not specific to nontherapeutic PGD, it can (in conjunction with other studies) be helpful in illuminating current practices. The remainder of this section surveys the availability and incidence of nontherapeutic selections for sex, disability, and eye, hair, and skin color. It should be noted at the outset that these traits have been chosen as the focus of this Article because they all influence identity: Sex creates presumptions about gender identity; disability creates a range of presumptions that vary depending on the specific condition; and eye, hair, and skin color create presumptions about racial identity. While prospective parents cannot use PGD to select disabilities for which they are not carriers or to select races entirely different from their own, they could, in theory, use it to emphasize or deemphasize traditional markers of race—for example, by selecting an embryo with slightly darker or lighter eye, hair, and skin color—and potentially of disability.

65 Andrea L. Kalfoglou et al., *PGD Patients’ and Providers’ Attitudes to the Use and Regulation of Preimplantation Genetic Diagnosis*, 11 REPROD. BIOMEDICINE ONLINE 486, 486 (2005), http://download.journals.elsevierhealth.com/pdfs/journals/1472-6483/PIIS1472648310611455.pdf (“Virtually all participants supported the use of PGD to avoid severe, life-threatening genetic illness or to select embryos that are a tissue match for a sick sibling, but their attitudes varied significantly over the appropriateness of using PGD to avoid adult-onset genetic disease, to select for sex, or to select for other non-medical characteristics.”); see also Klipstein, *supra* note 44, at 1349.

66 Baruch et al., *supra* note 57, at 1054.


68 Baruch et al. *supra* note 57, at 1054.

69 *Id.* at 1055–56.
(a) Sex Selection

Many clinics allow prospective parents to engage in nontherapeutic sex selection. Of the clinics that participated in Baruch’s study, 137 offered PGD, and 42% of those 137 allowed prospective parents to engage in nontherapeutic sex selection.70 Of that 42%, almost half (47%) allowed prospective parents to select for sex under any circumstances, 41% allowed prospective parents to select for sex only if they already had one or more children (presumably, of the sex they were selecting against), and 7% allowed prospective parents to select for sex only if PGD was already being performed for an independent therapeutic reason.71 Of the clinics that offered PGD but had never allowed nontherapeutic sex selection, over three-quarters reported that their decision not to allow it was motivated by ethical concerns.72 Some, however, reported that they had never received any requests.73 While Baruch’s study did not reveal the percentage of patients who sought nontherapeutic sex selection, other studies have shown that it is substantial. Doctor Tarun Jain reported in 2005 that there was “significant demand for preimplantation sex selection for nonmedical reasons . . . .”74 Doctor Richard Sharp found that those reasons ranged from a “desire for gender balance in the family” to an “interest in specific parenting experiences.”75 Despite the substantial demand, patients who engage in nontherapeutic sex selection recognize that their decisions are ethically complex: as Doctor Sharp further found, “even the most motivated of couples pursuing IVF/PGD for sex selection tend to express moral reservations.”76

---

70 Id. at 1056.
71 Id.
72 BARUCH ET AL., supra note 67, at 5; see also Kalfoglou et al., supra note 65, at 487 (reporting that “88% of respondents to a survey of American Society for Reproductive Medicine (ASRM) members believed that PGD is an acceptable clinical procedure, but 69% did not think patients should be able to use it for non-medical sex selection”).
73 BARUCH ET AL., supra note 67, at 5.
74 Jain, supra note 50, at 657 (adding that “a significant portion of this demand [came] from patients who do not have any children or have children all of one sex”). Doctor Jain reports, “Of 561 respondents, 229 (40.8%) wanted to select the sex of their next child for no added cost.” Id. at 651; see also Stacey A. Missmer & Tarun Jain, Preimplantation Sex Selection Demand and Preferences Among Infertility Patients in Midwestern United States, 24 J. ASSIST. REPROD. & GENET. 451, 451 (2007) (“Of respondents, 49% wanted to select the sex of their next child for no added cost.”).
76 Id. at 845. Doctor Sharp further reported that several couples expressed “anxiety about the prospect of telling their other children about their plans to use IVF/PGD for sex selection.” Id. at 844.
(b) Disability Selection

While many clinics allow prospective parents to engage in nontherapeutic sex selection, only a few allow prospective parents to engage in nontherapeutic disability selection. While requests to select for a disability may be less common than requests to select for sex, they are far more controversial. In a 2005 study performed by Doctor Andrea Kalfoglou of Johns Hopkins’s Genetics and Public Policy Center, several PGD providers expressed discomfort with selecting for traits they felt would disadvantage children, such as phenylketonuria, achondroplasia, and deafness. Some prospective parents do, however, seek to select for such traits, and some clinics do accommodate their requests. Indeed, 3% of the PGD clinics surveyed in Baruch’s 2006 study reported that they had assisted parents in selecting for disabilities. Even PGD providers with ethical reservations about selecting for such traits may, in some cases, be persuaded by persistent patients. Professor Kirsten Smolensky recounts an anecdote about a couple with achondroplasia, who “told their physician that if he refused to help them select a child with achondroplasia, they would go to another IVF clinic, refuse PGD testing, get pregnant, have the fetus tested via amniocentesis for achondroplasia, and abort any child not carrying the gene.” The physician, “[n]ot wanting to be the cause of an unnecessary abortion and recognizing that the end result would be the same with or without his assistance (a child with achondroplasia), . . . agreed to help the parents utilize PGD to select for a child with achondroplasia.”

77 Phenylketonuria is a condition that prevents the breakdown of the amino acid phenylalanine. It is associated with light eye, hair, and skin color, as well as seizures and mental retardation. It can, however, be treated by a diet low in phenylalanine. See Phenylketonuria, PUBMED HEALTH, http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002150/ (last reviewed June 17, 2011).


79 Kalfoglou et al., supra note 65, at 493.

80 Deaf parents have periodically sought to use PGD to select for deafness. See Gray, supra note 47. Parents may seek to select for a variety of disabilities beyond those listed above.

81 BARUCH ET AL., supra note 67, at 5 (“Some prospective parents have sought PGD to select an embryo for the presence of a particular disease or disability, such as deafness, in order that the child would share that characteristic with the parents. Three percent of IVF-PGD clinics report having provided PGD to couples who seek to use PGD in this manner.”).

82 Smolensky, supra note 61, at 305 (describing “[o]ff-the-record conversations with reproductive endocrinologists[, which] suggest that patients may be ‘strong-arming’ physicians into agreement”).

83 Id.
(c) Eye-, Hair-, and Skin-Color Selection

While many clinics allow sex selection and at least a few allow disability selection, none currently allow eye-, hair-, or skin-color selection. While the Fertility Institutes (a Los Angeles based ART clinic) briefly advertised such a service in 2009, there was an outpouring of opposition. The clinic’s director, Doctor Jeff Steinberg, who was part of the team that had reported the first successful use of IVF in 1978, initially said that that he was willing to live with the opposition and that several prospective parents had already inquired about the service. He noted that new technology is often controversial—in 1978, someone had left a note on his windshield saying “test tube babies have no soul,” yet IVF is now widely accepted. Other doctors, however, expressed disapproval. Doctor William Kearns, whose work helped to enable trait selection, said “Steinberg has jumped on my research but I’m totally against this. My goal is to screen embryos to help couples have healthy babies free of genetic diseases. Traits are not diseases.” Similarly, Doctor Mark Hughes, who was part of the team that reported the first successful use of PGD in 1990, condemned trait selection as “ridiculous and irresponsible.” The Fertility Institutes’ advertisement, which had initially appeared in February of 2009, was withdrawn in early March with a statement that, “[i]n response to feedback . . . , an internal, self regulatory decision has been made to proceed no further with this project.”

3. Current Regulation

At present, PGD clinics in the United States are governed by very little beyond internal self-regulatory decisions. While a number of countries have banned some nontherapeutic uses of PGD, the United States has allowed all uses of PGD—and, indeed, IVF and ART more generally—to remain almost entirely...

84 The Center for Genetics and Society, however, reports, “Some clinics have even offered [PGD] for purely cosmetic traits including eye color, hair color, and skin complexion.” About Genetic Selection, CENTER FOR GENETICS & SOC’Y, http://www.geneticsandsociety.org/section.php?id=82 (last visited Nov. 20, 2012).
86 Id. (reporting that Dr. Steinberg had “received ‘five or six’ requests from couples for the new service”).
87 Id.
88 Id.
89 Id.
91 See Zafran, supra note 6, at 215–16.
Neither Congress nor any of the federal agencies that could ostensibly regulate PGD have provided any guidance. While the Centers for Disease Control and Prevention (CDC) do require IVF clinics to report certain data, such as pregnancy success rates, they do not ask them to report what tests (such as PGD) are being performed. And, inasmuch as the only penalty for not reporting is a “noncompliance” listing on the CDC’s website, even this regulation is far from robust. While the Food and Drug Administration (FDA) does regulate some of the drugs, devices, and substances (e.g., reproductive tissues) used in IVF and PGD, “there is no uniform system to assure test accuracy or validity.” Furthermore, while the Centers for Medicare and Medicaid Services regulate clinical laboratories, they have taken the position that PGD “is an assessment of a product and therefore falls under FDA’s oversight of reproductive tissue,” which means that they do not regulate laboratories that perform PGD. Finally, while some states oversee the genetic tests associated with PGD, none regulate PGD directly.

Despite the lack of formal legal regulation, clinics are not completely without guidance. Several professional organizations—most notably, the American Society

92 Id. at 215.
93 See Baruch et al., supra note 57, at 1056 & n.19 (citing Fertility Clinic Success Rate and Certification Act of 1992, 42 U.S.C. § 263a-1 (2006)).
94 Id. at 1056.
96 Baruch et al., supra note 57, at 1056 & n.23 (citing Clinical Laboratory Improvement Amendments of 1988, 42 U.S.C. §§ 263–263b (2006)).
97 Id. at 1056 (citation omitted) (internal quotation marks omitted).
98 Id. at 1056–57 & n.30; see N.Y. STATE DEP’T OF HEALTH, CLINICAL LABORATORY STANDARDS OF PRACTICE: PART 2—SPECIALTY REQUIREMENTS 4–5, 71–75 (2012), available at http://www.wadsworth.org/labcert/clep/files/NYSDOH_Standards_Part_2_Specialty_Requirements_08_12.pdf (establishing technical standards for genetic testing). Professor Radhika Rao has discussed other related state laws, such as a Louisiana law regulating the disposition of embryos and the (arguably unconstitutional) Pennsylvania and Illinois laws prohibiting sex-selective abortion. Radhika Rao, Equal Liberty: Assisted Reproductive Technology and Reproductive Equality, 76 GEO. WASH. L. REV. 1457, 1485–86 (2008) (citing 720 ILL. COMP. STAT. ANN. 510/6(8) (West 2010) (“No person shall intentionally perform an abortion with knowledge that the pregnant woman is seeking the abortion solely on account of the sex of the fetus.”); LA. REV. STAT. ANN. §§ 9:129–:130 (2009) (“A viable in vitro fertilized human ovum is a juridical person which shall not be intentionally destroyed by any natural or other juridical person or through the actions of any other such person”); 18 PA. CONS. STAT. ANN. § 3204(c) (West 2000) (“No abortion which is sought solely because of the sex of the unborn child shall be deemed a necessary abortion.”)).
The ASRM’s position has, however, changed over time, particularly with respect to the use of nontherapeutic sex selection. In 1994, the ASRM’s Ethics Committee stated, “[W]hereas preimplantation sex selection is appropriate to avoid the birth of children with genetic disorders, it is not acceptable when used solely for nonmedical reasons.” Ten years later, in 2004, the Committee retained its position on therapeutic sex selection but took a somewhat more equivocal view of nontherapeutic sex selection. With respect to patients who were already undergoing IVF for therapeutic reasons, the Committee said nontherapeutic sex selection “should not be encouraged.” With respect to patients wishing to initiate IVF to engage in nontherapeutic sex selection, however, the Committee said nontherapeutic sex selection “should be discouraged.” While expressing concerns that nontherapeutic sex selection could lead to sex discrimination, individual and social harm, and the inappropriate use of medical resources, the Committee emphasized that it did not see sufficiently certain “grave wrongs or sufficiently predictable grave negative consequences” to warrant legal prohibition.

**B. The Perceived Liberty/Equality Binary**

As IVF and nontherapeutic PGD have become more accessible, the debate over their legal regulation has intensified. Scholars have made a variety of arguments against and in favor of legal regulation, many of which can be understood as manifestations of the perceived liberty/equality binary discussed in Part I. Liberty is generally cited as a reason for allowing nontherapeutic PGD, whereas equality is generally cited as a reason for banning nontherapeutic PGD.

---


101 Id. at S248.

102 Id.

103 Id.

104 Id.

105 This Article focuses on the arguments from liberty and equality, but it should be noted that these are not the only arguments in the literature.

106 See infra Part II.B.1.

107 See infra Part II.B.2.
This section reviews the most common liberty and equality arguments and illustrates how they have driven the debate over nontherapeutic PGD.

1. Liberty Arguments Against Banning Nontherapeutic PGD

Liberty arguments against banning PGD arise from the Due Process Clauses of the Fifth and Fourteenth Amendments, which collectively prevent the federal and state governments from depriving any person of “liberty . . . without due process of law.” The Supreme Court has interpreted the liberty protected by the Due Process Clauses broadly, such that it encompasses not only physical but also decisional autonomy. Its protection of decisional autonomy extends to decisions about procreation (i.e., “whether to bear or beget a child”) as well as to decisions about childrearing (i.e., how to “direct the upbringing and education of [one’s] children”). While the Court has not had occasion to consider whether liberty specifically protects the use of ART, one could argue that, since decisions about procreation and childrearing are protected, decisions about ART (and, by extension, nontherapeutic PGD) should similarly be protected. This section examines arguments that the liberty protected by the Due Process Clauses is broad enough to protect nontherapeutic PGD based on its association with decisions about either procreation or childrearing.

The right to make decisions about procreation—or, in the Court’s words, about “whether to bear or beget a child”—was articulated in dicta in the 1972 decision of *Eisenstadt v. Baird*, which invalidated a state ban on distributing contraceptives to unmarried persons as violating the Equal Protection Clause. It was subsequently reaffirmed in the 1973 decision of *Roe v. Wade*, as well as the 1992 decision of *Planned Parenthood of Southeastern Pennsylvania v. Casey*, both of which involved the invalidation of state abortion laws. While the right was referenced in the 2003 decision of *Lawrence v. Texas*—which provides the most recent evidence of its existence—it was called into question by the 2007 decision of *Gonzales v. Carhart*, where the Court upheld a highly restrictive

---

108 U.S. CONST. amends. V, XIV.
111 405 U.S. 438.
112 Id. at 453.
113 410 U.S. 113.
114 505 U.S. 833.
115 539 U.S. 558.
116 Id. at 565.
federal abortion law. Assuming that, despite Carhart, some species of the right still exists, there is some evidence that it applies not only to plaintiffs wishing to avoid procreation (as in Eisenstadt and its progeny), but also to plaintiffs wishing to procreate, as in the Court’s 1942 decision of Skinner v. Oklahoma. In Skinner, while invalidating a state policy of sterilizing certain habitual criminals, the Court described procreation as a “basic civil right” that is “fundamental to the very existence and survival of the race.” While Skinner was decided on equal protection rather than due process grounds, it nevertheless suggests that the Constitution protects decisions to procreate as well as decisions not to procreate.

Professor John Robertson has been one of the primary proponents of the argument that the liberty protected by the Due Process Clauses is broad enough to encompass the use of ART and, specifically, of PGD. He wrote, in a 1996 article, “[T]he main support for a right to engage in prebirth selection rests on the close connection between the expected characteristics of offspring and the decision whether or not to reproduce.” Relying on cases like Skinner and Eisenstadt, Professor Robertson explained:

If reproductive decisions are fundamental, then access to information material to the decision to reproduce should be equally fundamental. If a person would choose not to reproduce if she knew that the child would have a disability or some other undesired characteristic, then she should be entitled to have that information and to act on it. Her right to avoid reproduction for any reason would entitle her to avoid reproduction for a particular reason. Similarly, her right to have offspring generally should entitle her to have offspring only if she thinks that offspring will have particular characteristics.

On this reasoning, prospective parents would have a right to use PGD to select for any trait that is so “material to [their] reproductive decision [that it] determines whether reproduction will occur.” This could potentially include both

---

118 Id. at 168 (upholding the Partial-Birth Abortion Ban Act against a facial challenge on the basis that it was “void for vagueness . . . [and] impose[d] an undue burden on a woman’s right to abortion”).
119 Although one can read Gonzales as calling the right to abortion into question, there is as yet little evidence that it called the right to contraception into question.
120 316 U.S. 535 (1942).
121 Id. at 541.
123 Id.
124 Id. at 427 (citation omitted).
125 Id. at 429 (stating that the court must determine whether “the characteristic in question is one that is central or material to a reproductive decision—whether the characteristic determines whether reproduction will occur[, and if so, the law must give
therapeutic and nontherapeutic uses of PGD. \textsuperscript{126} Part II.A, above, illustrates that there may be prospective parents for whom sex and disabilities (e.g., achondroplasia) are but-for conditions of reproduction.

Even if the right to use nontherapeutic PGD is not protected by virtue of its association with decisions about procreation, it may be protected by virtue of its association with decisions about childrearing. The right to make decisions about childrearing was first articulated in the 1923 case of \textit{Meyer v. Nebraska}, \textsuperscript{127} where the Court invalidated a statute barring young children from learning foreign languages, in part, because it burdened parents’ rights “to control the education of their own.” \textsuperscript{128} The right was subsequently reaffirmed in the 1925 decision of \textit{Pierce v. Society of Sisters}, \textsuperscript{129} where the Court invalidated a statute requiring children to attend public schools, in part, because it burdened parents’ rights “to direct the upbringing and education of children under their control.” \textsuperscript{130} While many rights recognized during the early twentieth century did not survive the New Deal, parental rights were an important exception. \textsuperscript{131} They were protected—albeit in conjunction with religious rights—by the 1972 decision of \textit{Wisconsin v. Yoder}, \textsuperscript{132} where the Court exempted Amish parents from a law requiring children to attend either public or private school until the age of sixteen. \textsuperscript{133} The Court held that the Amish parents had a fundamental right to direct the “religious upbringing of their children,” and that the state’s interest in education was not sufficiently compelling to justify infringing that right. \textsuperscript{134} Parental rights were most recently recognized in this choice the same respect and weight it gives to other decisions about whether or not to reproduce”).

\textsuperscript{126} See, e.g., \textit{supra} Part II.A.2.b (discussing achondroplasic parents).
\textsuperscript{127} 262 U.S. 390 (1922).
\textsuperscript{128} \textit{Id.} at 401. The Court also stated that liberty encompasses the right to “marry, establish a home and bring up children.” \textit{Id.} at 399.
\textsuperscript{129} 268 U.S. 510 (1924).
\textsuperscript{130} \textit{Id.} at 534–35.
\textsuperscript{131} Modern cases in which parental rights were recognized—even if they did not always prevail—include, among others: \textit{Santosky v. Kramer}, 455 U.S. 745, 753 (1982) (noting that parents have a “fundamental liberty interest . . . in the care, custody, and management of their child”); \textit{Parham v. J.R.}, 442 U.S. 584, 602 (1979) (noting that “the family [is generally viewed] as a unit with broad parental authority over minor children”); \textit{Stanley v. Illinois}, 405 U.S. 645, 651 (1972) (mentioning the “the interest of a parent in the companionship, care, custody, and management of his or her children”); and \textit{Prince v. Massachusetts}, 321 U.S. 158, 166 (1944) (“It is cardinal with us that the custody, care and nurture of the child reside first in the parents . . . .”).
\textsuperscript{132} 406 U.S. 205 (1972).
\textsuperscript{133} \textit{Id.} at 234. It should be noted that the result would likely have differed had the claim not sounded in both parental \textit{and} religious freedom. The Court stated, “A way of life, however virtuous and admirable, may not be interposed as a barrier to reasonable state regulation of education if it is based on purely secular considerations . . . .” \textit{Id.} at 215.
\textsuperscript{134} \textit{Id.} at 214. The Court emphasized that the case did not present the question of what might happen if a child wished to continue her schooling and her parent wished to
the 2000 decision of *Troxel v. Granville*, where the Court stated, “[I]t cannot . . . be doubted that the Due Process Clause of the Fourteenth Amendment protects the fundamental right of parents to make decisions concerning the care, custody, and control of their children.”

Professor Robertson has argued that, although the “main support” for a right to PGD arises from its association with decisions about procreation, secondary support arises from its association with decisions about childrearing. He and others, including Professor Maxwell Mehlman, have observed that, but for the fact that decisions arising from PGD occur before birth whereas decisions about healthcare, education, and socialization occur after birth, they are quite similar. Professor Mehlman has emphasized the breadth of the liberty that parents currently enjoy in directing their children’s healthcare, education, and socialization. With respect to healthcare, he notes, “Parents have been allowed to withhold consent to corrective surgery for a child’s heart defect; refuse to consent to chemotherapy; . . . and donate a child’s kidney to a sibling.” While parental discretion over healthcare is not entirely unlimited, it is certainly substantial. With respect to education and socialization, Professor Mehlman notes that “parents routinely choose or change their children’s futures . . . [by] picking their . . . schools and selecting their extra-curricular activities.” Because the law gives parents such substantial control over their existing children, one can make the argument that it should give them similar control over their prospective children.

withdraw her for religious reasons. The Court explained, “Our holding in no way determines the proper resolution of possible competing interests of parents, children, and the State . . . .” *Id.* at 231.

*Id.* at 66 (holding that a Washington statute, which allowed any person to petition for visitation with a child at any time and did not require courts to give any deference to parental wishes, unconstitutionally burdened the parental rights of the plaintiff mother). It should be emphasized that none of the above cases presented the question of what would happen if a child disagreed with her parents’ decision. The children in these cases are presumed to have agreed with—or at least acquiesced in—their parents’ decisions.


*Mehlman, supra* note 138, at 111 (citations omitted).

*Id.* (“[C]ourts have intervened when parents failed to obtain treatment for a five-week-old infant with two broken arms; when a mother exposed a child to secondhand cigarette smoke during visitations; [etc.] . . . .”).

*Id.* at 109. Professor Mehlman further observes, “Parents also place their children at risk when they permit them to play dangerous sports.” *Id.* at 111.

One could even make the argument that, if a parent is going to be permitted to require her child to practice the piano for five hours every day, it is in the child’s interest to have a genetic predisposition to musicality.
Not everyone, however, is willing to accept the argument that the liberty protected by the Due Process Clauses protects the use of nontherapeutic PGD. Professor Radhika Rao, for example, has written that, although such an argument is “plausible,” it cannot ultimately succeed. Applying the traditional test for determining whether a given right is fundamental—that is, whether it is deeply rooted in our nation’s history and tradition—Professor Rao concludes that there is no fundamental right to ART. She reaches this conclusion, in part, because ART has “been in existence for too short a time for there to have developed any tradition of legal protection” and, in part, because “such an expansive reading of the privacy cases is unwarranted.” Professor Rao explains the latter point by stating that “the Constitution does not guarantee reproductive autonomy all by itself, disentangled from concerns about bodily integrity and inequality.” Her article, which illustrates the weaknesses of the dominant liberty arguments, is one of the few that addresses the complex interrelationship between our dual commitments to liberty and equality.

2. Equality Arguments in Favor of Banning Nontherapeutic PGD

Even assuming that the liberty of the Due Process Clauses protects a fundamental right to nontherapeutic PGD, that right could be burdened by a law “narrowly tailored” to a “compelling” government interest. While the government may have a variety of interests in banning nontherapeutic PGD, some of the most compelling relate to equality. Our constitutional commitment to equality is reflected in a number of provisions, but most notably in the

143 Rao, supra note 98, at 1462 (“[T]he ‘liberty’ protected under the Due Process Clause of the Fourteenth Amendment doesn’t appear to include a fundamental right to use ARTs.” (citation omitted)).
144 Id. (citing Washington v. Glucksberg, 521 U.S. 702, 710–22 (1997)).
145 Id. (noting also that “IVF itself has been around for only thirty years, while PGD has been practiced for less than two decades” (citation omitted)).
146 Id. at 1464.
147 Id.
148 Id. at 1460 (contending “that there is no general right to use ARTs as a matter of reproductive autonomy, but there may be a limited right to use ARTs as a matter of reproductive equality”).
149 Glucksberg, 521 U.S. at 721 (“[T]he Fourteenth Amendment ‘forbids the government to infringe . . . fundamental liberty interests at all, no matter what process is provided, unless the infringement is narrowly tailored to serve a compelling state interest.’” (quoting Reno v. Flores, 507 U.S. 292, 302 (1993))).
150 Professor Robertson enumerates the following interests: “[D]estruction of embryos and fetuses, harm to offspring, instrumentalizing or commodifying human life, discrimination on the basis of gender or disability, and easing the way to nonmedical enhancement.” Robertson, supra note 122, at 429.
151 See, e.g., U.S. CONST. amend. XIII (abolishing slavery); U.S. CONST. amend. XIV (guaranteeing, among other things, equal protection of the laws); U.S. CONST. amend. XV
Fourteenth Amendment’s Equal Protection Clause,\textsuperscript{152} which has been used to curtail discrimination based on (among other traits) race, sex, and disability.\textsuperscript{153} Notwithstanding the fact that courts scrutinize race discrimination more rigorously than sex discrimination, and sex discrimination more rigorously than disability discrimination, purely animus-based discrimination is always unconstitutional.\textsuperscript{154} And, while the Equal Protection Clause itself applies only to government discrimination,\textsuperscript{155} Section 5 of the Fourteenth Amendment, which gives Congress power to enforce the Equal Protection Clause, has been used (in conjunction with the Commerce Clause\textsuperscript{156}) as a ground for banning private discrimination based on race, sex, and disability.\textsuperscript{157} One could, therefore, argue that preventing discrimination on these bases ought to be accepted as the type of compelling government interest that could justify infringement on a fundamental right, such as the arguably fundamental right to nontherapeutic PGD. While the Court has not always accepted the creation of social equality as a compelling interest,\textsuperscript{158} one

\textsuperscript{152} U.S. CONST. amend. XIV, §1 (“No state shall . . . deny to any person within its jurisdiction the equal protection of the laws.”).

\textsuperscript{153} It was used to curtail race discrimination in \textit{Loving v. Virginia}, 388 U.S. 1, 11 (1967) (applying strict scrutiny to race-based classifications); sex discrimination in \textit{United States v. Virginia}, 518 U.S. 515, 533–34 (1996) (applying intermediate or skeptical scrutiny to gender-based classifications); and disability discrimination in \textit{Cleburne v. Cleburne Living Center}, 473 U.S. 432, 442, 450 (1985) (applying only rational basis review to disability-based classifications, but invalidating discrimination against individuals with disabilities on the basis that it was irrational).

\textsuperscript{154} Romer v. Evans, 517 U.S. 620, 632 (1996) (“[T]he amendment [at issue] seems inexplicable by anything but animus toward the class it affects; it lacks a rational relationship to legitimate state interests.”); U.S. Dep’t of Agric. v. Moreno, 413 U.S. 528, 534 (1973) (“[I]f the constitutional conception of ‘equal protection of the laws’ means anything, it must at the very least mean that a bare . . . desire to harm a politically unpopular group cannot constitute a legitimate governmental interest.”).

\textsuperscript{155} The Civil Rights Cases, 109 U.S. 3, 10–11 (1883) (“Individual invasion of individual rights is not the subject-matter of the [Equal Protection Clause].”).

\textsuperscript{156} U.S. CONST. art. I, § 8, cl. 3.


\textsuperscript{158} See, e.g., Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 723–25 (2007); see also Osamudia R. James, \textit{Dog Wags Tail: The Continuing Viability of
could certainly argue that it should. The remainder of this section explores three of the most common arguments from equality, all of which view banning nontherapeutic PGD as a narrowly tailored means of achieving a compelling interest in creating social equality.

The first argument from equality arises from concerns about race- and sex-based marginalization. If nontherapeutic PGD were consistently used to select against traits associated with racial minorities or women, those groups might be marginalized. Proponents of this argument often make the secondary argument that selecting against certain racial markers or a certain sex could also create an imbalance in the overall population. Because this secondary argument is more properly framed as a diversity argument than an equality argument, it will be addressed in Part III.B.

Concerns about race-based marginalization are rarely voiced, except in conjunction with concerns that racial minorities are disproportionately unable to afford the high costs of IVF and PGD. Yet nontherapeutic PGD could lead to race-based marginalization for reasons entirely unrelated to socioeconomic status. If ART clinics started allowing prospective parents to select for eye, hair, and skin color, racial markers could be emphasized or deemphasized. While, as previously mentioned, prospective parents using only their own genetic material could not select children of races entirely different from their own, they could select for lighter eyes, hair, or skin, and thereby affect the apparent racial makeup of the population. Whether prospective parents would actually make such selections is unknown, since clinics do not currently give them the opportunity—yet reports from sperm banks and adoption agencies, which will be discussed in Part III.B, suggest that they would not. Indeed, many prospective parents who use sperm banks or adoption agencies want children who resemble themselves.

In contrast, concerns that nontherapeutic PGD could marginalize women are regularly voiced. Professor Amartya Sen has warned that sex selection produces

---

Minority-Targeted Aid in Higher Education, 85 IND. L.J. 851, 881 & n.249 (2010) (explaining that “the [Parents Involved] Court implicitly reaffirmed earlier pronouncements that attempts to address the manifestations of societal discrimination are not compelling interests that justify affirmative action” and that, ultimately, “the Court failed to clearly address the legitimacy of social equality as a compelling interest”).


160 Such concerns over unequal access to ARTs will be discussed later in this section.

161 See supra Part II.A.2.

162 Id.

163 See infra Part III.B.

164 Id.
natality inequality,” a phenomenon that is particularly pronounced in male-dominated nations such as China, South Korea, and India. The President’s Council on Bioethics, in its 2003 report, Beyond Therapy: Biotechnology and the Pursuit of Happiness, expressed similar concerns about sex selection. It explained that, if the practice is allowed, “the private choices made by individuals, once aggregated, could produce major changes in a society’s sex ratio . . . .” It elaborated, “Over the past several decades, disturbing evidence has accumulated of the widespread use of various medical technologies to choose the sex of one’s child, with a strong preference for the male sex.” While the Council acknowledged that this preference for males does not currently exist within the United States, where the overall sex ratio remains at 104.8 (very close to the natural sex ratio of 105), it worried about rising sex ratios among certain groups—specifically, Chinese and Japanese Americans. While it is unclear whether prospective parents would consistently select for males—and some evidence, which will be discussed in Part III.B, suggests they would not—it is clear that such selections would be problematic.

The second equality-based argument in favor of banning nontherapeutic PGD arises from concerns about stereotyping. If prospective parents were allowed to engage in trait selection, they might force their children to conform to (or defy) the social stereotypes associated with the traits they have selected. Prospective parents who select males may, for example, expect conventionally “masculine” behavior. Policymakers, scientists, and academics from across the globe have voiced such concerns. As the President’s Council on Bioethics stated, “[A] host of powerful expectations go into the selection of a boy or a girl[, and, in] choosing one sex over the other, we are necessarily making a statement about what we expect of that child . . . .” Doctor David King, Director of Human Genetics Alert (a London-based watchdog organization), similarly stated that children who have been selected from multiple embryos “are likely to feel that the essence of themselves,

---

167 Id. at 58.
168 Id.
169 Id. at 58, 61 (observing that “[t]he natural sex ratio at birth is 105 baby boys born for every 100 baby girls,” and that “the sex ratio in the United States has remained stable at 104.8”).
170 Id. at 61 (“In 1984 the sex ratio for Chinese-Americans was 104.6 and for Japanese Americans 102.6; in 2000, these ratios had risen respectively to 107.7 and 106.4.”).
171 Id. at 70 (“As fathers, we may want a son to go fishing with; or as mothers, we may want a daughter to dress for the prom. The problem goes deeper than sexual stereotyping, however. For it could also be the case that we may want a daughter who will become president to show that women are the equal of men.”).
in an important sense no longer belongs solely to them, [and that they are merely expressions of] their parents’ aspirations, desires and whims.”¹⁷² Doctor Ruth Zafran, a law lecturer in Israel, wrote, “[T]he more the parents’ preferences for a child’s genetic makeup are met through active, external, and calculated intervention, the greater their expectation will be for that child to fulfill the ‘genetic promise,’ namely to lead a life consistent with what the genes are supposedly intended to achieve.”¹⁷³ Concerns about stereotyping, indeed, merit substantial consideration.

The third equality-based argument in favor of banning nontherapeutic PGD arises from concerns about unequal access. Unless and until nontherapeutic PGD becomes more affordable, only the wealthy will realistically have access. At present, IVF alone costs between $10,000 and $12,000 per cycle, and PGD adds an additional $2,500 to $7,000.¹⁷⁴ When interviewed about the expected cost of eye-, hair-, and skin-color selection, Doctor Jeff Feinberg of the Fertility Institutes estimated that it would be approximately $18,000.¹⁷⁵ Since IVF is generally not covered by insurance (except in a handful of states that mandate coverage),¹⁷⁶ most patients pay out-of-pocket for the procedures. Those who use IVF and PGD, thus, tend to be affluent and, in many cases, well educated and politically powerful.¹⁷⁷ The American Medical Association’s Council on Ethical and Judicial Affairs has spoken about unequal access to genetic enhancement—which, while different from genetic selection, raises similar concerns.¹⁷⁸ Genetic enhancement is, according to


¹⁷³ See Zafran, supra note 6, at 204 (“This may cover both physical characteristics and personality traits.”). Doctor Ruth Zafran is a lecturer at the Radzyner School of Law in Israel.


¹⁷⁵ Sherwell, supra note 85; see discussion supra Part II.A.2.

¹⁷⁶ King, supra note 174, at 296–97 (“While a handful of states require insurance companies to cover all or a portion of the costs associated with IVF, a substantial percentage of IVF patients remain uncovered by insurance and are forced to pay for the procedure out of pocket.”); see also June Carbone & Naomi Cahn, Embryo Fundamentalism, 18 WM. & MARY BILL RTS. J. 1015, 1036 & n.137 (2010) (citing The Nat’n Infertility Ass’n, Insurance Coverage in Your State, RESOLVE, http://www.resolve.org/family-building-options/insurance_coverage/state-coverage.html (last visited Jan 21, 2013) (observing that since the 1990s, only fifteen states have required insurance to cover infertility treatment).

¹⁷⁷ Carbone & Cahn, supra note 176, at 1018.

the Council’s 1994 report, “permissible only in severely restricted situations,” and not unless all prospective parents have “equal access . . . irrespective of [their] income or other socioeconomic characteristics.”

Scholars, cognizant of the rough correlation between socioeconomic status and race, have observed that the high costs of IVF and PGD tend to perpetuate existing privilege by “allowing wealthy white women to reproduce themselves.” Professor Dorothy Roberts, who has written extensively on these issues, observes that new reproductive technologies such as PGD are used “almost exclusively by white people” and are avoided disproportionately by African Americans. Professor Roberts has observed a “harsh dichotomy,” in which “policies punish poor black women for bearing children but advanced technologies assist mainly affluent white women not only to have genetically related children, but to have children with preferred genetic qualities.” Concerns about socioeconomic inequality are thus problematic not only in their own right but also in their overlap with concerns about racial marginalization.

These dominant arguments from equality, which suggest that a ban on nontherapeutic PGD would be justified, are in direct conflict with the dominant arguments from liberty, which suggest that such a ban would be unconstitutional. While there is no easy answer to this dilemma, Part III will illustrate that the dominant arguments discussed above do not necessarily represent the full spectrum of plausible liberty and equality arguments.

III. DECONSTRUCTING THE LIBERTY/EQUALITY BINARY IN NONTHERAPEUTIC PGD

The arguments discussed in Part II suggest that, in assessing the constitutionality of a ban on nontherapeutic PGD, one must make a choice between protecting either liberty or equality. They suggest that protecting liberty requires destroying equality, and protecting equality requires destroying liberty. Part III begins from the premise that this is a false—and, indeed, constitutionally problematic—choice, inasmuch as Section 1 of the Fourteenth Amendment protects both liberty and equality. Part III.A, drawing on the postmodern literary theory discussed in Part I, illustrates that the arguments set forth in Part II do not

non-disease-related traits would be inappropriate.” *Id.* at 638; see also Mehlman, *supra* note 138, at 120 (discussing the Council’s report).


183 *See supra* Part II.B.1.
represent the full spectrum of plausible arguments from liberty and equality. Part III.B, drawing on the postmodern political theory discussed in Part I, suggests that, when faced with multiple plausible arguments from liberty and equality, we should choose those that are most conducive to revealing the diverse identities that currently populate our public forum. The Article concludes by arguing that, in so doing, we will promote robust debate, accurate representation, and ultimately, democracy.

A. Multiple Arguments from Liberty and Equality

This Part draws on the postmodern literary theory introduced in Part I to illustrate that the dominant liberty and equality arguments set forth in Part II.B are not the only plausible liberty or equality arguments. In addition to the dominant liberty arguments, which oppose banning nontherapeutic PGD, there are other currently marginalized liberty arguments that favor banning nontherapeutic PGD. Similarly, in addition to the dominant equality arguments, which favor banning PGD, there are other currently marginalized equality arguments that oppose banning nontherapeutic PGD. While the dominant arguments set forth in Part II.B position liberty and equality as binary opposites, the arguments set forth below deconstruct the binary opposition and thereby give us the option of repositioning the two rights as fully compatible partners.

1. Liberty Arguments in Favor of Banning Nontherapeutic PGD

Liberty, as discussed in Part I, is susceptible to multiple interpretations. While the dominant liberty arguments discussed in Part II.B.1 focus on the rights of prospective parents to make decisions about procreation and childrearing—and, accordingly, protect the use of nontherapeutic PGD—other currently marginalized liberty arguments focus on the rights of prospective children to construct their own identities—and, accordingly, condemn the use of nontherapeutic PGD. These marginalized arguments might be framed as asserting that, even if a ban on nontherapeutic PGD did burden the liberty of prospective parents, the burden would be justified because the ban would be narrowly tailored to achieve the compelling interest in protecting the liberty of prospective children.184

While the Supreme Court has never held that children (or, for that matter, adults) have a right to fully autonomous identity construction, several decisions suggest that children have at least a limited right to construct their own identities. The 2003 decision of Lawrence v. Texas established that adults have the right “to define [their] own concept of existence, of meaning, of the universe, and of the

184 Proponents of such arguments would, however, likely begin by arguing that a ban on nontherapeutic PGD would not burden any fundamental liberty of prospective parents.
mystery of human life,"\textsuperscript{185} and that the liberty protected by the Due Process Clauses “presumes an autonomy of self that includes freedom of thought, belief, expression, and certain intimate conduct.”\textsuperscript{186} While the Lawrence Court was careful to note that the facts before it did not involve minors,\textsuperscript{187} the 1979 decision of Bellotti v. Baird\textsuperscript{188} established that minors enjoy at least limited liberty rights, including the right to make decisions about procreation.\textsuperscript{189} The rights of minors are, however, far weaker than—and often superseded by—the rights of their parents.\textsuperscript{190} The Bellotti Court stated, “[T]he guiding role of parents in the upbringing of their children justifies limitations on the freedoms of minors.”\textsuperscript{191} Yet it nevertheless held that minors seeking abortions are entitled to judicially bypass parental consent requirements.\textsuperscript{192} Based on decisions like Lawrence and Bellotti, minors arguably have the right to construct their own identities. That right is limited (at least under existing doctrine and dominant understandings), however, by the superior rights of parents to make decisions about their children’s upbringing.

Professor Dena Davis’s work has been instrumental in revealing previously marginalized liberty arguments that protect children’s rights to construct their own identities. Her seminal article, Genetic Dilemmas and the Child’s Right to an Open Future, shifts the focus from the liberty of prospective parents to the liberty of prospective children—which she (following Professor Joel Feinberg) refers to as the child’s “right to an open future.”\textsuperscript{193} Professor Davis argues that prospective parents who use nontherapeutic PGD inevitably deprive their children of liberty by forcing them to live out (or, in some cases, deliberately defy) the stereotypes

\begin{footnotesize}

\textsuperscript{186} Id. at 562.

\textsuperscript{187} Id. at 578 (“The present case does not involve minors.”).

\textsuperscript{188} 443 U.S. 622 (1979).

\textsuperscript{189} See id. at 643 (explaining that parental consent laws are constitutionally permissible so long as they are accompanied by sufficient judicial bypass proceedings). The Bellotti Court stated, “A child, merely on account of his minority, is not beyond the protection of the Constitution.” Id. at 633.

\textsuperscript{190} See id. at 633–34.

\textsuperscript{191} Id. at 637. The Bellotti Court cited Pierce v. Society of Sisters and Wisconsin v. Yoder, both of which are discussed supra Part II.B.1. The Court “recognized three reasons justifying the conclusion that the constitutional rights of children cannot be equated with those of adults: the peculiar vulnerability of children; their inability to make critical decisions in an informed, mature manner; and the importance of the parental role in child rearing.” Id. at 634. It stated that “the tradition of parental authority is not inconsistent with our tradition of individual liberty . . . .” Id. at 638.

\textsuperscript{192} Id. at 643.

\end{footnotesize}
associated with their selected traits. Professor Davis writes, “Parents whose preferences for one sex or the other are compelling enough for them to take active steps to control the outcome must, logically, be committed to certain strong gender-role expectations of the children they will raise.” This often marginalized liberty argument is quite consistent with the dominant equality argument that parents who engage in nontherapeutic trait selection will force their children to conform to the social stereotypes associated with their selected traits. By reframing a dominant equality argument as a previously unarticulated liberty argument, Professor Davis’s article helps to deconstruct the liberty/equality binary.

2. Equality Arguments Against Banning Nontherapeutic PGD

Equality, like liberty, is susceptible to multiple interpretations. Equality arguments can, therefore, weigh both in favor of and against bans on nontherapeutic PGD. While the dominant equality arguments discussed in Part II.B.2 weigh in favor of bans on nontherapeutic PGD, there are at least two currently marginalized equality arguments that weigh against such bans: the first attacks discrimination between those who wish to make nontherapeutic trait selections and those who wish to make therapeutic trait selections; the second attacks discrimination between those who wish to make nontherapeutic trait selections and those who wish to make selections from banks of donor sperm or ova. While such forms of discrimination trigger only rational basis review under the Equal Protection Clause, this section asserts that they cannot survive even that lowest level of scrutiny. After elaborating on these two currently marginalized arguments, this section concludes by offering a possible rebuttal to one of the dominant equality arguments: It suggests that the socioeconomic discrimination resulting from nontherapeutic PGD could be addressed either by banning nontherapeutic PGD as discussed in Part II.B.2 or by affirmatively subsidizing nontherapeutic PGD.

The first marginalized equality argument, which attacks discrimination between those who wish to make nontherapeutic trait selections and those who wish to make therapeutic trait selections, posits that such discrimination is irrational due to the inherent moral instability of the therapeutic/nontherapeutic binary. Inasmuch as it is, without moral judgment, difficult to definitively characterize certain trait selections as therapeutic or nontherapeutic—particularly when racial markers, sex, or disabilities are involved—a ban that targets

---

194 Id. at 587–91.
195 Id. at 587. She continues, “If they want a girl that badly, whether they are hoping for a Miss America or the next Catherine MacKinnon, they are likely to make it difficult for the actual child to resist their expectations and follow her own bent.” Id.
196 See supra Part II.B.2. It should be noted that framing the argument as a liberty argument rather than an equality argument may have implications for abortion politics.
nontherapeutic PGD but allows therapeutic PGD could not survive rational basis review.\footnote{While many would say that therapeutic selections are made for medical reasons and nontherapeutic selections are made for nonmedical reasons, defining the terms therapeutic and nontherapeutic is difficult. Indeed, as Professor Brigham Fordham has observed, the terms “have meaning only by reference to some subjective standard for what makes a person whole.” Brigham A. Fordham, Disability and Designer Babies, 45 VAL. U. L. REV. 1473, 1517 (2011). Professor Leonardo De Castro has defined therapeutic experiments as those in which “the benefits are intended to accrue directly to the subjects,” and nontherapeutic experiments as those in which “the benefits are expected to be enjoyed by others.” Leonardo D. De Castro, Ethical Issues in Human Experimentation, in A Companion to Bioethics 380, 380 (Helga Kuhse & Peter Singer eds., 2001). He has, however, recognized that “[o]ne and the same procedure may be therapeutic to some subjects but nontherapeutic to others.” Id.}

The therapeutic/nontherapeutic binary is, in many ways, similar to the “therapy-versus-enhancement distinction” discussed in the report of the President’s Council on Bioethics.\footnote{PRESIDENT’S COUNCIL ON BIOETHICS, supra note 166, at 17.} The report defines therapy as “the use of biotechnical power to treat individuals with known diseases, disabilities, or impairments, in an attempt to restore them to a normal state of health and fitness.”\footnote{Id. at 13.} It defines enhancement as the “use of biotechnical power to alter . . . the ‘normal’ workings of the human body and psyche, to augment or improve their native capacities and performances.”\footnote{Id.} Yet the report recognizes the limitations of the distinction, noting that “‘therapy’ and ‘enhancement’ are overlapping categories,” both of which are “bound up with . . . the inherently complicated idea of health and the always-controversial idea of normality.”\footnote{Id. at 14–15.} It explains, “While in some cases—for instance, a chronic disease or a serious injury—it is fairly easy to point to a departure from the standard of health, other cases defy simple classification,” in part, because “[m]ost human capacities fall along a continuum . . . .”\footnote{Id. at 15–16.} It asks, “Is it therapy to give growth hormone to a genetic dwarf, but not to a short fellow who is just unhappy to be short?”\footnote{Id. at 16. (“And if the short are brought up to the average, the average, now having become short, will have precedent for a claim to growth hormone injections.”).} This question has broad implications, inasmuch as many trait selections could plausibly have both therapeutic and nontherapeutic aspects.

Prospective parents might, for example, select darker or lighter skin for either therapeutic or nontherapeutic reasons. Selecting darker skin to reduce the risk of
skin cancer\textsuperscript{204} or selecting lighter skin to facilitate the synthesis of vitamin D\textsuperscript{205} could arguably be viewed as therapeutic. In contrast, selecting darker or lighter skin to create a particular social status would conventionally be viewed as nontherapeutic. Yet challenges to this conventional view exist. As Rabbi Mark Popovsky has observed, having a certain skin color may, in certain social contexts, confer such significant “advantages that [it] could potentially affect someone’s overall health and well-being far more than the presence or absence of one specific genetic disease.”\textsuperscript{206} If this is true, then selecting a certain skin color to create a particular social status could be characterized as at least partly therapeutic. Furthermore, it is possible that prospective parents might choose a certain skin color for a combination of reasons, some conventionally viewed as therapeutic and others conventionally viewed as nontherapeutic—or they might articulate a therapeutic reason as a pretext to achieve a nontherapeutic goal. In sum, absent moral judgment, characterizing skin-color selection as definitively therapeutic or nontherapeutic is problematic.

A similar analysis can be applied to sex selection. Prospective parents might select a female or male child for either conventionally therapeutic or conventionally nontherapeutic reasons. Choosing a female child to avoid an X-chromosome-linked disease would conventionally be viewed as therapeutic, while choosing a female child to “balance” one’s family or create a certain social status would conventionally be viewed as nontherapeutic. Yet, again building on Rabbi Popovsky’s observation, one could argue that being of a certain sex might, in certain social contexts (e.g., a family that desperately wanted a child of a certain sex or a society that consistently privileged members of a certain sex), confer such significant advantages that being of that sex would improve a child’s health far more than the absence of any particular disease. Characterizing sex selection as definitively therapeutic or nontherapeutic is thus, absent moral judgment, also problematic.


\textsuperscript{205} Nina G. Jablonski & George Chaplin, \textit{The Evolution of Human Skin Coloration}, 39 J. HUM. EVOLUTION 57, 58 (2000) (“The more lightly pigmented skins of peoples inhabiting latitudes nearer the Arctic have been explained as adaptations to the lower UV radiation regimes of those regions and the importance of maintaining UV-induced biosynthesis of vitamin D\textsubscript{3} in the skin.” (citations omitted)).

\textsuperscript{206} Mark Popovsky, \textit{Jewish Perspectives on the Use of Preimplantation Genetic Diagnosis}, 35 J.L. MED. & ETHICS 699, 709 (2007) (“[O]ne can argue convincingly that the line between traits affecting health and those that do not is impossible to draw. For example, the advantages that may follow in our society from a lighter shade of skin color or from being straight could potentially affect someone’s overall health and well-being far more than the presence or absence of one specific genetic disease.”).
While similar arguments can illustrate the problems of characterizing selections related to skin color and sex, a different set of arguments is necessary to illustrate the problems of characterizing selections related to disability. While selecting against disability is relatively common and generally viewed as therapeutic, selecting for disability is relatively uncommon and generally viewed as nontherapeutic—as well as unethical, according to some. Yet defining the term disability can in itself be quite problematic, since many prospective parents who wish to select for traits conventionally viewed as disabilities would not themselves characterize those traits as disabilities. When parents of a child with Down syndrome wish to select for Down syndrome,207 when achondroplasic parents wish to select for achondroplasia,208 or when deaf parents wish to select for deafness,209 they may not necessarily characterize themselves as selecting for disabilities. A deaf father, when questioned about his desire to select a deaf child, explained: “Being deaf is not about being disabled. It’s about being part of a linguistic minority. We’re proud of the language we use and the community we live in.”210 For such parents, selecting for deafness is not a means of imposing a disability but of sharing an identity and ensuring the continued presence of that identity in the general population. Due to the problems with defining the term disability, it is again difficult, absent moral judgment, to determine whether selecting for or against a trait conventionally viewed as a disability is definitively therapeutic or nontherapeutic.211

One could argue, for the above reasons, that a ban exclusively targeting nontherapeutic PGD could not survive rational basis review, inasmuch as it could not be rationally related to any legitimate government interest.212 While the government may have legitimate interests in both protecting prospective children and preserving the existing gene pool, these interests—when effectuated through a ban exclusively targeting nontherapeutic PGD—rest upon moral judgments. While, under rational basis review, the government is permitted to solve problems incrementally,213 it is not permitted to rest legislation solely upon moral

207 See Smolensky, supra note 61, at 304.
208 See id. at 305.
209 See Gray, supra note 47.
210 Id.
211 Even assuming that we could reach general consensus on the list of disabilities, the fact that selecting against disabilities is common and selecting for disabilities is uncommon creates inequality. This argument will be discussed below, in Part III.B.
213 See Williamson v. Lee Optical of Okla., Inc., 348 U.S. 483, 489 (1955) (“[T]he reform may take one step at a time, addressing itself to the phase of the problem which seems most acute to the legislative mind. The legislature may select one phase of one field and apply a remedy there, neglecting the others.” (citations omitted)).
And a ban exclusively targeting nontherapeutic PGD would rest on such judgments inasmuch as it would protect prospective children based on a particular moral judgment—that what has conventionally been viewed as nontherapeutic is less beneficial—and preserve the gene pool by singling out nontherapeutic PGD as an immoral means of altering that pool. Somewhat similar morality-based interests have, on prior occasions, failed rational basis review. Finally, though it is not the focus of this Article, it should be noted that the inability to definitively characterize a given use of PGD as therapeutic or nontherapeutic could also render a ban exclusively targeting nontherapeutic PGD unconstitutionally vague under the Due Process Clause.

If the first marginalized equality argument were unpersuasive, the second could be raised to attack discrimination between those who wish to make selections via nontherapeutic PGD (or, for that matter, any PGD) and those who wish to make selections from banks of donor sperm or ova. This argument posits that discrimination between nontherapeutic PGD users and sperm or ova bank users is irrational because any danger of skewing the gene pool that arises from PGD is far less than that which arises from the use of donor sperm or ova, and the latter is currently allowed with very few restrictions. Thus, a ban targeting PGD

---

214 See Lawrence v. Texas, 529 U.S. 558, 578 (2003) (invalidating a morality-based ban on same-sex sodomy by stating that it “further[ed] no legitimate state interest”). Justice Scalia, indeed, in his dissent, read the majority opinion as holding that morality cannot suffice as a legitimate state interest. Id. at 599 (Scalia, J., dissenting) (lamenting that the majority opinion “effectively decrees the end of all morals legislation,” and characterizing the majority opinion as asserting that “the promotion of majoritarian sexual morality is not even a legitimate state interest”). But see Gonzales v. Carhart, 550 U.S. 124, 158 (2007) (upholding an abortion ban, in part, on the basis that “Congress could . . . conclude that the type of abortion proscribed by the Act requires specific regulation because it implicates additional ethical and moral concerns that justify a special prohibition”).

215 See, e.g., Romer, 517 U.S. 620 (invalidating an animus-based law under rational basis review); City of Cleburne v. Cleburne Living Ctr., Inc., 473 U.S. 432, 449–50 (1985) (invalidating a local land use decision under rational basis review because it rested on irrational prejudice toward the mentally disabled); U.S. Dep’t of Agricul. v. Moreno, 413 U.S. 528, 534 (1973) (striking down an amendment to the federal Food Stamp Act defining the term “household” so as to include only groups of related persons under rational basis review on grounds that “a bare congressional desire to harm a politically unpopular group cannot constitute a legitimate governmental interest”).

216 See, e.g., Lifchez v. Hartigan, 735 F. Supp. 1361, 1363–64 (N.D. Ill. 1990) (challenging a law that banned experimentation unless it was therapeutic to the fetus on vagueness grounds).

217 The distinction between nontherapeutic and therapeutic uses is less relevant within the context of this argument.

218 Michelle Dennison, Revealing Your Sources: The Case for Non-Anonymous Gamete Donation, 21 J.L. & HEALTH 1, 15 (2008) (“[T]here is very little federal or state regulation in the donor industry.”). Dennison explains, “[E]xisting laws almost exclusively center around the parentage of children born through ART methods, providing in almost all
while continuing to allow the unrestricted selection of donor sperm and ova could not survive rational basis review.\footnote{While the government may act incrementally to solve a problem, see \textit{Williamson v. Lee Optical of Okla., Inc.}, 348 U.S. 483, 489 (1955), targeting an unlikely danger while continuing to allow a likely danger is irrational.} 

Sperm and ova banks may, but are not required to, limit the number of times a donor can donate or the number of clients who can receive the same donor’s sperm.\footnote{Dennison, \textit{supra} note 218, at 15.} Many clinics, however, report “most-requested” donors,\footnote{Id.} and at least one clinic has allowed a single donor to genetically father over 150 children, potentially all in the same geographic area.\footnote{Id. at 10. But see 42 U.S.C. § 263a-1 (2006).} English philosopher Mary Warnock, known for her contributions to the \textit{Warnock Report on Human Fertilization and Embryology}, has stated, “It is quite unpredictable what the ultimate effect on the gene pool of a society might be if donors were permitted to donate as many times as they chose . . . .”\footnote{Id.} In addition to the possibility of single donors skewing the gene pool, consistent preferences in favor of donors of a certain race could also skew the gene pool. Yet many clinics facilitate the race-based selection of donors.\footnote{Jacqueline Mroz, \textit{One Sperm Donor, 150 Offspring}, N.Y. TIMES, Sept. 6, 2011, at D1 (late ed.), available at http://www.nytimes.com/2011/09/06/health/06donor.html?pagewanted=all.} \footnote{Dov Fox, \textit{Choosing Your Child’s Race}, 22 HASTINGS WOMEN’S L.J. 3, 3–4 (2011) (“[T]wenty-three of the twenty-eight sperm banks currently operating in the United States provide aspiring parents with the sperm donors’ self-reported racial identity. The largest among these—including the world’s leading sperm bank, California Cryobank—organize catalogs into separate sections on the basis of donor race.” (citations omitted)); \textit{see also} Fox, \textit{supra} note 44, at 1849–50 (discussing how California Cryobank, Inc., the world’s leading sperm bank, “includes donor information across a wide range of traits, including height, weight, education, occupation, religion, ethnic origin, facial features, eye and hair color, hair texture, skin tone, and race”).} If this drive to self-replicate diminishes concerns about the use of donor material skewing the gene pool, it must further diminish concerns about the use of PGD skewing the gene pool. The possibility of skewing is, indeed, much greater when prospective parents select from large banks of donor material than when they select from among several embryos created from one male’s sperm and one female’s ova. While prospective parents might, of course, use donor sperm and ova to create several embryos and then use PGD to select among those embryos, the greatest danger of skewing the gene pool derives
from the selection of the donor material rather than from the selection among embryos. It is, therefore, irrational to ban PGD but allow the unrestricted use of donor material.

One might relatedly observe that it is irrational to impose more restrictions on those who wish to use nontherapeutic PGD than on those who wish to undergo nontherapeutic abortions, which are currently permitted under certain circumstances. Prospective parents who lack access to nontherapeutic PGD—either because a clinic has deemed their desired trait selection inappropriate or because they cannot afford the high costs associated with IVF and PGD—may become pregnant through other means, undergo prenatal testing such as chorionic villus sampling or amniocentesis, and if they dislike the test results elect to undergo a nontherapeutic abortion. If such prospective parents were given affordable access to nontherapeutic PGD, abortion rates could decrease.

Finally, according to one of the dominant equality arguments, nontherapeutic PGD is undesirable because it exacerbates existing socioeconomic inequality. Because IVF and PGD are so costly, only the wealthy can realistically afford them—which, due to the continuing correlation between wealth and race, means that whites are more likely than African Americans to utilize them. It should be emphasized, however, that concerns about socioeconomic inequality could be addressed either by banning nontherapeutic PGD or by subsidizing it. Professor Maxwell Mehlman—who has warned that lack of equal access to the technology necessary for genetic enhancement (which, again, raises concerns similar to those associated with genetic selection) could create a “genobility”—has argued in favor of government subsidies. He has written, “Everyone should be able to access technologies that significantly enhance human capabilities, and... the government should subsidize access if need be for those less well-off.” While the Supreme Court has consistently declined to require subsidies for the exercise of even fundamental rights, legislatures could elect to provide them. Subsidizing

226 See Gonzales v. Carhart, 550 U.S. 124, 146 (2007) (“Before viability, a State ‘may not prohibit any woman from making the ultimate decision to terminate her pregnancy.’ It also may not impose upon this right an undue burden, which exists if a regulation’s ‘purpose or effect is to place a substantial obstacle in the path of a woman seeking an abortion before the fetus attains viability.’” (citations omitted)).

227 Consider, for example, the achondroplasic parents discussed supra Part II.A.2.b.

228 See supra Part II.B.2.

229 Mehlman, supra note 138, at 120 (citing MAXWELL J. MEHLMAN & JEFFREY BOTKIN, ACCESS TO THE GENOME: THE CHALLENGE TO EQUALITY 99 (1998)).

230 Id.

231 Id. Professor Mehlman adds the somewhat radical sentiment that, “This should apply as well to directed evolutionary techniques that parents employ to give their children significant social advantages.” Id.

IVF and PGD may not be realistic in the current economy, but the possibility of such subsidies illustrates that the dominant wealth-based arguments do not lead inevitably to banning nontherapeutic PGD.

B. Interpreting Liberty and Equality to Promote Diversity

Just as postmodern literary theory can help us to identify multiple meanings of liberty and equality, postmodern political theory can help us to select among those meanings. As described in Part I.A.2, postmodern political theory reveals the diverse array of identities that always already exist within our public forum. Rather than reifying the traditionally dominant and marginalized identity categories—for example, white/black, male/female, abled/disabled—postmodern political theory exposes their inherent fallacy. It illustrates that identity is never so singular or so stable that it can be definitively categorized and that identity-based hierarchies are never so simplistic that the terms “dominant” and “marginalized” can serve as adequate descriptors. An individual is never merely white, male, or nondisabled, and an identity group is never absolutely dominant or absolutely marginalized. Individual identity is multifaceted, and dominant identity groups depend for their continued dominance (and, indeed, for their existence) upon multiple marginalized identity groups. On both individual and societal levels, then, postmodern political theory reveals that identity categories are always incomplete. They obscure and, in some cases, even diminish existing diversity. Yet democracy can succeed only when diverse identities are both expressed and represented. Debates are rarely robust when diverse opinions are not expressed, and policies are rarely just when diverse identities have not been represented in their formulation. Ensuring that the diverse identities that populate our public forum are expressed and represented is, in short, crucial to the success of our democracy. This Part thus assesses the full spectrum of liberty and equality arguments and attempts to determine those most conducive to recognizing existing diversity.

The liberty and equality arguments set forth in Parts II.B and III.A suggest at least three possible policy outcomes. First, one might conclude that nontherapeutic PGD should be banned because allowing it could deprive racial minorities and women of equality and prospective children of liberty. Second, one might conclude that nontherapeutic PGD should be allowed because banning it could deprive prospective parents of liberty to make decisions about procreation and childrearing and of equality vis-à-vis therapeutic PGD users and sperm or ova bank users. Third, one might conclude that nontherapeutic PGD should be affirmatively

---

233 Binary identity categories can diminish existing diversity by causing people to conform to social norms and fit into one of the two common categories rather than expressing their own unique identities. See Jessica Knouse, From Identity Politics to Ideology Politics, 2009 UTAH L. REV. 749, 761–64 (observing, in Part II.D, that understanding individuals solely as members of identity groups forces conformity and prevents fully autonomous self-determination).

234 See supra Part I.A.2.
subsidized because allowing it without subsidy could deprive the poor (and, by extension, the racial minorities that are most likely to experience poverty) of equality. This section will assess how each of these policies could enhance or impede the expression of diverse identities and, accordingly, benefit or harm our democracy.

Banning nontherapeutic PGD could potentially diminish diversity. While there is little reason to think that it would significantly diminish genetic diversity—except insofar as it might diminish the incidence of certain disabilities—there is ample reason to think that it would diminish ideological diversity. The dominant concerns about genetic diversity relate to the marginalization of racial minorities and women.235 As discussed in Part II.B.2, one might argue that nontherapeutic PGD should be banned to prevent prospective parents from consistently selecting against certain racial markers—for example, dark eyes, dark hair, and dark skin—and against females. Such selections would, indeed, create an imbalance in the overall population and diminish genetic diversity. Yet there is little evidence to suggest that prospective parents would, if given the opportunity, make such selections.

With respect to racial markers, no clinics presently allow selection on the basis of eye, hair, or skin color. And if they started allowing such selections, there is little other than prevailing norms about beauty and attractiveness—which are, of course, far from monolithic—to suggest that a significant proportion of prospective parents would actually use nontherapeutic PGD to consistently select against children with dark eyes, hair, and skin. Indeed, reports from sperm banks and adoption agencies suggest that prospective parents would not make such selections. While the considerations involved in selecting among embryos may differ in relevant respects from those involved in selecting among sperm donors or adoptive children, they are the best evidence we currently have of prospective parents’ likely preferences. With respect to sperm banks, Ole Schou of Cryos International has reported that his clients tend to favor donors who resemble themselves.236 And, although Cappy Rothman of California Cryobank reported in 2001 that his clients tend to favor donors with brown eyes and blond hair,237 this

235 See supra Part II.B.2.
236 The World: No Redheads or Blondes Wanted at Sperm Bank (PRI radio broadcast Sept. 20, 2011), available at http://www.theworld.org/2011/09/redheads-sperm-bank/; see also Fox, supra note 224, at 4–5 (noting that many prospective parents select donors who have racial markers similar to their own). In an interview by Lisa Mullins, Ole Schou said, “The heterosexual couples always look for a donor who could look like the male partner. If it’s lesbians, they try to match as close as possible to one of the females. And then we have the single segmentation, they do not have a husband or partner to match it, so they probably try to find something close to themselves and their dream prince in life.” The World: No Redheads or Blondes Wanted at Sperm Bank, supra.
may translate into some form of self-reproduction, since other sources indicate that
most sperm-bank users are white.\textsuperscript{238} With respect to adoption agencies, scholars
often speak of a “same-race preference,”\textsuperscript{239} which similarly suggests that most
prospective parents wish to self-reproduce rather than to reify racial hierarchies.
There is, accordingly, insufficient evidence to substantiate concerns that allowing
nontherapeutic PGD would exacerbate racial marginalization and diminish genetic
diversity. Yet, since PGD users’ preferences regarding eye, hair, and skin color
have not been fully studied, future studies could alter this analysis.

There is also little evidence to suggest that a significant proportion of
prospective parents would use nontherapeutic PGD to select against females—and
sex preferences, unlike eye, hair, and skin color preferences, have been extensively
studied. One study found that, while PGD patients of Chinese, Indian and
Arab/Muslim origin were indeed inclined to select against females, those of
Western origin exhibited a “slight but not significant preference for females.”\textsuperscript{240}
The authors therefore concluded that “sex selection by PGD in an ethnically
diverse Western society, like the USA, does not have any significant effect on
population sex ratio [and] does not discriminate against female embryos . . . .”\textsuperscript{241}
Other studies have found that many prospective parents who select for sex are
acting to “balance” their families rather than to produce all male or all female
children.\textsuperscript{242} If this is accurate, then allowing sex selection might actually help to
maintain the current sex ratio.

One might nevertheless argue that, even though current users of PGD do not
exhibit a preference for males, a preference could emerge if PGD became more
widely available and affordable. A 2011 Gallup Poll lends some credence to this
concern, but ultimately fails to substantiate it. When asked whether, if they could
have only one child, they would prefer a boy or girl, 40% of Americans indicated

\begin{itemize}
\item “ideal donor” is “6 feet tall, [with a] college degree, brown eyes, blond hair, and
dimples.”
\item Courtney Megan Cahill, \textit{Regulating at the Margins: Non-Traditional Kinship and
the Legal Regulation of Intimate and Family Life}, 54 ARIZ. L. REV. 43, 63 (2012) (citing
\item Barbara Fedders, \textit{Race and Market Values in Domestic Infant Adoption}, 88 N.C. L.
REV. 1687, 1695–96, 1706 (2010) (explaining that “[m]any agencies classify babies
available for adoption according to race” and describing the “same-race preference[]”
by stating that “more white adults enter the formal private adoption process as prospective
adoptive parents than do adults of any other race, and most of these individuals express a
preference for adopting white babies”).
\item Pere Colls et al., \textit{Preimplantation Genetic Diagnosis for Gender Selection in the
USA}, 4 ETHICS, BIOSCIENCE & LIFE 16, 20 (2009) (emphasis added) (studying parents who
had used PGD in 2007 or 2008).
\item \textit{Id.} at 18, 21.
\item \textit{Id.} at 17 (“[S]everal studies have demonstrated that . . . gender preferences are
usually the result of a desire to have a family with children of both genders (family
balancing) . . . .”).
\end{itemize}
that they would prefer a boy, and 28% indicated that they would prefer a girl.243

There are, however, at least two reasons why this does not prove that allowing sex selection would marginalize women. First, the poll asked which sex respondents would prefer if they could have only one child, whereas most Americans who have any children have more than one244 and reportedly perceive the “ideal” number of children to be 2.5.245 Coupled with the preference for “balanced” families, this suggests that Americans would not consistently favor males. Second, the poll did not ask whether respondents would actually use PGD to achieve their preferences, and other studies report that “only a marginal percentage of the population has expressed willingness to deploy technology in order to select the sex of the embryo.”246 Concerns that allowing sex selection would marginalize women and diminish genetic diversity, thus, similarly remain unsubstantiated.

There is, in sum, little evidence that a ban on nontherapeutic PGD would diminish genetic diversity with respect to racial markers or sex. It is, however, possible that a ban on nontherapeutic PGD would diminish genetic diversity with respect to traits conventionally understood as disabilities, because prospective parents often select against—and only rarely select for—such traits.247 While there are deep ethical concerns about allowing prospective parents to select for disabilities, one could argue that allowing such selection has the potential to preserve genetic diversity and prevent the further marginalization of individuals with disabilities. Professor Dorothy Roberts has written that the “availability of prenatal diagnosis for a disorder may discourage government funding for research and social services for people who have the disorder.”248 Professor Judith Daar has

243 A 2011 Gallup Poll asked “Suppose you could only have one child. Would you prefer that it be a boy or a girl?” Forty percent of respondents said they would prefer a boy, 28% said they would prefer a girl, 26% said it did not matter, and the remainder were unsure or had no opinion. The slight preference for boys is driven entirely by men: 49% of men said they would prefer a boy and 22% of men said they would prefer a girl; in contrast, 31% of women said they would prefer a boy and 33% said they would prefer a girl. Frank Newport, Americans Prefer Boys to Girls, Just as They Did in 1941, GALLUP POLL (June 23, 2011), http://www.gallup.com/poll/148187/Americans-Prefer-Boys-Girls-1941.aspx.

244 Table ST-F1-2000: Average Number of Children Per Family and Per Family With Children, by State: 2000 Census, U.S. CENSUS BUREAU (Sept. 15, 2004), http://www.census.gov/polling/soedemo/hh-fam/tabST-F1-2000.pdf (reporting that, among families who have any children under 18, the national average number of children is 1.86).


246 See Zafran, supra note 6, at 195.

247 See generally Baruch et al., supra note 57 (discussing the number of clinics that provide various PGD services and reasons why those clinics do or do not provide those services).

248 Roberts, supra note 182, at 1355.
expressed similar concerns and has written that those “afflicted with a genetic disorder that was susceptible to detection by PGD will likely be part of a lower socioeconomic class.” Such individuals, she explains, having been born to parents who “lacked the resources to access PGD,” will themselves lack the resources to obtain “adequate health care to manage the disease.” If nontherapeutic PGD were banned, therapeutic PGD might be used to dramatically diminish or entirely eliminate the presence of certain disabilities from the population. While the elimination of some disabilities might be desirable, the elimination of others might be detrimental.

While reasonable people could disagree on whether banning nontherapeutic PGD would diminish genetic diversity, there is a strong argument that it would diminish the expression of ideological diversity. To prevent prospective parents from using nontherapeutic PGD would be to deprive them of the ability to act on their own diverse ethical and ideological viewpoints. Such a deprivation seems unwarranted in light of the fact that those who use nontherapeutic PGD appear to be aware of the ethical complexity of their decisions. Furthermore, the high level of commitment to trait selection that some prospective parents have exhibited suggests that, if nontherapeutic PGD were unavailable, they might resort to prenatal testing in combination with nontherapeutic abortion. Allowing prospective parents access to nontherapeutic PGD could, therefore, have the dual benefits of promoting ideological diversity and decreasing the number of nontherapeutic abortions. While ideological diversity should not be promoted at all costs—for example, if its promotion would lead to a problematic decrease in another form of diversity—it should be promoted in the absence of costs.

One might counter this argument about ideological diversity by pointing out that, while allowing nontherapeutic PGD might promote ideological diversity among prospective parents, it might diminish ideological diversity among their prospective children, who will be forced to live out (or, in some cases, defy) the stereotypes associated with their selected traits. But it seems unlikely that banning nontherapeutic PGD will substantially impact the lives of children whose parents would use the technology if it were available. It is, indeed, odd to think that prospective parents, if allowed to engage in trait selection, would force their children to live out the stereotypes associated with their traits, while the same parents, if barred from selecting their children’s traits, would allow their children to self-determine. While one may think that allowing parents to standardize their children by forcing them to live out or defy stereotypes is problematic, banning nontherapeutic PGD will not independently solve this problem.

250 Id. at 259–60.
251 Id. at 260.
252 See Sharp et al., supra note 75, at 844–45.
253 Consider, for example, the achondroplasic parents discussed supra Part II.A.2.b.
254 Resolving this problem would, indeed, require greater respect for children’s rights.
It therefore appears that the first policy option—banning nontherapeutic PGD—would be undesirable. Yet further inquiry is necessary to determine whether the second or third policy options—allowing nontherapeutic PGD or subsidizing it—would in fact be preferable. While allowing nontherapeutic PGD would promote ideological diversity, it would not do so on an equal basis. Due to the high cost of IVF and PGD, only the wealthy can realistically engage in trait selection; the poor (as with abortion) are left unable to operationalize their ideologies and express their diversity. Subsidizing nontherapeutic PGD, though inconsistent with current policies and (under existing doctrine) not constitutionally required, would address this concern. This is not to say that subsidies are unequivocally the best policy, simply that they appear most likely to enable the expression of genetic and ideological diversity. If future studies revealed that this were not accurate—because prospective parents were, for example, consistently selecting against given eye, hair, and skin colors, or against a given sex—reassessment would of course be required.

CONCLUSION

This Article draws on postmodern theory to develop a framework for analyzing debates in which liberty and equality appear to conflict, such as the debate over whether to ban nontherapeutic PGD. If concerns about liberty cause legislatures to allow nontherapeutic PGD to remain virtually unregulated, wealthy white couples might select nondisabled, light-skinned, blond-haired, blue-eyed males, and equality might be diminished. Conversely, if concerns about equality cause legislatures to ban nontherapeutic PGD, prospective parents will be prevented from making decisions about procreation and childrearing, and liberty might be diminished. While these dominant views of liberty and equality are in conflict, other currently marginalized views are fully compatible. The fact that the Fourteenth Amendment protects both rights suggests that that we ought to choose the latter (compatible) views. We should, when selecting among compatible views, choose those most likely to promote diversity. While it may in some cases be difficult to determine which views are most likely to promote diversity, careful consideration is worthwhile because, in promoting diversity, we promote democracy.

255 There would, of course, be no constitutional concerns unless state actors were involved. See supra Part II.B.2.
DISSENT INTO CONFUSION: THE SUPREME COURT, DENIALISM, AND THE FALSE “SCIENTIFIC” CONTROVERSY OVER SHAKEN BABY SYNDROME

Joëlle Anne Moreno* & Brian Holmgren**

I. INTRODUCTION

Some scientific controversies are real; some are false. Challenges to the existence of global warming, and arguments about childhood vaccines causing autism or intelligent design versus evolution are false controversies because there is near consensus in the global scientific community on these questions.¹ Near consensus in science means that, as with all legitimate scientific research, there are unsettled questions that merit future investigation and reasonable experts may differ over select issues, but these unresolved matters do not threaten core scientific foundations. In contrast, false scientific controversies have been fabricated and are a form of denialism²—the rejection of scientifically sound

¹ See generally Leah Ceccarelli, Manufactured Scientific Controversy: Science, Rhetoric, and Public Debate, 14 RHETORIC & PUB. AFFAIRS 195 (2011) (discussing the false “manufactured” controversies of global warming skepticism, dissent over AIDS being caused by HIV, and intelligent design); see also Understanding Evolution, UNIV. CAL. MUSEUM PALEONTOLOGY, http://evolution.berkeley.edu (last visited Mar. 10, 2013) (discussing the false “scientific controversy” of intelligent design versus evolution); Vaccine Safety, CENTER FOR DISEASE CONTROL & PREVENTION, http://www.cdc.gov/vaccinesafety (last visited Mar. 10, 2013) (discussing the false “scientific controversy” regarding whether childhood vaccines can cause autism). As Professor Susan Haack has observed:

At any time there is a whole continuum of scientific ideas, claims, and theories: some so well-warranted by such strong evidence that it is most unlikely they will have to be revised; some not quite so well-warranted but still pretty solidly established; some promising but as yet far from certain; some new and exciting but highly speculative and as yet untested; and some so wild that few mainstream scientists are willing even to listen to them. (The proportion of the well-warranted to the highly-speculative varies, obviously, across fields and sub-fields.) A few of the exciting but as yet untested ideas, and a very, very few of the wildest ideas, will eventually turn out to be warrantable; but most will not.


² See generally Michael Specter, Denialism: How Irrational Thinking Hinders Scientific Progress, Harms the Planet, and Threatens Our Lives (2009) (describing the phenomenon of denialism and discussing its dangers); Martin McKee &
information in favor of purported “truth” claims that cannot be empirically supported. In her excellent work on manufactured controversies, Professor Leah Ceccarelli explains that a false scientific controversy is typically marked by an announcement that “there is an ongoing scientific debate in the technical sphere about a matter for which there is actually overwhelming scientific consensus.”

Professor Ceccarelli also identifies several common strategies used to manufacture false controversies, including (1) the use of mercenary scientists, (2) the use of cherry-picked data and manipulation of statistical methods, (3) the manufacture and promotion of doubt and uncertainty, and (4) the use of rhetoric to manufacture controversy in addition to uncertainty.

In the United States Supreme Court’s first opinion on the merits from its 2011–2012 term, three members of the Court contributed their authoritative voices to one of the most recent—and one of the most deadly—false scientific controversies, the purported scientific debate over the medical diagnosis of shaken baby syndrome (SBS), a prevalent form of abusive head trauma (AHT).

On October 31, 2011, in Cavazos v. Smith, the Supreme Court upheld Shirley Ree Smith’s conviction for causing the death of her seven-week-old grandson, Etzel. This conviction was based on the jury finding that Etzel died from SBS, a diagnosis that has been recognized as clinically valid and evidence-based by an overwhelming majority of pediatric medical specialists for almost half a century.

Pascal Diethelm, *How the Growth of Denialism Undermines Public Health*, 341 BMJ 1309, 1310 (2010) (noting that “denialism” in the medical arena is characterized by several features, including (a) “identification of conspiracies,” (b) “use of fake experts,” (c) “selectivity of citation,” (d) “[c]reation of impossible expectations of research,” (e) “[m]isrepresentation and logical fallacies,” and (f) “[m]anufacture of doubt”). Denialism and, more specifically, explorations of manufactured “scientific controversies” such as Professor Ceceralli’s work, *supra* note 1, at 195, provide generally useful tools for understanding much of the specific scientific, pseudoscientific, and legal academic information discussed below.

See McKee & Diethelm, *supra* note 2, at 1309.

Ceccarelli, *supra* note 1, at 196 (emphasis added).

Id. at 197.

132 S. Ct. 2 (2011) (per curiam).

Id. at 8. The relevant statute provides, “Any person who, having the care or custody of a child who is under eight years of age, assaults the child by means of force that to a reasonable person would be likely to produce great bodily injury, resulting in the child’s death, shall be punished by imprisonment . . . .” *CAL. PENAL CODE § 273ab* (West 2008).

substantiated by the bulk of the medical research in a range of scientific disciplines,9 recognized and defined by the Centers for Disease Control and Prevention,10 and widely accepted by courts in the United States11 and numerous foreign countries. Following a brief review of the evidence presented at trial, the Supreme Court issued a scientifically accurate per curiam decision. The Court held that the Ninth Circuit had improperly “substituted its judgment for that of a California jury on the question of whether the prosecution’s or defense’s expert witnesses more persuasively explained the cause of a death.”12

Justice Ginsburg, joined by Justices Breyer and Sotomayor, seized this opportunity to issue an unusual and scientifically inaccurate dissenting opinion. Based on their review of the medical and nonmedical evidence presented in this case, the dissenters opined that “[f]ew of the signs of SBS were present.”13 More generally, the Court’s summary adjudication of the Smith case was “untoward” because the dissenters believed that “[d]oubt has increased in the medical

_________


11 See cases cited infra notes 47, 71; see also JOHN E.B. MYERS, MYERS ON EVIDENCE OF INTERPERSONAL VIOLENCE: CHILD MALTREATMENT, INTIMATE PARTNER VIOLENCE, RAPE, STALKING, AND ELDER ABUSE (5th ed. 2011) (discussing the issues surrounding expert medical testimony in this arena and citing numerous cases as examples).


13 Smith, 132 S. Ct. at 9 (Ginsburg, J., dissenting)
community ‘over whether infants can be fatally injured through shaking alone.’”¹⁴ Because these three justices purport to describe the shifting opinion of the “medical community,” one might initially mistake this finding for a distillation of the relevant medical literature on abusive head trauma/shaken baby syndrome (AHT/SBS).¹⁵ Nothing could be further from the truth. Instead the dissenters engage in two different, significant, and interrelated jurisprudential errors.

The first jurisprudential error, which will be fully addressed in this Article, is the dissenters’ conclusion that few of the signs of SBS were present in this case which misconstrues the medical and nonmedical evidence presented by the prosecution and defense. The second is their sweeping conclusion that doubt has increased within the medical community regarding SBS which is based on the dissenters’ careless and irresponsible independent extrarecord judicial fact-finding and contradicted by over seven hundred AHT/SBS medical articles written over the past four decades.¹⁶ Ignoring the overwhelming medical evidence, the justices support their opinion with single-sentence quotations from seven cherry-picked sources. This type of faux fact-finding, which perpetuates the false SBS

---

¹⁴ Id., at 9–10 (Ginsburg, J., dissenting) (quoting State v. Edmunds, 746 N.W.2d 590, 596 (Wis. Ct. App. 2008)).

¹⁵ This Article refers to AHT/SBS because the American Academy of Pediatrics has recently revised its own position paper on SBS to be more inclusive of the multiple mechanisms by which AHT may be inflicted. See Cindy W. Christian & Robert W. Block, Abusive Head Trauma in Infants and Children, 123 PEDIATRICS 1409, 1409–11 (2009) (setting forth the American Academy of Pediatrics (AAP) position and noting that the Academy determined it was necessary to modify the terminology for describing inflicted head trauma to recognize the multiple mechanisms by which the spectrum of injuries could be inflicted, including shaking, impact, a combination, and additional mechanisms). Contrary to the representations made by many defense-retained witnesses and legal academic commentators, the 2009 position statement does not do away with shaking as a mechanism of injury but reaffirms it. According to the AAP, “Shaken baby syndrome is a subset of AHT. Injuries induced by shaking and those caused by blunt trauma have the potential to result in death or permanent neurologic disability” and “[t]he goal of this policy statement is not to detract from shaking as a mechanism of AHT but to broaden the terminology to account for the multitude of primary and secondary injuries that result from AHT . . . .” Id. at 1409–10; see also Stephen Lazoritz et al., The Whiplash Shaken Infant Syndrome: Has Caffey’s Syndrome Changed or Have We Changed His Syndrome?, 21 CHILD ABUSE & NEGLECT 1009, 1010–13 (1997) (suggesting that in 1997 Dr. Caffey’s terminology and diagnostic criteria for “Whiplash Shaken Infant Syndrome” should be changed to encompass broader understandings of the mechanisms of injury based on medical research on the syndrome over the past twenty-five years).

¹⁶ This conclusion might logically follow because no amicus briefs were filed in the Supreme Court and only one of the articles relied upon by the dissenters was cited in the defendant’s brief, suggesting that the dissenters conducted an independent analysis of the extant medical literature. See Respondent’s Brief in Opposition at 35, Cavazos v. Smith, 132 S. Ct. 2 (2011) (No. 10-1115) (citing Faris A. Bandak, Shaken Baby Syndrome: A Biomechanics Analysis of Injury Mechanisms, 151 FORENSIC SCI. INT’L 71, 71–79 (2005)).
contrroversy, will be addressed in detail in a subsequent companion article.\(^{17}\) It is enough to note here that the dissenters base their conclusion on a handful of papers: (1) written by a tiny group of “mercenary scientists” whose regular testimony as defense-retained witnesses in child abuse and child homicide cases undermines the objectivity, legitimacy, and validity of their work;\(^{18}\) (2) that contain little original research and instead reflect manipulation of data and statistical methods,\(^{19}\) (i.e., opinion pieces, nonrandomized retrospective case series/reports, scientifically unsubstantiated opinions of other “mercenary” witnesses, and mischaracterizations of earlier AHT/SBS research);\(^{20}\) (3) written not for academic and research purposes, but for use in legal proceedings; and (4) riddled with blatant methodological flaws and discredited by pediatric expert medical research and peer-reviewed scientific publications in a wide range of fields.\(^{21}\) Although the Smith dissenters could not garner a majority for their empirically unwarranted assertions regarding the purportedly unsettled state of the science of AHT/SBS, the scope and power of their view became undeniable on April 6, 2012. On that day, California Governor Jerry Brown commuted Smith’s sentence, echoing Justice Ginsburg’s conclusion that the AHT/SBS controversy made it “clear that significant doubts surround Ms. Smith’s conviction.”\(^{22}\)

\(^{17}\) Joëlle Anne Moreno & Brian Holmgren, The Supreme Court Screws Up the Science: There Is No Abusive Head Trauma/Shaken Baby Syndrome Controversy, 2013 UTAH L. REV. (forthcoming).

\(^{18}\) According to Dr. Daniel Lindberg of Brigham and Women’s Hospital, the AHT/SBS controversy has been manufactured based “exclusively on the opinions and work of ‘experts’ who derive substantial income from lucrative court testimony on behalf of the accused perpetrators of child abuse” and “rarely, if ever, provide medical care for children.” Carey Goldberg, The Real Consensus on Shaken Baby Syndrome?, WBUR’S COMMONHEALTH REFORM & REALITY (Sept. 27, 2010, 5:12 PM), http://commonhealth.wbur.org/2010/09/shaken-baby/; see also Narang, supra note 8, at 593–94 (“[T]he pecuniary interest in providing expert testimony cannot be underestimated. It has posed and continues to pose a significant risk to the presentation of unbiased medical information. . . . [I]n addition to pecuniary interest, as discussed above, personal prejudices can also affect scientific analysis. This can result in the adherence to disproven theories and the presentation of skewed information.”); Kenneth Feldman, Commentary on Congenital Rickets Article, 39 PEDIATRIC RADIOLOGY 1127, 1128–29 (2009) (providing example of pecuniary interests of repeat expert witnesses).

\(^{19}\) Ceccarelli, supra note 1, at 197.

\(^{20}\) Narang, supra note 8, at 541 (noting that all of the medical papers “‘questioning’ the validity of AHT (save two or three) are non-randomized, retrospective case series/reports, and without comparative control groups. In fact, many are single case reports.”).

\(^{21}\) See Moreno & Holmgren, supra note 17; Sandeep Narang et al., A Daubert Analysis of Abusive Head Trauma/Shaken Baby Syndrome, Part II: An Examination of the Differential Diagnosis, 13 HOUS. J. HEALTH L. & POL’Y (forthcoming 2013).

A. The Supreme Court Enters an Ongoing “Controversy”

The dissenters’ conclusions and choice of sources were not novel. Their findings echo the work of a small, partisan, and vocal group of law professors and law students who have recently challenged the diagnostic validity of AHT/SBS and advanced their own specious claims of a “scientific controversy.” Justice Ginsburg notably did not cite any of these law review articles. However, because her opinion closely mirrors these works, she grants an unwarranted imprimatur of legitimacy to legal academic arguments that SBS “quite possibly does not exist,” may be “junk science,” that “SBS science in its current dissenting opinion may have had on Governor Brown’s decision. The remaining sections of this Article should clarify whether the medical and nonmedical evidence relied on (or ignored) by the Ninth Circuit and the Smith Court’s three dissenting justices actually raises “significant doubts” about the guilt of the defendant.


24 Rachel Burg, Note, Un-Convicting the Innocent: The Case for Shaken Baby Syndrome Review Panels, 45 U. MICH. J.L. REFORM 657 (2012); Molly Gena, Comment, Shaken Baby Syndrome: Medical Uncertainty Casts Doubts on Convictions, 2007 Wis. L. REV. 701; Genie Lyons, Comment & Note, Shaken Baby Syndrome: A Questionable Scientific Syndrome and a Dangerous Legal Concept, 2003 UTAH L. REV. 1109, 1132 (asserting that “[f]or many years now, attorneys have been willing to prosecute, and juries have been willing to convict, people whose only clearly established mistake was caring for a baby that died”); Daniel G. Orenstein, Comment, Shaken to the Core: Emerging Scientific Opinion and Post-Conviction Relief in Cases of Shaken Baby Syndrome, 42 ARIZ. ST. L.J. 1305 (2011); Lauren Quint, Note, Bridging the Gap: An Application of Social Frameworks Evidence to Shaken Baby Syndrome, 62 HASTINGS L.J. 1839 (2011). A few defense practitioners have authored similar articles, which suffer from the same shortcomings of selective, improper, or incomplete citation. See, e.g., Matthew D. Ramsey, A Nuts and Bolts Approach to Litigating the Shaken Baby or Shaken Impact Syndrome, 188 MIL. L. REV. 1 (2006); Elizabeth A. Walker, Shaken Baby Syndrome: Daubert and MRE 702’s Failure to Exclude Unreliable Scientific Evidence and the Need for Reform, 210 MIL. L. REV. 1 (2011).

25 See infra Part II.C.

26 Lyons, supra note 24, at 1109.

27 Imwinkelried, supra note 23, at 158 (“The question here is whether shaken baby syndrome evidence is ‘junk’ science presenting an intolerable risk of a wrongful conviction . . . .”).
conflicted state . . . does not support criminal convictions,”28 and that the medical community has “deliberately discarded a diagnosis defined by shaking.”29 These attention-grabbing claims fundamentally misconstrue and misstate the basic science involved in the medical diagnosis of child abuse and the type of medical expert testimony offered in legal proceedings. Like the Smith dissenters, law professors and students who claim to have “ripped the lid off” the scandal of false AHT/SBS convictions base their assertions on selective or improper citation to outlier medical papers that: (1) rely on unscientific methods; (2) are written almost exclusively by self-interested and highly-paid defense witnesses; and (3) ignore the vast quantity of valid, easily accessible, evidence-based medical research and the many public and professional statements that substantiate AHT/SBS as a clinically valid diagnosis.30

Supreme Court justices, law professors, and law students are generally not scientists. As Professor Susan Haack astutely observed, when courts rely on information beyond their ken they must assume that “[g]iven the investigative character of the scientific enterprise and the pervasive reliance of individual scientists on evidence discovered by others, the core values are honesty (both with yourself and with other people) about what the evidence is and where it leads.”31 Thus, especially when wading into a purported “controversy,” courts should hesitate before relying on any so-called child abuse “expert” whose public statements include “[a]s far as I’m concerned, every goddam convictions in this country over the past 25 years which is based on testimony regarding shaking has to be overturned,”32 inviting skepticism regarding the expert’s accuracy, honesty, and objectivity.

The problem of biased and scientifically unsound defense witness testimony in the AHT/SBS context has increasingly aroused the attention of various medical associations. For example, a recent article in the Journal of the American Medical Association described how legal cases involving AHT/SBS have been harmed by

28 Orenstein, supra note 24, at 1306.
29 Tuerkheimer, supra note 23, at 11.
30 See Narang, supra note 8, at 574–76 (listing organizations endorsing the validity of AHT/SBS promulgated by the World Health Organization, the American Academy of Pediatrics, the American Academy of Ophthalmology, the American Association for Pediatric Ophthalmology and Strabismus, the American College of Radiology, the American Academy of Family Physicians, the American College of Surgeons, the American Association of Neurologic Surgeons, the Pediatric Orthopedic Society of North America, the American College of Emergency Physicians, the American Academy of Neurology, the Royal College of Pediatrics and Child Health, the Royal College of Radiologists, the Royal College of Ophthalmologists, and the Canadian Pediatric Society).
31 Haack, supra note 1, at 998.
“physicians with variable credentials [who] have a willingness to disparage scientifically grounded and accepted testimony, use unique theories of causation, omit pertinent facts or knowledge, use unique or unusual interpretations of medical findings, make false statements, or engage in flagrant misquoting of medical journals.”

In a similar effort, physicians involved in the diagnosis and treatment of AHT/SBS have recently called for new certification programs and specific rules of ethical expert conduct that would “provide some degree of certainty that physicians testifying both for the prosecution and defense as AHT medical experts are indeed expert, experienced, and unbiased.”

In the interim, judges can use Federal Rule of Evidence 702 (or its state equivalent) to exclude or limit the testimony of partisan experts who “have developed their opinions expressly for the purposes of testifying.”

B. The AHT/SBS “Controversy” is False

Despite all the ballyhoo, there has been no paradigm shift in the scientific support for the diagnosis of AHT/SBS. The empirical evidence includes a continuously growing body of “evidence-based, peer-reviewed medical literature with 40 years of contributions by pediatricians, neuroradiologists, clinical and forensic pathologists, ophthalmologists, and physiologists clearly supporting the construct of a medical diagnosis of AHT.” At the clinical level, “[m]edical experts agree with the physical, laboratory, and imaging findings associated with the medical construct of AHT, which can include subdural hemorrhage, retinal hemorrhage, encephalopathy, and often evidence of previous trauma or other bodily injury.”


FED. R. EVID. 702 advisory committee’s note (quoting Daubert v. Merrill Dow Pharm., Inc., 43 F.3d 1311, 1317 (9th Cir 1995)).

Neuropathologists, radiologists, hematologists, and biomechanicians have also contributed to the literature in support of AHT/SBS. See sources cited supra notes 8–9.

Albert et al., supra note 33, at 39; see also Sabine A. Maguire et al., Which Clinical Features Distinguish Inflicted from Non-Inflicted Brain Injury? A Systematic Review, 94 ARCHIVES DISEASE IN CHILDHOOD 860, 865–66 (2009) (noting the positive predictive value of retinal hemorrhages and apnea in AHT, but explaining that, as in most cases of physical child abuse, there is no diagnostic test for inflicted brain injury and the diagnosis is made on the basis of probability after careful exclusion of other possible causes for the clinical findings, including accidental injury and other medical conditions). Contrary to the representations made by most defense witnesses and legal scholars challenging the diagnosis of AHT/SBS, the diagnosis of this form of trauma is not made exclusively on the basis of the triad of injuries, but instead as part of an extensive differential diagnostic process that considers all aspects of the medical evaluation along with all of the investigative information from other relevant nonmedical sources. See infra notes 61–64 and accompanying text.
Of course, as with any area of science, the existence of extensive substantiating medical evidence does not mean that every question has been answered, that there are no areas of legitimate uncertainty, or that research should not continue. But false claims of a scientific paradigm shift completely mischaracterize the existing medical evidence. Law professors and students who seek to analogize these claims to the false convictions uncovered using DNA evidence embrace a false science and paradoxically reject the core lesson of the Innocence Project—that good science makes good law.

C. How False “Controversies” Create Real Problems

Unlike most of the academic arguments that consume legal scholars, it is a matter of life or death when judges who must decide child abuse cases mistake biased, poorly substantiated, and outlier advocacy for legitimate medical information. Proponents of the false AHT/SBS controversy transcend academic discourse to undermine real world public health and child abuse prevention efforts by suggesting that shaking an infant is not dangerous and cannot cause serious injury. The flaws in Justice Ginsburg’s dissenting opinion may not be obvious, so

38 See Tuerkheimer, The Next Innocence Project, supra note 23, at 2 (claiming that her article “identifies a criminal justice crisis”); Burg, supra note 24, at 660 (suggesting that the medical community has shifted towards skepticism of SBS, but that innocent people continue to be falsely convicted on the basis of this scientifically questionable diagnosis); Lyons, supra note 24, at 1132 (“For many years now, attorneys have been willing to prosecute, and juries have been willing to convict, people whose only clearly established mistake was caring for a baby that died.”); Orenstein, supra note 24, at 1305–07 (alleging that “the American criminal justice system . . . [must] address concerns that SBS theory has potentially sent wrongfully convicted persons to prison”).

39 The vast majority of the exonerations secured by the Innocence Project have been based on the legitimate science of DNA sampling and testing. Thus, the comparison between the valid and widely accepted science of DNA analysis and outlier litigation-driven challenges to the validity of AHT/SBS is especially inapt. Peter J. Neufeld, an Innocence Project cofounder, has specifically bemoaned courts’ reliance on specious litigation-driven science and advocated for more stringent screening including more careful application of the criteria outlined in Daubert v. Merrell Dow Pharmaceuticals Inc., 509 U.S. 579 (1993). See Peter J. Neufeld, The (Near) Irrelevance of Daubert to Criminal Justice and Some Suggestions for Reform, 95 AM. J. PUB. HEALTH S107 (2005).

40 As Chief Justice John Roberts lamented in his June 2011 address to the Fourth Circuit Judicial Conference, “Pick up a copy of any law review that you see and the first article is likely to be, you know, the influence of Immanuel Kant on evidentiary approaches in 18th century Bulgaria.” Richard Brust, The High Bench vs. The Ivory Tower, A.B.A. J. (Feb. 17, 2012, 7:12 CDT), http://www.abajournal.com/mobile/article/the_high_bench_vs._the_ivory.

41 See infra notes 47, 69–71, 74, 81–91 and accompanying text (noting judicial decisions on SBS).

42 See Christian & Block, supra note 15, at 1409–11 (setting forth the American Academy of Pediatrics position paper on AHT and discussing the public health and prevention efforts encouraged by the Academy to educate parents about the dangers of
they must be corrected before these mistakes gain traction with future courts, the media, and the public.

This correction must be prompt because incidences of AHT in children are increasing. A new multicenter study reveals that the overall rate of AHT, which is the leading cause of death from child abuse, has increased over 60%, from 8.9 to 14.7 per 100,000 children under five years old, over the past five years.43 The current incidence of AHT in children under one year is 20 to 30 cases per 100,000, with a fatality rate of 20% and a significant disability rate of 66%.44 As noted above, the promulgation of scientifically unsubstantiated claims that shaking cannot harm an infant sends a dangerous message to parents and caregivers and undermines important child abuse prevention efforts.45

This correction must be accurate and empirically sound. As shown below, the Smith dissent reveals how courts carelessly or inadvertently rely on pseudoscientific information resoundingly rejected within multiple scientific fields.46 This is more likely to occur when the court engages in its own fact finding and especially when it fails to use valid and transparent source selection criteria. These problems are compounded when brief quotations are excerpted without attention to the remainder of the source (including the description of methods), concurrent or subsequently published critical responses to the selected work, or other articles on the same topic that reach different conclusions. It is vitally important that future courts do not uncritically rely on the Smith dissent, or the sources cited therein, to make similar unscientific mistakes.

This correction must be clear and accessible to nonscientists. In a growing number of child homicide and abuse cases involving a medical diagnosis of AHT/SBS (based on clear clinical findings and extensive medical research), scientific-sounding information of dubious validity has increasingly been paraded before trial courts and offered to support postconviction claims of “factual innocence” or “newly discovered evidence.”47 Ironically, this may be attributed, in...
part, to the Daubert standard, which encourages judges to consider publication as a measure of scientific validity. The problem is that judges mistake publication in any medical journal for a determination that the author has used scientifically sound methods to reach valid conclusions. What courts routinely fail to understand is that not all medical journals are the same, running the gamut from prestigious peer-reviewed journals to undiscriminating pay-to-publish outlets. More importantly, even well regarded journals will sometimes publish an article that presents an outlier view specifically to expose the article to critique from others in the field.48

For a recent concrete example of an outlier theory published along with critical responses, see Kathy A. Keller & Patrick D. Barnes, Rickets vs. Abuse: A National and International Epidemic, 38 PEDIATRIC RADIOLOGY 1210 (2008) (proposing that “congenital rickets” could account for multiple fractures in several alleged child abuse cases). That article was published not as an accepted peer-reviewed article but instead as a “comment” along with invited critiques from numerous other doctors and the editors of the journal in which it was published. Id.; see, e.g., Thomas L. Slovis & Stephen Chapman, Vitamin D Insufficiency/Deficiency—A Conundrum, 38 PEDIATRIC RADIOLOGY 1153 (2008); Thomas L. Slovis & Stephen Chapman, Evaluating the Data Concerning Vitamin D Insufficiency/Deficiency and Child Abuse, 38 PEDIATRIC RADIOLOGY 1221 (2008) (providing the editor’s comments about the lack of scientific support for the conclusions made by Drs. Keller and Barnes); Carole Jenny, Rickets or Abuse?, 38 PEDIATRIC RADIOLOGY 1219 (2008) (criticizing the methodology used by Drs. Barnes and Keller and their selection bias based on their extensive experience as expert witnesses); Feldman, supra note 18, at 1127 (noting that several cases presented by Drs. Barnes and Keller contained significant omissions, including findings not seen by several other radiologists who reviewed the films and the authors’ failure to disclose their role as defense experts who routinely testify in cases where this defense is advanced).

Although the focus of this Article is the diagnosis of AHT/SBS, similar problems arise in child abuse and child homicide cases involving different types of injuries. See, e.g., Joëlle Anne Moreno, Einstein on the Bench?: Exposing What Judges Do Not Know About Science and Using Child Abuse Cases to Improve How Courts Evaluate Scientific Evidence, 64 OHIO ST. L.J. 531, 535–36 (2003) (exploring the unscientific diagnosis of “temporary brittle bone disease” offered by defense witnesses to explain fracture injuries in children). In a recent child abuse trial in San Diego involving multiple fractures to a baby, Dr. Patrick Barnes testified for the defendant and attributed the child’s injuries to congenital rickets. See Reporter’s Transcript of Proceedings at 12–13, People v. Sanders,
Finally, this correction must embrace both the science and law of AHT/SBS. In many AHT cases, including the Smith case, the defendant makes admissions to medical professionals, social services personnel, or to the police. There is a substantial and growing body of research demonstrating that incriminating statements and confessions (made to family members, doctors, social service personnel, or police officers) often describe a mechanism of injury that is consistent with the clinical findings, thereby confirming the medical evidence. But some law professors have recently argued that defendant confessions and admissions should be ignored because such statements are the product of a pervasive child abuse prosecution bias. Despite the lack of any evidence to support this “bias,” one medical author (cited approvingly by the Smith dissenter), simply states that incriminating statements and confessions in AHT/SBS cases are inherently unreliable and should be irrelevant to any determination of causation. While arguments regarding cognitive biases are a recent academic fad, it is ridiculous to posit that every medical professional, social worker, law enforcement personnel, or anyone else to whom a child abuse/homicide suspect makes incriminating statements is so tainted by a pro-prosecution bias that they are blind to all other evidence and cannot exercise independent judgment.

As discussed below, before Smith, most courts accurately concluded that the diagnosis of AHT/SBS is supported by extensive valid relevant medical evidence and is generally accepted in the relevant medical community. The Smith dissent, which provides a model for examining the above-described problems in detail and

---

49 See, e.g., Catherine Adamsbaum et al., Abusive Head Trauma: Judicial Admissions Highlight Violent and Repetitive Shaking, 126 PEDIATRICS 546, 550, 554 (2010); Suzanne P. Starling et al., Abusive Head Trauma: The Relationship of Perpetrators to Their Victims, 95 PEDIATRICS 259, 261 (1995) [hereinafter Starling et al., Abusive Head Trauma]; Suzanne P. Starling et al., Analysis of Perpetrator Admissions to Inflicted Traumatic Brain Injury in Children, 158 ARCHIVES PEDIATRICS & ADOLESCENT MED. 454, 456–57 (2004) [hereinafter Starling et al., Analysis of Perpetrator Admissions].

50 See, e.g., Keith Findley, Clinical Professor, Presentation at Twelfth International Conference on Shaken Baby Syndrome / Abusive Head Trauma: What Role Should Confessions Play in Diagnosing Abusive Head Trauma? (Oct. 1, 2012); see also Symposium, supra note 23, at 232 (statement of Professor Keith Findley) (proposing that confessions and adjudications are not reliable for supporting the “hypothesis” of SBS).


52 See infra note 71 and accompanying text.
context, if uncorrected, will redound to the lower courts and distort public awareness and opinion regarding AHT/SBS.  

D. An Evidence-Based Approach to AHT/SBS

It is against this backdrop of increasing incidence of child abuse and child homicide, decreasing understanding of the clinical diagnostic criteria and medical literature, and a deliberate effort to politicize AHT/SBS by mischaracterizing it as prosecutorial overreaching that this Article proposes an evidence-based approach to these interdependent scientific and legal questions.

Evidence-based medicine is a widely used and commonly misunderstood concept. According to Dr. David Sackett, the originator of the concept, “[e]vidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research.”  

This Article’s approach adopts the fundamental principle of evidence-based review, which is that not all evidence is of equal validity. Thus, in Smith we compare the clinical experience of the five diagnosing physicians and examine whether their methods and conclusions are consistent with the evidence base that supports the accuracy of the AHT/SBS diagnosis. This methodology is a direct response to the Smith dissenters’ myopic review of the facts, which implicitly endorses efforts to “misrepresent the state of knowledge in the medical community regarding the reality of child abuse, and, in particular, AHT/SBS.”  

This methodology also

---

53 See infra Part II.B (discussing State v. Edmunds, 746 N.W.2d 590 (Wis. Ct. App. 2008)).


55 Goldberg, supra note 18 (including a comment from Lucy B. Rorke-Adams, M.D., Senior Neuropathologist, The Children’s Hospital of Philadelphia, Consultant Neuropathologist, Office of the Medical Examiner of Philadelphia). Dr. Rorke-Adams continued,

There has been a growing body of well-done scientific work confirming the various aspects of this syndrome, specifically, the pathogenesis of subdural hematomas, the nature of the retinal hemorrhages and more serious retinal injury, the pathophysiology of the concussion which results from shaking, the reality of the severe spinal cord injury[,] etc. It is well-recognized that there may be mimics of SBS and physicians dealing with the clinical and pathological diagnosis of babies presenting with signs of abuse meticulously explore all other possibilities before concluding that the illness or death is not a consequence of natural causes.

The small number of “experts” who challenge this enormous body of evidence are, for the most part, not involved in the day-to-day practice of pediatrics, neurosurgery or pediatric forensic pathology/neuropathology.
reclaims and clarifies the evidence-based approach by identifying the *Smith* dissenters’ unwarranted reliance on work that purports to be “evidence-based,” but is not.56

Toward this end, Part II provides a brief historical perspective on the genesis of the false AHT/SBS controversy. Part III addresses how the dissenting justices distorted the medical evidence introduced by the prosecution and the defense and provides an evidence-based response to each of the five specific “problems” with the medical evidence identified by the *Smith* dissenters. Part IV examines how the Supreme Court justices who dissented in *Smith* also misinterpreted the nonmedical evidence introduced at trial. Finally, the conclusion identifies the *Smith* dissent as an example of the more general “tail wagging the dog” problem that arises during postconviction review when judges fail to properly apply the legal standard, and opt instead to engage in extrarecord fact-finding used to support the speculation that a conviction was wrongful. A companion article explores the additional and more general problems created by the *Smith* dissenters’ independent review of the medical literature.57

II. THE HISTORY OF THE FALSE AHT/SBS “CONTROVERSY”

A thorough examination of the history of the false AHT/SBS controversy is beyond the scope of this Article. It is helpful, however, to explore the critical events that brought us to where we are today.

As noted above, the original articles recognizing SBS were published in the early 1970s.58 Since then, the empirical evidence supporting the validity of the AHT/SBS diagnosis has grown to include a body of “evidence-based, peer-reviewed medical literature with 40 years of contributions by pediatricians, neuroradiologists, clinical and forensic pathologists, ophthalmologists, and physiologists clearly supporting the construct of a medical diagnosis for AHT.”59

At the clinical level, “[m]edical experts agree with the physical, laboratory, and imaging findings associated with the medical construct of AHT, which can include subdural hemorrhage, retinal hemorrhage, encephalopathy, and often evidence of previous trauma or other bodily injury.”60 Contrary to the inaccurate representations regularly made by many defense-retained medical witness and legal academics, an AHT/SBS diagnosis is not the automatic result of the so-called triad of injury findings.61 Instead, it is the product of an extensive differential

---

56 See Mark Donohoe, *Evidence-Based Medicine and Shaken Baby Syndrome*, 24 AM. J. FORENSIC MED. & PATHOLOGY 239 (2003); *infra* text accompanying note 114.
57 See Moreno & Holmgren, * supra* note 17.
58 See * supra* note 8 and accompanying text.
59 Albert et al., * supra* note 33 at 39.
60 Id.
61 In the opinion of the Medill Innocence Project and Professor Alec Klein, “If the classic triad of shaken-baby syndrome symptoms is present—retinal bleeding, brain swelling and brain bleeding—it is often assumed that a caregiver caused them. But science
diagnostic process that considers all aspects of the medical evaluation (i.e., clinical evaluation, clinical history, diagnostic and laboratory testing, evidence of cerebral trauma, evidence of ocular trauma, imaging studies, evidence of additional injuries, and, in the event of death, pathology and autopsy results) along with all of the investigative information provided by relevant nonmedical sources. Thus, the repeated law professor/law student complaint that criminal convictions are based exclusively on so-called triad evidence is false. More recent academic claims that has evolved and some of the assumptions are being challenged.” Rethinking Shaken Baby Syndrome Convictions: Medill Innocence Project Begins Investigating Potential Miscarriages of Justice, NW. UNIV. (Oct. 2, 2012), http://www.northwestern.edu/newscenter/stories/2012/10/rethinking-shaken-baby-syndrome-convictions.html; see also Symposium, supra note 23, at 222–23 (statement of Professor Keith Findley) (asserting that the traditional theory is that the triad is caused exclusively by shaking and is used to establish all the elements of the crime, and agreeing with Professor Tuerkheimer’s assertion that the traditional theory amounts to “medically diagnosed murder”); Tuerkheimer, The Next Innocence Project, supra note 23, at 6–18 (claiming the medical diagnosis of SBS and criminal convictions are based almost exclusively on the triad); Tuerkheimer, Science-Dependent Prosecution, supra note 23 at 515 (claiming “the classic formulation of SBS is based exclusively on the diagnostic ‘triad’”).

62 See Maguire et al., supra note 37 (noting the positive predictive value of retinal hemorrhages and apnea in AHT, but explaining that, as in all cases of physical child abuse, there is no diagnostic test for inflicted brain injury and the diagnosis is made on the basis of probability after careful exclusion of other possible causes for the clinical findings including accidental injury and other medical conditions). For an excellent discussion of the differential diagnostic process engaged in by medical professionals, and the consideration of alternative causes, see Stephen C. Boos, Abusive Head Trauma as a Medical Diagnosis, in ABUSIVE HEAD TRAUMA IN INFANTS AND CHILDREN: A MEDICAL, LEGAL, AND FORENSIC REFERENCE, supra note 9, at 49; Andrew P. Sirontak, Medical Disorders that Mimic Abusive Head Trauma, in ABUSIVE HEAD TRAUMA IN INFANTS AND CHILDREN: A MEDICAL, LEGAL, AND FORENSIC REFERENCE, supra, at 191. This differential diagnostic process is not new and has been well described for over two decades. See generally Carolyn J. Levitt, Wilbur L. Smith & Randell C. Alexander, Abusive Head Trauma, in CHILD ABUSE, MEDICAL DIAGNOSIS AND MANAGEMENT 1–23 (Robert M. Reece ed., 1994) (describing the differential diagnostic process).

63 See Symposium, supra note 23, at 224–25 (statement of Professor Keith Findley); Tuerkheimer, The Next Innocence Project, supra note 23, at 5 (explaining that “[w]ith rare exception, the case turns on the testimony of medical experts” and “SBS comes as close as one could imagine to a medical diagnosis of murder: prosecutors use it to prove the mechanism of death, the intent to harm, and the identity of the killer” because “all elements of the crime—mens rea and actus reus (which includes both the act itself and causation of the resulting harm)—are proven by the science”). These are strong allegations, which make it surprising that the only support provided by Professor Tuerkheimer is a citation to two cases, neither of which represents a triad-only prosecution. See Tuerkheimer, The Next Innocence Project, supra note 23, at 7 (citing Mitchell v. State, No. CACR 07-472, 2008 WL 316166, at *1 (Ark. Ct. App. Feb. 6, 2008), and State v. Edmunds, 598 N.W.2d 290, 293 (Wis. Ct. App. 1999)). Professor Findley also mischaracterizes Edmunds as a triad-based prosecution, see Symposium, supra note 23, at 224 (statement of Professor Keith Findley), despite the fact that having handled her postconviction petition he is well aware
pediatricians and emergency room physicians who diagnose AHT/SBS are tainted by a “prosecution bias” are empirically unsupportable, a transparent attempt to garner attention with faddish behavioral science jargon, and patently absurd.64

A. Commonwealth v. Woodward65

In 1997, Louise Woodward’s prosecution brought the term “shaken baby syndrome” into the national spotlight. The highly publicized case of the English au pair who fatally shook and slammed eight-month-old Matthew Eappen brought international public attention to the reality of infant AHT. The hotly contested trial also brought national attention to the use of highly paid defense medical witnesses to challenge the accuracy of a child abuse diagnosis and to advance outlier and highly controversial “alternative theories” of causation that purport to explain traumatic or fatal infant injuries including, in this case, a two-inch skull fracture. Woodward marks the origin of the false AHT/SBS controversy—at least in part because the defendant, who was convicted of second-degree murder by a jury, was later freed by the judge.66 This fact alone could explain the resulting public uncertainty regarding the weight of the prosecution’s medical evidence.

of the extensive additional evidence of Edmunds’s guilt, including evidence that she was previously witnessed assaulting another child in her care, Edmunds’s own inculpatory trial testimony, and the medical evidence of impact injury. See Edmunds v. Deppisch, 313 F.3d 997 (7th Cir. 2002), cert. denied, 538 U.S. 1066 (2003) (commenting on the findings made by the trial judge on this issue). See also infra notes 103–107 and accompanying text noting Professor Tuerkheimer’s erroneous assertions in this regard. Brian Holmgren, a coauthor on this Article, has twenty-five years of experience prosecuting child abuse cases. He has never prosecuted a triad-only case and has yet to find one reported in the appellate reports.

Symposium, supra note 23, at 232–33 (statement of Professor Keith Findley) (arguing that doctors make two types of proprosecution mistakes: (1) “selection bias” mistakes when they consider confession evidence as relevant to an abuse diagnosis, and (2) “observer bias” mistakes because doctors, in their view, prefer to diagnose abuse and “you’re more likely to see something when you want to see it”); Tuerkheimer, The Next Innocence Project, supra note 23 at 6 (claiming the medical diagnosis of SBS has been “corrupted by a too-close medical-legal nexus”).


Carey Goldberg, Massachusetts High Court Backs Freeing Au Pair in Baby’s Death, N.Y. TIMES, June 17, 1998, at A1. At trial, the prosecutor in the Woodward case challenged Judge Hiller Zoebel’s reduction of Woodward’s second degree murder conviction to manslaughter, the lighter associated sentence allowed her to return to England. The prosecution alleged judicial bias, in part based on the fact that prior to the trial, Judge Zoebel inappropriately proposed that the prosecutor offer a plea to manslaughter. See Commonwealth v. Woodward, 694 N.E.2d 1277, 1277 (Mass. 1998). But the sentence reduction does not imply that Woodward did not cause the injuries or death of Matthew Eappen, or that the prosecution’s medical proof was questionable. Id. (setting forth Judge Zoebel’s reasons for denying Woodward’s motion for acquittal and a new trial).
Irresponsible journalists, however, including Mike Wallace of 60 Minutes, exacerbated the confusion.\textsuperscript{67}

The public confusion that began with Woodward has been transformed into an apparent AHT/SBS “controversy” by some of the doctors who testified in Woodward\textsuperscript{68} who, along with a handful of others, have spent the past fifteen years providing a plethora of defense-supported medical challenges in child homicide and child abuse cases. These include contesting the admission of the diagnosis of AHT/SBS in pretrial hearings, Daubert/Frye evidentiary challenges,\textsuperscript{69} the increased use of medical witnesses to testify to outlier or unsubstantiated causation theories at trial, and a concomitant increase in postconviction challenges based on claims of a paradigm shift in the medical community.\textsuperscript{70} To date, most courts have accurately concluded that the diagnosis of AHT/SBS is supported by valid medical evidence and is generally accepted in the relevant medical community.\textsuperscript{71}

\textsuperscript{67} See Bill Carter, Media Talk; Irate Boston Media Fault '60 Minutes' For Nanny Report, N.Y. TIMES, Mar. 15, 1999, at C9. In March 1998, following Louise Woodward’s conviction, Mike Wallace of 60 Minutes broadcast an interview with Dr. Floyd Gillis, which included the doctor’s opinion that Matthew Eappen did not die of AHT/SBS, but instead was strangled to death, an opinion offered by none of the witnesses at trial. Id. Mr. Wallace failed to disclose, however, that Dr. Gillis’s opinion was based on the opinion of Dr. Marvin Nelson—a paid consultant for the defense in the Woodward case. Id. Thus, Mike Wallace’s misleading report directly contributed to the growing national confusion regarding the validity of the medical evidence offered to support the AHT/SBS diagnosis that had been presented at trial. Id. For published discussion of those medical findings, see Patrick D. Barnes & Caroline D. Robson, CT Findings in Hyperacute Nonaccidental Brain Injury, 30 PEDIATRIC RADIOLOGY 74, 75–80 (2000) [hereinafter Barnes & Robson, CT Findings]; Patrick D. Barnes & Caroline D. Robson, An Unresponsive Infant in the Emergency Room, 6 SEMINARS IN PEDIATRIC NEUROLOGY 225, 225–27 (1999). Dr. Barnes was a prosecution expert at Woodward’s trial and these papers discuss the medical findings from the case. For an expression of the medical consensus on these injuries and a response to the proffered claims of the defense witnesses at the Woodward trial, see David L. Chadwick et al., Shaken Baby Syndrome—A Forensic Pediatric Response, 101 PEDIATRICS 321, 321–23 (1998) (critiquing the accuracy of medical testimony provided by defense witnesses during the Woodward trial and signed by over seventy physicians).

\textsuperscript{68} Two of these doctors are Dr. Jan Leestma and Dr. Ronald Uscinski, who have been frequent defense witnesses since Woodward and whose papers are cited by the Smith dissenters. See Moreno & Holmgren, supra note 17 (discussing the controversial positions and testimony of Drs. Leestma and Uscinski).

\textsuperscript{69} See, e.g., Commonwealth v. Martin, 290 S.W.3d 59, 69 (Ky. Ct. App. 2008) (reversing the trial court’s exclusion of testimony on SBS following a Daubert hearing at which Dr. Ronald Uscinski testified for the defense and challenged the scientific reliability of SBS); Transcript of Motion Hearing at 5–51, State v. Mendoza, No. 071908696 (Utah Dist. Ct. Nov. 10, 2008) (testimony of Dr. Janice Ophoven) (supporting a motion to exclude an SBS diagnosis on Daubert grounds).

\textsuperscript{70} State v. Edmunds, 746 N.W.2d 590, 592–93, 599 (Wis. Ct. App. 2008).

\textsuperscript{71} See, e.g., Mitchell v. State, No. CACR 07-472, 2008 WL 316166, at *3 (Ark. Ct. App. Feb. 6, 2008) (rejecting the defense claim that a Daubert hearing was required before testimony on SBS may be admitted); Grant v. Warden, No. TSRCV030004233S, 2008
knowledge, not a single appellate court has rejected evidence of AHT/SBS as scientifically unreliable under the Daubert or Frye evidentiary standard—an outcome one would expect if the science was truly unsound.

At least one trial judge with extensive experience in child abuse cases has opined that witnesses (and even legal academics) who misrepresent the AHT/SBS medical evidence create real problems for the courts.\textsuperscript{72} According to Judge Gill, “It is disconcerting, if not frightening, when a law professor professes factual, technical, and legal misleading statements in public and professional publications.”\textsuperscript{73} The Smith dissent has the capacity to create far greater problems as the most influential, if not the first, appellate decision to endorse the false AHT/SBS controversy.\textsuperscript{74}

\begin{itemize}

\textsuperscript{72} See Goldberg, supra note 18.

\textsuperscript{73} Charles Gill, Comment to Goldberg, supra note 18.

\textsuperscript{74} See infra Part II.B (discussing Edmunds, 746 N.W.2d at 590); see also People v. Rector, 226 P.3d 1170, 1173–75 (Colo. Ct. App. 2009), rev’d, 248 P.3d 1196 (Colo. 2011) (reversing because trial judge failed to conduct hearing to determine whether doctor’s testimony on nonaccidental head trauma was based on reliable scientific principles, failing to appropriately consider holding in People v. Martinez, 74 P.3d 316 (Colo. 2003) approving similar testimony); Hamilton v. Commonwealth, 293 S.W.3d 413, 420 (Ky. Ct. App. 2009) (holding that it was error to permit testimony on SBS without first conducting a Daubert hearing because no Kentucky case had specifically determined it was a reliable theory); State v. Schoonmaker, 176 P.3d 1105, 1114–16 (N.M. 2008) (reversing conviction based on ineffective assistance of counsel claim for defense attorney’s failure to get an expert to support defense that a five-week-old baby fell off couch and citing defense expert
B. State v. Edmunds

Seven years ago, Professor Keith Findley enlisted the help of Wisconsin Innocence Project students in his effort to seek postconviction review for Audrey A. Edmunds. Edmunds had been convicted of reckless homicide based on evidence that seven-month-old Natalie Beard suffered both shaking and impact head trauma while in the defendant’s care. Her conviction was affirmed on direct appeal, and her subsequent federal habeas petition was denied. In 2006, she filed a motion for a new trial asserting “there were significant developments in the medical community around ‘shaken baby syndrome’ in the ten years since her trial that amounted to newly discovered evidence.” In support of this petition, Edmunds presented affidavits and testimony from several medical witnesses who claimed that the science of AHT/SBS had “changed” since the time of her trial and now supported alternative theories of causation for many of the medical findings relied upon by the jury. The trial judge, Daniel Moeser, denied Edmunds’ petition holding that, although some of the defense medical witnesses were “credible,” the medical science supporting the conviction was, if anything, stronger a decade later, and that evidence of the defendant’s guilt was extensive. Judge Moeser detailed his findings in an extensive written order.

testimony in other cases and inapplicable defense literature as support for holding); State v. Louis, 798 N.W.2d 319 (Wis. Ct. App. 2011) (unpublished table decision) (upholding trial court’s order for new trial based on erroneous determinations by the trial and appellate court “that medical community is sharply divided” on diagnostic findings for AHT).

75 746 N.W.2d 590 (Wis. Ct. App. 2008).
76 See id. at 592. In fact, Judge Dykman, writing for the court, actually thanks the Wisconsin law school students by name in the text of the decision. Id. at 592 n.1.
77 See id.
78 State v. Edmunds, 598 N.W. 2d 290, 299 (Wis. Ct. App. 1999)
79 See Edmunds v. Deppisch, 313 F.3d 997, 997–98 (7th Cir. 2002).
80 Edmunds, 746 N.W.2d at 593.
81 See id. Edmunds presented affidavits and testimony from the following defense medical witnesses (none of whom regularly treat infants with traumatic injuries): Dr. Patrick Barnes (a pediatric neuroradiologist), Dr. Horace Gardner (a retired military ophthalmologist), Dr. John Galaznik (a university health services physician), Dr. Peter Stephens (a forensic pathologist), and Dr. George Nichols (a forensic pathologist). Edmunds also presented evidence from the original medical examiner, Dr. Robert Huntington, who apparently had modified his position on the potential timing of Natalie’s fatal head injuries but maintained that her cause of death was AHT. See id.
82 Id. The state court summarized the defense testimony as arguing that “there is now a significant debate in the medical community as to whether Natalie’s symptoms were necessarily indicative of shaking or shaking combined with head trauma in infants. The experts explained that there was not a significant debate about this issue in the mid-1990s and that the opinions offered in Edmunds’s first postconviction motion would have been considered minority or fringe medical opinions.” Id.
84 See id.
The Wisconsin Court of Appeals reversed Judge Moeser, finding that Edmunds was entitled to a new trial because the “newly discovered evidence in this case shows that there has been a shift in mainstream medical opinion since the time of Edmunds’s trial as to the causes of the types of trauma Natalie exhibited.”85 According to the court of appeals, Judge Moeser erred because he “expressly found that Edmunds’[s] new evidence and the State’s new evidence were both credible . . . [and] then weighed the evidence and concluded that the State’s evidence was stronger.”86 The court of appeals held that “it was not the [trial] court’s role to weigh the evidence . . . it was required to determine whether there was a reasonable probability that a jury, hearing all the medical evidence, would have a reasonable doubt as to Edmunds’s guilt.”87 Thus, the court of appeals’s narrow and technical ruling was based solely on speculation that “a jury could have a reasonable doubt as to a defendant’s guilt even if the State’s evidence is stronger,”88 so Edmunds was entitled to a new trial.89

A thorough discussion of Edmunds is beyond the scope of this Article, but because this case is invariably cited as evidence that courts are starting to reject the science of AHT/SBS, some clarification is required. The trial judge’s opinion that a few defense witnesses were “credible” provides the only support for the sweeping and erroneous conclusion by the appellate court that “newly discovered evidence in this case shows that there has been a shift in mainstream medical opinion,”89 which in turn, presages the similarly erroneous conclusion of the Smith dissenters. This conclusion must be understood in context as the ipse dixit of a few defense witnesses who testified in Edmunds’s post-conviction hearing for the purpose of advancing self-serving (but empirically unsupportable) claims that a shift in “mainstream medical opinion” is underway.

Edmunds should not be misunderstood as heralding a significant change in medical opinion on SBS/AHT. The mere fact that a single state court judge found a

85 Edmunds, 746 N.W.2d at 598–99.
86 Id. at 597.
87 Id.
88 Id.
89 Id. at 599. Of course, this ruling was made under the court’s unique interpretation of Wisconsin’s procedural rules for postconviction petitions alleging “newly discovered evidence.” Thus, this ruling has little or no precedential value for other jurisdictions that use different rules. Moreover, the Wisconsin rules create an odd outcome by mandating that any time a trial court determines that a defendant’s postconviction “expert” is credible, or even that the evidence presented by the state is stronger, a new trial is required. The problems created by these state rules should be readily apparent to anyone familiar with postconviction litigation—especially in AHT/SBS cases. They were certainly obvious to the judge in Grant v. Warden, No. TSRCV030004233S, 2008 Conn. Super. LEXIS 1402 (Conn. Super. Ct. June 4, 2008), who immediately after Edmunds opined that “[t]he Edmunds case presents a potential quagmire of epic proportions: the strong likelihood of constant renewed prosecution and relitigation of criminal charges as expert opinion changes and/or evolves over time” and that “the strong interest in the finality of judgments is significantly undermined by reasoning employed by the Edmunds court.” Id. at *2 n.1.
90 Edmunds, 746 N.W.2d at 598–99.
small group of doctors (of whom only one was actively engaged in the diagnosis of child abuse) “credible” is not evidence of a paradigm shift—especially when the judge’s opinion is contradicted by four decades of scientific consensus on the AHT/SBS diagnosis across a wide range of pediatric medical subspecialties and countless physicians who are more credible because they actually diagnose abuse as part of their clinical medical practice. The extensive biomechanical research on infant head trauma also provides sound additional reasons to suspect the accuracy of Judge Moeser’s credibility conclusion.91

91 For example, several of the purportedly credible defense witnesses testified that shaking alone could not have caused Natalie’s injuries and that biomechanics research did not support the SBS diagnosis. See, e.g., Transcript of Evidentiary Hearing (Day 1) at 29–30, 37, State v. Edmunds, No. 96-CF-555 (Wis. Cir. Ct. 1997) (testimony of Dr. Patrick Barnes); Transcript of Evidentiary Hearing (Day 2) at 133–35, 147–48, State v. Edmunds, No. 96-CF-555 (Wis. Cir. Ct. 1997) (testimony of Dr. George Nichols). This evidence should not have confirmed the witnesses’ credibility, but should instead have raised concerns for three reasons. First, Natalie had evidence of blunt impact trauma to her head, making her case not a “shaking alone” case, but a “shaking plus impact” case, which (as the defense witnesses surely must have known but opted not to share with the court) made these biomechanics arguments irrelevant. Second, because the biomechanics literature predated Edmunds’s trial, it was not “newly discovered evidence.” See generally Ann-Christine Duhaime et al., The Shaken Baby Syndrome: A Clinical, Pathological and Biomechanical Study, 66 J. NEUROSURGERY 409 (1987) (reviewing forty-eight cases of infants and young children with SBS, including scans showing brain hemorrhaging, subdural or subarachnoid hemorrhaging, autopsies, and reporting on various biomechanical experiments of shaking and impact with a surrogate infant). Third, none of the well-documented critiques of the biomechanical research were acknowledged or addressed by the defense witnesses. See generally SHAKING AND OTHER NONACCIDENTAL HEAD INJURIES IN CHILDREN (Robert A. Minns & J. Keith Brown eds., 2005) (summarizing multiple limitations regarding Dr. Duhaime’s conclusions and listing contrary clinical and biomechanical evidence); C.Z. Cory & M.D. Jones, Can Shaking Alone Cause Fatal Brain Injury?: A Biomechanical Assessment of the Duhaime Shaken Baby Syndrome Model, 43 MED. SCI. & L. 317 (2003) (discussing the Duhaime model and concluding “[t]here must now be sufficient doubt in the reliability of the Duhaime et al. (1987) biomechanical study to warrant the exclusion of such testimony in cases of suspected shaken baby syndrome.”); D.R. Wolfson et al., Rigid-Body Modelling of Shaken Baby Syndrome, 219 J. ENGINEERING MED. 63 (2005) (discussing the use of rigid-body modeling to investigate neck stiffness on head motion and head-torso impacts as a possible mechanism of injury). In addition, defense witnesses opined that if Natalie were shaken violently, she should have sustained neck injuries, see Transcript of Evidentiary Hearing (Day 2), supra, at 153 (testimony of Dr. George Nichols), an outlier view supported by a single paper rife with methodological errors and the subject of extensive scathing criticism. See also supra note 17 (discussing critiques of Bandak, supra note 16). Finally, the defense witnesses’ credibility should have been further undermined by the fact that they proposed numerous alternative theories to explain Natalie’s injuries (i.e., a dysphagic choking episode, a “rebleed” of a prior subdural hematoma, irritation of her respiratory centers from a subarachnoid hemorrhage, hypoxia), see id., despite the fact that these theories could not accurately account for Natalie’s injuries, they were radically divergent and logically inconsistent, and none were substantiated by accepted medical literature.
C. Law Professors and Students Add to the Growing Confusion

If Woodward marks the advent of public awareness and confusion regarding AHT/SBS, Edmunds marks the tipping point for the new false controversy. Starting in 2009, a small group of law professors and law students began to advance the view that wrongful convictions in child homicide and child abuse cases were creating a “criminal justice crisis.”

Law professors’ and students’ recent self-serving attempts to garner attention with arguments that AHT/SBS “quite possibly does not exist,” may be “junk science,” that “SBS science in its current conflicted state . . . does not support criminal convictions,” and that the medical community has “deliberately discarded a diagnosis defined by shaking,” have provoked extensive criticism from all corners of the pediatric medical community. According to Allison Scobie-Carroll, program director for the child protection program at Children’s Hospital Boston, “For those [of] us who actually see these children, there is no debate.”

---


93 See sources cited supra note 23.

94 See sources cited supra note 24.

95 See Tuerkheimer, Science-Dependent Prosecution, supra note 23, at 550 (describing the “national crisis in forensic science” as including “the SBS context [because] there is a tension between the criminal law’s treatment of forensic science claims and what scientists now know about the validity of these [SBS] claims”).

96 Lyons, supra note 24, at 1109.

97 Imwinkelreid, supra note 23, at 158 (“The question here is whether shaken baby syndrome evidence is ‘junk’ science presenting an intolerable risk of a wrongful conviction . . . .”).

98 Orenstein, supra note 24, at 1306.


100 Carey Goldberg, Pediatrics Academy President-Elect on “Shaken Baby Syndrome,” WBUR’S COMMONHEALTH REFORM & REALITY (Oct. 4, 2010, 2:53 PM),
According to Dr. Robert W. Block of the American Academy of Pediatrics, the so-called AHT/SBS controversy was created by a few people who ignore the known science and excuse the confessional literature, the clinical experience that many of us have working with babies who are injured or killed and the people who hurt them, and choose instead to come up with alternative hypotheses, none of which are substantiated by reasonable science.101

Dr. Desmond Runyan, M.D., Professor of Social Medicine and Pediatrics at the University of North Carolina School of Medicine, specifically criticized Professor Tuerkheimer’s effort to discredit the diagnosis of AHT/SBS as based on “factual an[d] conceptual errors” including her misuse of his own research. 102 According to Dr. Runyan:

Where Professor Tuerkheimer states that there are more than 1000 babies labeled with the diagnosis of shaken baby, she used an estimate derived from a study that I helped design and conduct. In North Carolina we had 80 cases of children less than 2 years of age who were diagnosed as having suffered an inflicted traumatic brain injury in 2000 and 2001. This yielded a rate of 17/100,000 children in the first 2 years of life. In that study we undertook a careful blinded secondary review of the case circumstances and found that the diagnoses were accurate and well done and the[re were] only two cases that we found mis-identified. In one case we determined it was most likely an accident and in another we concluded [t]he case was likely abusive[.] In both cases the physician involved called it “undetermined.” From the 80 North Carolina cases that constituted our sample over two years, only 54 cases proceeded to the courts. In 80% of those cases there was a guilty plea by the perpetrator. Thus, there were six trials each year and five convictions.103

Thus, according to Dr. Runyan, Professor Tuerkheimer improperly based her claim of 1,000 SBS diagnoses on his work. Dr. Runyan indicated that, when Professor Tuerkheimer cited his research, she omitted any data that would have undermined her argument. Professor Tuerkheimer failed to acknowledge the critical fact that in the twelve North Carolina prosecutions that went to trial, “[t]he information the juries used went far beyond Prof. Turkheimer’s ‘triad’ and included careful histories, other injuries, and diagnostic testing to rule-out other diagnoses.”104


101 Id.
102 Desmond Runyan, Comment to Goldberg, supra note 18.
103 Id.
104 Id.
Thus, according to Dr. Runyan, his research does not support “[h]er premise that there are many hundreds of people languishing in jails with false convictions.”

A more general criticism of Professor Tuerkheimer and everyone else who has opined that there is a “criminal justice crisis” of false AHT/SBS convictions, is that child physical abuse cases are significantly underreported and therefore underprosecuted. The fact that law professors are, at least in part, promulgating the false AHT/SBS controversy is deeply troubling to Judge Gill who believes that Professor Tuerkheimer’s over-extension of the highly questionable medical minority view on the subject into the legal world . . . are not medically, scientifically or legally correct. They suggest a legal tilting at her new “innocence project.” . . . But her project is guilty of existing pretty much in her own mind. Her sources are scant and wrong.

With no support from the pediatric subspecialist communities, legal academics seeking to advance the AHT/SBS false controversy have been (unsurprisingly) forced to rely on the same handful of defense-employed witnesses who regularly testify for the defense in child abuse and child homicide cases for all of their “scientific” support. Since the Woodward trial, these medical witnesses have published articles challenging the diagnosis of AHT/SBS and proposing a range of alternative causal theories to explain traumatic and deadly brain injuries in infants. These articles provide publications that can be cited from the witness stand to support defense-sponsored testimony opining that an infant’s traumatic brain injury was not caused by abuse. However, judges and juries are apparently unaware of the fact that, with few exceptions, these papers have encountered overwhelming evidence-based critique from a broad range of medical

---

105 Id.
106 See Albert et al., supra note 33, at 40 (“The reported rates of child abuse resulting in homicide have been documented to be underreported by as much as 50% to 60%.”).
107 Gill, supra note 73.
108 Dr. Robert W. Block explained that it is critical to recognize that “the real experts are the physicians who work every day with these cases and have both authored and read voluminous literature that substantiates the existence of abusive head trauma, and those are the folks who are the most capable of informing the public about what the issue really is” so as to distinguish them from the law professors or students or both who “create[e] these sham media blasts that [cause] great confusion.” See Goldberg, supra note 100.
professionals. They are also generally viewed as being written for the purpose of maintaining or increasing the authors’ lucrative defense witness appearances.  

As the preceding discussion illustrates, the outlier views expressed by defense medical witnesses and parroted in law review articles is self-validating. The

110 Many articles that are written and then cited by defense experts fall far short of the type of peer-reviewed scientific literature standards contemplated by the Supreme Court in Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993). As discussed above, frequent defense witness Dr. Patrick Barnes has, at least once, misleadingly suggested that an article was published in a peer-reviewed journal, while failing to explain that the article, because it was a commentary piece, had not been subject to peer review. See Keller & Barnes, supra note 48 and accompanying text. Dr. Barnes also published articles (single case reports) based on cases where he served as a defense witness. See Patrick D. Barnes et al., Infant Acute Life-Threatening Event—Dysphagic Choking Versus Nonaccidental Injury, 17 SEMINARS PEDIATRIC NEUROLOGY 7 (2010) [hereinafter Barnes et al., Dysphagic Choking]; Patrick D. Barnes et al., Traumatic Spinal Cord Injury: Accidental Versus Nonaccidental Injury, 15 SEMINARS PEDIATRIC NEUROLOGY 178 (2008). Both articles were published in a topical journal that has publishing guidelines that do not include any form of peer-review process before publication. See NEUROLOGY ADVANCE, http://www.neurologyadvance.com/content/neurologyadvance-journalspage-paedia (last visited June 13, 2013). Dr. Christopher S. Greeley has addressed the problems with Barnes’s Dysphagic Choking at some length, noting that Dr. Barnes and his coauthors (1) omitted evidence of additional salient abuse injuries to the child; (2) omitted the fact that the case resulted in a child abuse trial, see Thomas v. State, No. 03-07-00646-CR, 2009 WL 1364348 (Tex. Ct. App. May 14, 2009); (3) omitted the fact that the defendant was convicted of child abuse; (4) omitted the fact that the defendant’s conviction was affirmed on appeal; and (5) failed to reveal their own roles as defense witnesses at trial. See Dr. Christopher S. Greeley, Letter to the Editor, 17 SEMINARS PEDIATRIC NEUROLOGY 275 (2010). There is also evidence that some authors, whose work has been published in non-peer-reviewed journals or has been subjected to critical review, try to deter readers from discovering this type of discrediting evidence. For example, in a more recent paper, Dr. Barnes again postulated “dysphagic choking” as a mimic of AHT/SBS. Patrick D. Barnes, Imaging of Nonaccidental Injury and the Mimics: Issues and Controversies in the Era of Evidence-Based Medicine, 49 RADIOLOGIC CLINICS N. AM. 205 (2011) [hereinafter Barnes, Imaging of Nonaccidental Injury]. However, rather than citing to his published paper on this topic (which would inevitably reveal Dr. Greeley’s discrediting critique), he cited instead to a similarly titled conference presentation and failed to address (or even mention) Dr. Greeley’s critique. Id. at 220 n.167.

111 A review of the medical sources cited in the recent spate of law review articles cited supra notes 23–24 reveals a common practice of citing to the same handful of defense witnesses for “scientific” support. Notably absent is any recognition of the abundant medical literature supporting the clinical and empirical validity of the AHT/SBS diagnosis. This form of selective citation is, as Professor Ceccarelli notes, a hallmark of manufactured scientific controversy. See Ceccarelli, supra note 1, at 196 and accompanying text. While Professor Tuerkheimer was writing her first article, The Next Innocence Project: Shaken Baby Syndrome and the Criminal Courts, supra note 23, she contacted one of this Article’s authors, Brian Holmgren, for information and she was provided with over two-hundred pages of materials documenting the medical literature supporting the diagnosis of AHT/SBS and critiquing many of the sources she relied upon in her work. This material is
academics cite the same handful of defense medical witnesses, the media cites both, the defense medical witnesses benefit from the publicity and are hired in more cases, and the cycle begins anew.

The most recent law review articles eliminate the circularity in favor of direct collaboration.112 With everyone on the same page, law professors and defense experts work together to promote doubt and uncertainty that (1) AHT/SBS is an anecdotal medical diagnosis without scientific proof;113 (2) there is no “evidence-based” medical research to support the AHT/SBS diagnosis;114 (3) adults cannot conspicuously absent from all of Professor Tuerkheimer’s academic work on AHT/SBS to date.

112 See Symposium, supra note 23 (transcribing remarks made during a joint presentation with frequent defense medical witness Patrick Barnes who Professor Findley had employed to testify on behalf of Audrey Edmunds); Findley et al., supra note 23.

113 Findley et al., supra note 23, at 299–300 (“[T]he real problem is that the literature cited in support of the SBS/AHT hypothesis falls at the bottom of the hierarchy of evidence and rests almost entirely on assumptions and hypotheses, combined with emotionally compelling demonstrations and anecdotal evidence, largely in the form of confessions.”); Barnes, Imaging of Nonaccidental Injury, supra note 110 (characterizing “much of the traditional literature on child abuse . . . [as] anecdotal case series, case reports, reviews, opinions, and position papers”); J.F. Geddes & J. Plunkett, The Evidence Base for Shaken Baby Syndrome, 328 BMJ 719, 719–20 (2004) (suggesting that the medical literature and diagnostic criteria supporting the AHT/SBS diagnosis are inadequate and flawed and expressing doubt about the existence of AHT/SBS); Jan E. Leestma, Child Abuse: Neuropathology Perspectives, in FORENSIC NEUROPATHOLOGY 561, 602 (Jan E. Leestma ed., 2d ed. 2008) (characterizing the medical literature linking shaking to pathology-injury findings as deficient and anecdotal).

114 See, e.g., Findley et al., supra note 23, at 237–38 (claiming that Dr. Mark Donohoe—the author of Evidence-Based Medicine and Shaken Baby Syndrome, supra note 56—“examined the research support for SBS through 1998 and concluded what others—including the NIH conference participants—had been saying privately for years: the research basis for shaken baby syndrome was remarkably weak” and citing Dr. Donohoe’s conclusion that “the commonly held opinion that the finding of [subdural hemorrhage] and [retinal hemorrhage] in an infant was strong evidence [of] SBS was unsustainable, at least from the medical literature”).

Defense medical witnesses and legal academics routinely rely on Dr. Donohoe’s single three-page article to support their assertions that AHT/SBS does not exist. For examples from the medical witness articles, see Barnes et al., Dysphagic Choking, supra note 110, at 1; Barnes, Imaging of Nonaccidental Injury, supra note 110, at 206; Geddes & Plunkett, supra note 113, at 719–20; Leestma, supra note 51, at 14; Leestma, supra note 109; Jan Leestma, The So-Called “Shaken Baby Syndrome”: A Concept Unsupported by Science and the Facts, IND. DEFENDER, March 2006, at 1; James LeFanu, Wrongful Diagnosis of Child Abuse—A Master Theory, 98 J. ROYAL SOC’Y MED. 249 (2005); Marvin Miller et al., A Sojourn in the Abyss: Hypothesis, Theory, and Established Truth in Infant Head Injury, 114 PEDIATRICS 326 (2004) (including as coauthors Patrick Barnes, Jan Leestma, John Plunkett, Ron Uscinski, and several others); Miller & Miller, supra note 109, at 169; Squier, supra note 109, at 11; Uscinski, supra note 109. For examples from the legal literature, see Burg, supra note 24, at 665 n.53–56 and accompanying text; Gena, supra note 24, at 706 n.56, 710 n.95–100, 711 n.101, 727 n.270 and accompanying text;
shake infants hard enough to cause injuries ascribed to AHT/SBS based on biomechanical research;\(^\text{115}\) (4) violent shaking would break the infant’s neck or result in other thoracic injuries;\(^\text{116}\) (5) suspects’ confessions are false, cannot explain the injuries, and result from coercive prosecution-biased interrogation;\(^\text{117}\) (6) AHT/SBS is routinely diagnosed solely on triad findings (i.e., retinal hemorrhages, subdural or subarachnoid hemorrhages, and brain encephalopathy);\(^\text{118}\) (7) the diagnostic triad is nonspecific;\(^\text{119}\) (8) alternative
medical conditions and accidental traumas account for injuries misdiagnosed as AHT/SBS;\textsuperscript{120} and (9) injuries cannot be timed to identify a perpetrator because children can have “lucid intervals” after severe injuries.\textsuperscript{121} The popular press is

(\textit{asserting that AHT/SBS is a “hypothesis that violent shaking may be reliably diagnosed based on the triad of subdural hemorrhage, retinal hemorrhage, and encephalopathy (brain damage) if the caretakers do not describe a major trauma . . . and no alternative medical explanation is identified”}; Tuerkheimer, \textit{The Next Innocence Project, supra} note 23, at 11 (\textit{asserting that “the triad of symptoms was believed to be distinctly characteristic—in scientific terms, pathognomonic—of violent shaking”}).

\textsuperscript{119} For an example from the medical witness articles, see Squier, \textit{supra} note 115, at 520 (characterizing the triad as \textit{not} diagnostic of AHT/SBS). For examples from the legal articles, see Burg, \textit{supra} note 24, at 663–64; Symposium, \textit{supra} note 23, at 224–25 (statement of Professor Keith Findley) (criticizing the validity of the AHT/SBS diagnosis as “based on the belief that the triad elements were . . . traumatic in origin” and that “[b]ecause the brain damage was often bilateral and widespread, it was assumed the force needed . . . was comparable to . . . that found in . . . motor vehicle accidents” and therefore “if the history provided by the caretakers did not include a major accident, the history was considered to be inconsistent with the findings, and abuse was considered to be the only plausible explanation”); Michele Nethercott, \textit{The Role of Forensic Science and Scientific Evidence in the Defense of Criminal Cases, in UTILIZING FORENSIC SCIENCE IN CRIMINAL CASES 7, 15 (2012)} (“[T]he original notion of [AHT/]SBS proponents that an infant who exhibited the triad of symptoms, including a subdural hematoma, retinal hemorrhaging, and a swollen brain, could exhibit this constellation of symptoms only as a result of a violent shaking, which had occurred shortly before the onset of the symptoms, has now been shown to be incorrect.”); Tuerkheimer, \textit{Science-Dependent Prosecution, supra} note 23, at 516–17 (asserting that convictions in cases involving AHT/SBS diagnoses are necessarily suspect because “research has shown that retinal hemorrhages and subdural hematomas can result from forces other than shaking,” thus, “the myth of pathonomony—which told that the diagnostic triad was necessarily and exclusively induced by shaking—has been debunked”).

\textsuperscript{120} For examples from the medical witness articles, see Barnes, \textit{Imaging of Nonaccidental Injury, supra} note 110, at 205–06. For an example from the legal articles see Symposium, \textit{supra} note 23, at 239–40 (statement of Dr. Patrick Barnes) (“[W]e have found with advanced technology . . . that there are a number of conditions that have nothing to do with trauma—for example medical illnesses including infections, bleeding or clotting problems—that can have findings that mimic abuse. And in some very young infants (e.g., under six months of age) who were thought to have been shaken or battered, their symptoms and signs actually extended from birth injuries or conditions. And this not only includes brain injury and bleeding, but also bone injuries or fractures, especially with the more recent revelations that nutritional deficiencies have come back, like vitamin D and vitamin C deficiencies, which can cause Rickets and Scurvy.’’); Squier, \textit{supra} note 115, at 519 (listing numerous alternative theories of causation).

\textsuperscript{121} For examples from the medical witness articles, see Leestma, \textit{supra} note 113, at 577–78, 604–05 (asserting that lucid intervals with subdural hematomas and head injury is a well-known phenomenon); Leestma, \textit{supra} note 51 (asserting that the case literature does not support the conclusion that symptoms are typically immediate). For examples from the legal articles see Symposium, \textit{supra} note 23, at 229–30 (statement of Professor Keith Findley) (claiming that “we have the evidence that lucid intervals are a distinct reality” and
complicit when it presents sympathetic or sensationalistic claims, but ignores the extensive empirical evidence that doctors rely on legitimate diagnostic criteria for AHT/SBS. Legitimate medical responses that utilize scientific "controversy" terminology may inadvertently add to the confusion.\textsuperscript{122}

The Smith dissent makes it more likely that these increasingly common but unreliable medical opinions will be admitted in future child homicide and abuse cases. No appellate court has ruled that AHT/SBS fails to satisfy standards of evidentiary reliability for admission. But the effect of defense challenges at trial are more difficult to measure, because acquittals are the only outcome likely to be reported on, if at all, in the media. There has been no systematic research designed to assess the impact of defense testimony attacking AHT/SBS on the adjudication of child homicide or child abuse cases. But abundant anecdotal evidence suggests that when such evidence is presented by defense-retained medical witnesses, jurors and judges in child abuse and child homicide cases are more inclined to acquit defendants or to convict them of less serious offenses.

These concerns are more profound during postconviction proceedings. The legal standards for raising postconviction challenges (i.e., factual innocence, newly discovered evidence, and ineffective assistance of counsel claims) provide effective mechanisms for advancing specious but scientific-sounding claims. The standards for scientific evidence in postconviction proceedings are murky. Every newly published medical article or law review article (regardless of content or quality) that challenges the “orthodoxy,” proposes an alternative causation theory, or purports to have uncovered a “paradigm shift” provides the opportunity to argue that the defendant is factually innocent, that there is newly discovered evidence, or that trial counsel was ineffective for failing to challenge the AHT/SBS diagnosis. Whenever an appellate court relies on unsubstantiated scientific claims, or a misunderstanding of precedent, to find that the AHT/SBS “controversy” requires a new trial, this decision garners significant media attention. The media reports focus, not on the science, but on “groundbreaking” revelations of blameless parents and caretakers languishing in jail and their brave fight against the intransigent pro-prosecution medical mainstream.

Ironically, the Smith dissent cannot be attributed to the normal shortcomings of the AHT/SBS postconviction challenge. Smith’s habeas petition did not include legal affidavits from defense medical witnesses attacking the medical testimony at trial. In fact, the petition contained just one citation to the routinely cited defense that “research shows lucid intervals of up to seventy-two hours or more” but citing only to testimony from Dr. Robert Huntington, from the Edmunds case, that “[t]he lucid interval is a distinct discomforting but real possibility”); Findley et al., supra note 23, at 250–51 (claiming that “there is no real dispute over whether lucid intervals can occur”).\textsuperscript{122}

literature and did not include any postconviction medical testimony from defense witnesses proposing that “newly discovered evidence” established Smith’s actual innocence. The legal briefs also did not include any of the evidence or arguments that had been omitted from the habeas petition. Thus, Smith is especially troubling because the dissenters themselves went looking outside the record for evidence to support their conclusion that “[d]oubt has increased in the medical community ‘over whether infants can be fatally injured through shaking alone.”’

III. CAZAVOS V. SMITH: 1996–2012

A. Fifteen Years of Litigation, Three Decisions from the Ninth Circuit Granting Habeas Relief, and Three Reversals from the United States Supreme Court

Shirley Ree Smith was arrested in 1996 and charged with assault on a child resulting in death. She was tried the following year, convicted, and sentenced to fifteen years to life. Following exhaustion of her state appellate challenges, the defendant brought a writ of habeas corpus in federal court.

In her federal habeas petition, Smith claimed that her due process rights had been violated because the evidence introduced at trial was “constitutionally insufficient” particularly with respect to “one element of the crime—the cause of the child’s death.” On February 6, 2006, the Ninth Circuit reversed her conviction based on its finding that no rational jury could conclude beyond a reasonable doubt that Smith caused the child’s death, and the defendant was released from prison. This ruling was reversed and remanded by the United States Supreme Court on April 30, 2007, for further consideration in light of

---

124 Id. at 4 (majority opinion); see also supra text accompanying note 7 (explaining that the defendant was convicted under CAL. PENAL CODE § 273ab (West 2008) and quoting the relevant statutory language).
125 Id. at 5 (“The jury found Smith guilty. Concluding that the jury carefully weighed the tremendous amount of evidence supporting the verdict, the trial judge denied Smith’s motion for a new trial and sentenced her to an indeterminate term of 15 years to life in prison.” (citation omitted) (internal quotation marks omitted)).
126 Id. at 5–6 (“On direct review, Smith contended that the evidence was not sufficient to establish that Etzel died from SBS. After thoroughly reviewing the competing medical testimony, the California Court of Appeal rejected this claim . . . . The California Supreme Court [also] denied review.” (citation omitted)).
127 Id. at 6 (“Smith then filed this petition for a writ of habeas corpus with the United States District Court for the Central District of California, renewing her claim that the evidence was insufficient to prove that Etzel died of SBS.”).
129 Id. at 890. Thus, Smith served only a portion of her imposed sentence.
130 Patrick, 550 U.S. at 915.
Carey v. Musladin, in which the Court emphasized the deference owed to state court juries in federal petitions for postconviction review.

On December 4, 2007, the Ninth Circuit reinstated its 2006 decision based on a finding that “the opinion of the prosecution experts that shaking of the infant had caused death was wholly unsupported by the physical evidence.” On January 19, 2010, the Supreme Court again reversed and remanded to the Ninth Circuit for further consideration in light of McDaniel v. Brown, a new case clarifying the standard for habeas review.

On October 29, 2010, the Ninth Circuit reinstated its 2006 decision based on its finding, once again, that “nothing in the physical evidence supported the prosecution experts’ testimony as to the cause of death.” The Supreme Court reversed this decision for the third time on October 31, 2011, and on February 3, 2012, the Ninth Circuit affirmed the judgment of the district court denying Smith’s petition for a writ of habeas corpus.

On April 6, 2012, California Governor Jerry Brown commuted Smith’s sentence stating, “It is clear that significant doubts surround Ms. Smith’s conviction.”

B. Smith Illustrates Problems Endemic to Postconviction Review of Fact-Intensive Cases

An unfortunate reality of the appellate process is that pertinent facts from the trial record are often lost, ignored, distorted, taken out of context, or omitted as the case is presented in later hearings. In some cases, decreased judicial reliance on the facts may be appropriate, especially where the facts have little bearing on the legal analysis. However, child abuse cases are inherently fact based and Cavazos v. Smith provides a classic example of the problems that arise as multiple layers of appeals move the courts further from the relevant facts the jury considered.

At every level of appeal, the critical issue in Smith was the defendant’s claim that there was insufficient evidence of her guilt. This claim pertains to both the sufficiency of the medical evidence regarding the victim’s cause of death and the nonmedical evidence implicating the defendant’s acts as the cause of the victim’s

---

132 Id. at 77.
133 Smith v. Patrick, 508 F.3d 1256, 1258 (9th Cir. 2007), vacated, 558 U.S. 1143 (2010).
137 Smith, 132 S. Ct. at 2.
138 Smith v. Cavazos, 667 F.3d 1308, 1308 (9th Cir. 2012) (mem.). At this point Smith had been out of prison for nearly six years.
139 See supra note 22.
140 See supra Part III.A.
fatal injuries.\(^\text{141}\) The most thorough and complete recitation of the medical and nonmedical evidence appears in the February 10, 2000, decision from the California Court of Appeal.\(^\text{142}\) Thus, this discussion of the facts is principally derived from that decision. However, possible sources of confusion and additional or conflicting facts contained in subsequent decisions from the Ninth Circuit and the United States Supreme Court have also been identified and explained.\(^\text{143}\) What follows is a review of the nonmedical and medical evidence presented at trial. Where appropriate, the evidence presented by the prosecution and the defense and relied upon by the reviewing courts will be compared to the relevant evidence base in the widely available research literature.

\(^{141}\) See Smith v. Mitchell, 437 F.3d 884, 889–90 (9th Cir. 2006), vacated sub nom. Patrick v. Smith, 550 U.S. 915 (2007). The California Court of Appeals had already rejected the defendant’s challenge to the sufficiency of the evidence during her direct appeal. Id. at 888.

\(^{142}\) See People v. Smith, No. B118869 (Cal. Ct. App. Feb. 10, 2000). In addition to this unpublished decision, the authors of this Article have also reviewed the trial transcripts of each of the five medical experts who testified. This review revealed that, although the testimony of the prosecution experts encompassed nearly six-hundred pages of transcript, the appellate courts focused on the defense witnesses and summarized the prosecution witnesses’ testimony in just a few paragraphs. Reporter’s Transcript on Appeal, People v. Smith, No. B118869 (Cal. Ct. App. Feb. 10, 2000).

\(^{143}\) Repeated appeals also provide courts the opportunity to correct factual errors created when lower courts ignore or omit relevant facts. The Ninth Circuit had two opportunities to present an accurate assessment of the evidence presented at trial. Unfortunately, on both occasions, the court opted instead to microfocus on the narrow set of facts they claimed supported the conclusion that “no rational jury could find beyond a reasonable doubt, in light of all the evidence, that Smith had shaken the baby to death.” Smith v. Mitchell, 624 F.3d 1235, 1239 (9th Cir. 2010), rev’d sub nom. Cavazos v. Smith, 132 S. Ct. 2 (2011). The Ninth Circuit also made the implausible claim that, to reach this conclusion, it had “not resolve[d] any disputes of historical fact against the prosecution[,] [but] . . . simply assessed the prosecution’s evidence on its own terms and concluded that it did not meet the Jackson standard . . . .” Id. (citing Jackson v. Virginia, 443 U.S. 307, 319 (1979) (“[T]he relevant question is whether, after viewing the evidence in the light most favorable to the prosecution, any rational trier of fact could have found the essential elements of the crime beyond a reasonable doubt.”)). As such, the prosecution’s evidence “was so lacking that the state court’s rejection of Smith’s argument over the insufficiency of the evidence was an unreasonable application of Jackson to the facts of this case.” Id. As the Supreme Court correctly observed, in fact the Ninth Circuit simply ignored “the plenitude of expert testimony in the trial record concluding that sudden shearing or tearing of the brainstem was the cause of Etzel’s death.” Smith, 132 S. Ct. at 6. Thus, “[t]he Ninth Circuit’s assertion that [the prosecution’s] experts reached [their] conclusion because there is no evidence in the brain itself of the cause of death is simply false.” Id. at 7 (internal quotation marks omitted) (third alteration in original).
C. The Nonmedical Evidence Presented to the Smith Jury

1. Etzel’s Living Arrangements and Caregivers

Etzel Glass was born on October 10, 1996, and died on November 29, 1996. Etzel had been born two weeks before his due date with jaundice and a heart murmur, both of which had fully resolved before his death. No additional complications surrounding his birth were reported.

At the time of his death, Etzel lived with his grandmother, the defendant, Shirley Ree Smith. Etzel’s mother, Tomeka, his sister, Yolanda (four years old), and his brother Yondale (fourteen months old) along with several other family members also lived with Smith. Smith helped Tomeka care for her children and she was generally described as a loving grandmother who was not harsh or abusive to her grandchildren.

2. Etzel Appeared Normal During the Day and Evening Before He Died

On November 29, 1996, Smith took Etzel and his siblings to visit her sister Renee Townsend. Etzel seemed normal during the day. He was eating, smiling, moving his arms and legs, and having normal urination and bowel movements. That night, at approximately 11:30 p.m., after Etzel had been fed, changed, and bathed by his mother, Tomeka put him to sleep on his stomach on a sofa in the living room of Renee’s apartment. Renee left for work when Etzel was put down to sleep. Etzel’s brother, Yondale, slept on the same sofa with

\[
\text{Mitchell, 437 F.3d at 884, 885. The California appellate decision erroneously listed Etzel’s date of birth as November 10, 2006. See People v. Smith, No. B118869, slip op. at 2.}
\]

\[
\text{Smith, 132 S. Ct. at 4.}
\]

\[
\text{People v. Smith, No. B118869, slip op. at 2.}
\]

\[
\text{See id.}
\]

\[
\text{Id.}
\]

\[
\text{Mitchell, 437 F.3d at 885. The opinion does not indicate who provided this testimony at the trial. However, Etzel’s mother (Tomeka Smith) or Smith’s sister (Renee Townsend) must have provided it. As discussed below, see infra Part III.C.4–5, although omitted from the decisions by the state and federal appellate courts, this trial testimony is inconsistent with statements made by both witnesses to the social services and police investigators.}
\]

\[
\text{People v. Smith, No. B118869, slip op. at 2.}
\]

\[
\text{Id. at 2–3.}
\]

\[
\text{Id. at 3. Etzel’s ability to feed appropriately indicated he had not yet sustained any fatal injuries. The Ninth Circuit also described Tomeka’s testimony as indicating that Etzel appeared perfectly healthy during the day and at the beginning of the evening when both she and Smith fed him. Mitchell, 437 F.3d at 885–86.}
\]

\[
\text{People v. Smith, No. B118869, slip op. at 3.}
\]

\[
\text{Id.}
\]
Etzel while his sister, Yolanda, slept on a love seat in the same room. Smith slept on the floor next to the sofa. Renee’s children slept in the room they shared. Etzel was fine at 11:30 p.m. when Tomeka put him to sleep. Tomeka remained in the living room for about an hour. During this time, she checked his diaper and noticed him moving. Although Tomeka normally slept in the living room, that night she went into Renee’s bedroom to listen to music and fell asleep. Thus, after midnight Smith was Etzel’s only caregiver.

3. Etzel’s Condition While in the Defendant’s Care

At approximately 1:30 a.m., Smith awoke and stated she found Etzel on the floor. “She picked him up, rocked him back to sleep, and placed him on the couch in the same position (stomach down, head to the side).” There was no indication that Smith noticed anything unusual about Etzel at that time.

The suggestion that Etzel’s injuries could have been caused by a fall from the couch to the floor is clearly contradicted by the medical evidence, general well-accepted information about pediatric development, and common sense. As discussed below, none of the five medical experts who testified during the trial suggested that Etzel’s injuries and death could have been attributable to a short fall of less than two feet. The medical literature likewise does not support this conclusion. Moreover, because a seven-week-old child can barely scoot and cannot roll over, it is developmentally highly unlikely that Etzel accidentally fell off a couch. Given these facts, it is reasonable to conclude that Smith provided a false story of a fall to explain Etzel’s injuries. The child abuse literature is replete with this type of “accidental fall” from household items as a history offered to explain a range of severe injuries. In fact, these caretaker explanations are so common they are referred to in the professional literature as “the killer couch” story. In fact, research has confirmed that a purported fall of fewer than three

156 Id.
157 Id.
158 Id.
159 Id.
161 Id.
162 Id.
163 See id.
164 Id.
165 Id.
166 Id. According to the Ninth Circuit, Smith related this information to Tomeka. Id.
167 See, e.g., David L. Chadwick et al., Annual Risk of Death Resulting from Short Falls Among Young Children: Less than 1 in 1 Million, 121 PEDIATRICS 1213, 1213, 1220 (2008) (summarizing decades of research on short falls and noting the extreme rarity of such events).
168 See Brian K. Holmgren, Prosecuting the Shaken Infant Case, in THE SHAKEN BABY SYNDROME: A MULTIDISCIPLINARY APPROACH, supra note 9, at 275, 287 (noting
feet, when offered to account for severe head trauma, is one of several false histories that permit a diagnosis of AHT to be made with statistical reliability.\(^\text{169}\)

Smith said that she awoke again at 3:20 a.m. because she had to go to the bathroom.\(^\text{170}\) After she returned from the bathroom, she noticed that Etzel had thrown up and had blood on his right nostril.\(^\text{171}\) When Etzel did not respond to her touch, Smith “rushed into the bedroom holding Etzel and wakened Tomeka.”\(^\text{172}\) “Etzel was limp and appeared to have vomit coming from his nose.”\(^\text{173}\) Smith said, “Tomeka, Tomeka. Something is wrong with Etzel. . . . [C]all 911.”\(^\text{174}\) Smith then passed the baby to Tomeka, who then called 911.\(^\text{175}\)

4. Defendant’s Various Statements to Investigators Regarding Etzel’s Death

Firefighters and paramedics responded to the apartment at 3:36 a.m.\(^\text{176}\) When emergency personnel arrived, Smith was “apprehensive” and stated that she thought Etzel had fallen off the couch.\(^\text{177}\) Etzel was clothed, laying on the bed, and had bright red blood in one nostril.\(^\text{178}\) He was warm, but he was not breathing and had no heartbeat.\(^\text{179}\) The firefighters administered CPR without success.\(^\text{180}\) Etzel

\(^\text{169}\) See Joeli Hettler & David S. Greenes, Can the Initial History Predict Whether a Child with a Head Injury Has Been Abused?, 111 PEDIATRICS 602, 604–07 (2003); see also C. Henry Kempe et al., The Battered-Child Syndrome, 181 JAMA 17, 18 (1962) (“A marked discrepancy between clinical findings and historical data as supplied by the parents is a major diagnostic feature of the Battered Child Syndrome.”). Ironically, the Supreme Court has itself dealt with the “killer couch” scenario in one of the few other AHT cases it has decided. See Estelle v. McGuire, 502 U.S. 62, 67–75 (1991) (approving admission of evidence of earlier injuries to refute this claim). Notably, neither the Ninth Circuit nor the Supreme Court’s opinions in Smith reference this relevant precedent.

\(^\text{170}\) Mitchell, 437 F.3d at 886, vacated sub nom. Patrick, 550 U.S. at 915.

\(^\text{171}\) Id.


\(^\text{173}\) Id.

\(^\text{174}\) Id.

\(^\text{175}\) Id.

\(^\text{176}\) Id.


\(^\text{178}\) People v. Smith, No. B118869, slip op. at 3.

\(^\text{179}\) Id.

\(^\text{180}\) Id.
was transported to the hospital and arrived at 3:50 a.m. At that point, he was in full cardiac arrest, and was pronounced dead shortly thereafter.

Doctors initially suspected that Etzel had died as a result of Sudden Infant Death Syndrome (SIDS). After the autopsy was concluded, however, the cause of death was found to be SBS.

On December 5, 1996, a social worker, Linda Reusser, went to Renee’s apartment and told Tomeka and Smith that the cause of death had been changed to SBS. Ms. Reusser asked Smith what happened the night of Etzel’s death. Smith told her that Yondale had awakened at about 3:20 a.m. and that she had gone to comfort him. Smith said she also went over to Etzel and saw that he was face down. Smith said she became worried when she touched Etzel and he did not respond so she picked him up and gave him “a little shake, a jostle” to waken him. Smith said that Etzel’s head was flopped back when she picked him up. Smith then demonstrated how she had picked Etzel up by his armpits with her hands even with her shoulders and gave what was described as “a quick jostle” making a smooth rather than a jerky motion. At that point, Smith stopped talking. The social worker then asked Smith, “What happened next?” and Smith replied “Oh my God. Did I do it? Did I do it? Oh, my God.” Ms. Reusser also testified that Tomeka told Smith, “If it wasn’t for you, this wouldn’t have happened.” Finally, the social worker told Tomeka she was going to remove Yondale and Yolanda while the investigation continued.

Los Angeles police officers interviewed Smith four days later on December 9, 1996. Smith stated that Etzel had been fine at 11:30 p.m. on the night of November 29. She said that she was awakened when Yolanda fell on her.

181 Mitchell, 437 F.3d at 886.
182 Id.
183 People v. Smith, No. B118869, slip op. at 3.
184 Id. at 3–4. The medical basis for this diagnosis is discussed infra in Part III.D.
185 Id. at 4.
186 Id.
187 Id.
188 Id.
189 Id.
190 Id.
191 Id. The Ninth Circuit described this testimony slightly differently stating, “Smith demonstrated picking up the baby under his arms and moving him quickly forward and back in a smooth motion.” Smith v. Mitchell, 437 F.3d 884, 886 n.4 (9th Cir. 2006), vacated sub nom. Patrick v. Smith, 550 U.S. 915 (2007).
192 People v. Smith, No. B118869, slip op. at 4. Apparently Ms. Reusser testified to Smith’s demonstration, but it was not videotaped. See id.
193 Id.; see also Cavazos v. Smith, 132 S. Ct. 2, 4 (2011) (per curiam) (noting defendant’s statements to the social worker).
194 Smith, 132 S. Ct. at 11 n.3 (Ginsburg, J., dissenting).
196 Id.
197 Id.
Smith then noticed that Etzel needed to be changed and when she picked him up, his head flopped back and he had vomit around his mouth. At first, Smith told the police that she had shaken Etzel, but she corrected herself claiming instead that she had “twisted” him to try to elicit a response from him. Smith said that she shook Etzel for “just a matter of a few seconds.” Smith also told the police that Etzel had “fallen off the couch” earlier in the evening, was fine, and was laid back down to sleep on the sofa.

As discussed below, the medical evidence directly contradicts the defendant’s statements regarding the sequence of events. It supports the inference that events happened in the reverse order—that Etzel became unresponsive after he had been shaken.

5. The Trial Testimony

At trial, Tomeka denied that she had blamed Smith during the interview with Ms. Reusser. Tomeka also testified that the social worker fabricated both Smith’s admissions that she had shaken or jostled Etzel and Smith’s demonstration of how she shook Etzel to wake him. According to Tomeka, Ms. Reusser had accused Smith of killing Etzel and Smith had responded by saying, “No I didn’t.” Finally, Tomeka testified that she could not remember any of the statements she made to the police. Renee Townsend testified that Smith did not have a temper and that she only corrected the children verbally.

---

198 Id.
199 Id.
200 Id.
201 Id.; see also Cavazos v. Smith, 132 S. Ct. 2, 4 (2011) (per curiam) (noting defendant’s two contradictory statements to the police).
203 Smith v. Mitchell, 437 F.3d 884, 886 (9th Cir. 2006), vacated sub nom. Patrick v. Smith, 550 U.S. 915 (2007). Smith’s changing accounts of the events that evening were likely understood by the jury as additional nonmedical evidence tending to establish her guilt.
204 Id.
205 See infra Parts III.D.1, III.D.2, III.D.3 and accompanying text.
206 Smith, 132 S. Ct. at 11 n.3 (Ginsburg, J., dissenting).
207 Id. at 4 (majority opinion).
208 Id. at 11 (Ginsburg, J., dissenting).
209 Id. The jury likely understood the changing accounts of the facts from the defendant’s family members as additional nonmedical evidence tending to establish her guilt.
death.\textsuperscript{211} The prosecutor did not subject the defendant to substantive cross-examination.\textsuperscript{212}

6. Summary of the Nonmedical Evidence

Medical evidence is a dominant aspect of most child abuse and child homicide cases. However, nonmedical evidence is an important component of these cases and may frequently be dispositive in the minds of jurors, especially in cases that involve conflicting expert witness testimony.

In \textit{Smith}, the nonmedical evidence established that Etzel was neurologically normal in the hours leading up to the time when he was left in Smith’s exclusive care, indicating that he had not sustained any severe head trauma up to that point.\textsuperscript{213} This uncontradicted evidence provided a clear timeline of events establishing that Etzel’s neurological symptoms and death occurred only after he was in Smith’s care.\textsuperscript{214} The jury also viewed this evidence in conjunction with Smith’s own changing, conflicting, and implausible stories offered to account for Etzel’s fatal injuries. These include, most notably, Smith’s admission to the fatal mechanism of injury—shaking seven-week-old Etzel—although she unsurprisingly attempted to minimize the force associated with her shaking her grandson. Given that the jury also heard extensive medical evidence, discussed in the next section, establishing that Etzel died from AHT caused by violent shaking shortly after these injuries were inflicted, it should be clear that a rational trier of fact could have found the prosecutor proved the essential elements of the crime beyond a reasonable doubt.

\textbf{D. The Medical Evidence Presented to the Smith Jury}

At trial, the medical examiner who conducted the autopsy, Dr. Stephanie Erlich, her supervisor Dr. Eugene Carpenter, and Dr. David Chadwick,\textsuperscript{215} a pediatric specialist in child abuse, all testified that the cause of death was SBS that tore or sheared portions of Etzel’s brain stem, resulting in death.\textsuperscript{216} These conclusions were supported by the autopsy and neuropathology findings discussed at trial and by the extensive medical literature referenced at trial and cited below.

\footnotesize
\begin{itemize}
  \item \textsuperscript{211} Smith v. Mitchell, 437 F.3d 884, 885 n.1 (9th Cir. 2006) (“Smith herself testified very briefly . . . and her entire trial testimony extends for fewer than three pages of transcript.”), \textit{vacated sub nom.} Patrick v. Smith, 550 U.S. 915 (2007).
  \item \textsuperscript{212} \textit{Id.} (noting that Smith “was subjected to almost no cross-examination”).
  \item \textsuperscript{213} \textit{Id.} at 886.
  \item \textsuperscript{214} \textit{Id.}
  \item \textsuperscript{216} \textit{Smith}, 132 S. Ct. at 4–5.
\end{itemize}
I. Testimony from the Medical Examiner Who Conducted Etzel’s Autopsy—Dr. Erlich

(a) Dr. Erlich’s Qualifications and Findings

Dr. Erlich, who was board certified in anatomic pathology and neuropathology and was completing a one-year residency in forensic pathology, conducted both the autopsy and neuropathology exam. Her autopsy examination revealed that Etzel had subdural hemorrhages and subarachnoid hemorrhages. This bleeding was located diffusely: (1) between the two halves of his brain, (2) in the posterior fossa (i.e., at the base of his skull), and (3) on the under surface of his brain. Diffuse bleeding within the brain (i.e., bleeding located in multiple areas of the brain) is a hallmark of rotational injury to the head and is consistent with shaking, whereas focal hemorrhage (i.e., bleeding located in a single location) is characteristic of a focal impact trauma (i.e., injury resulting from a blow to the head or the head impacting against an object including injury potentially associated with a significant fall).

The autopsy also revealed older blood, which indicated an earlier brain injury that could have occurred days or weeks prior to his death. Acute hemorrhages were also found in both optic nerve sheaths, as well as older blood in this region, again suggesting an earlier injury. The subdural hematoma was described as being relatively small in size, measuring between one and two tablespoons in volume. Both the Ninth Circuit and the Supreme Court dissenter would subsequently attach significance to this seemingly small volume of subdural blood. However, this reflects the Courts’ ignorance of the fact that subdural fluid collections in AHT/SBS cases are typically small in volume. At trial, Dr. Erlich

217 Id.
218 People v. Smith, No. B116669, slip op. at 8.
219 Smith, 132 S. Ct. at 7; Mitchell, 437 F.3d at 887.
220 Reporter’s Transcript on Appeal, supra note 142, at 717–22 (testimony of Dr. Stephanie Erlich).
221 See Mary Case, Forensic Pathology of Childhood Brain Trauma, 18 BRAIN PATHOLOGY 562, 564 (2008); Mary Case, Inflicted Traumatic Brain Injury in Infants and Young Children, 18 BRAIN PATHOLOGY 571, 572 (2008) [hereinafter Case, Inflicted Traumatic Brain Injury].
223 Id.
226 See Jennian F. Geddes et al., Neuropathology of Inflicted Head Injury in Children: I. Patterns of Brain Damage, 124 BRAIN 1290, 1292 (2001) (noting that out of 53 cases of inflicted fatal head trauma only four of the older children had large enough hematomas to
also testified that there was no evidence of severe swelling of the brain at autopsy and that the brain had not herniated from swelling. This finding was consistent with Etzel experiencing a rapid death, before the brain had time to swell following trauma. Etzel’s brain, however, did evidence some swelling indicating that some period of time elapsed between the initial trauma to his brain and his death.

Dr. Erlich and another doctor also performed a neuropathological examination of Etzel’s brain. That examination “confirmed the presence of fresh subarachnoid hemorrhages in the parietal, the occipital areas, the frontal areas, and in the left temporal area” of the brain. “They also confirmed the presence of fresh subdural hemorrhage in the parietal area, as well as an older subdural hemorrhage in the right parietal area.” Based on these findings, Dr. Erlich confirmed that the cause of death was “trauma to the brain.”

The autopsy also revealed a “recent small abrasion, approximately 1/16 by 3/16 of an inch, on [Etzel’s] lower skull, upper neck region, and a recent bruise beneath this abrasion.” The Ninth Circuit inaccurately understated the diagnostic importance of Etzel’s scalp abrasion and underlying bruise when it opined that the “scalp abrasion was minimal, and was not even discovered until well into the autopsy.” This bruise and abrasion clearly indicated that Etzel had sustained a blunt impact trauma to the head. Impact trauma to the head is frequently difficult to detect on visual inspection and may not be seen until the scalp is reflected during the autopsy. Accordingly, this evidence can be easily missed in children who do not die and are not autopsied.

act as space occupying lesions while 34 had only a “trivial” quantity of subdural blood invariably described as “thin film”).

See Mitchell, 437 F.3d at 890.

Email from Dr. David Chadwick, Vice-Chairman, Int’l Advisory Bd. Nat’l Ctr. on Shaken Baby Syndrome, to Dr. Randell Alexander, Member, Int’l Advisory Bd. Nat’l Ctr. on Shaken Baby Syndrome (Sept. 22, 2012, 4:12 PM) (on file with authors) (“My notes indicate that the neuropathologist found slit-like ventricles and optic nerve hemorrhage. . . . The slit[-]-like ventricles require some time between injury and death.”); Dr. David Chadwick, Personal Case Notes for Testimony in California v. Smith (Dec. 2, 1997) (on file with authors) (“Neuropathologist noted slit-like lateral ventricles indicating some brain swelling.”). In fact, the presence of brain swelling is further evidence of direct trauma to the brain itself, since the brain’s response to trauma is to swell. See generally, ABUSIVE HEAD TRAUMA IN INFANTS AND CHILDREN: A MEDICAL, LEGAL, AND FORENSIC REFERENCE, supra note 9.


Id.

Id.

Id.; see also Reporter’s Transcript on Appeal, supra note 142, at 727–29 (testimony of Dr. Stephanie Erlich).


Id.

See, e.g., Randall Alexander et al., Incidence of Impact Trauma with Cranial Injuries Ascribed to Shaking, 144 AM. J. DISEASES CHILD. 724, 724–26 (1990); Duhaime et
Dr. Erlich testified that she did not observe retinal hemorrhages at autopsy. But the defense medical experts at Smith’s trial and the Ninth Circuit emphasized the absence of retinal hemorrhages as a basis for their conclusion that Etzel did not sustain a shaking injury. But this opinion ignores the extensive medical literature establishing that retinal hemorrhages are not present in approximately 20% of AHT/SBS cases. As Dr. Erlich explained during her trial testimony, the lack of retinal hemorrhaging did not undermine her diagnosis because it had been well documented that retinal hemorrhages are not present in 15–30% of SBS cases.

Dr. Erlich further explained that the fresh blood on Etzel’s brain could not be the result of rebleeding from an earlier trauma nor could it have been caused by shaking after death. Dr. Erlich specifically stated that the medical literature cannot support the theory of rebleeding of chronic subdural hematomas in children. Because minor trauma cannot cause a rebleed of a chronic subdural in infants, the older blood on Etzel’s brain represented prior trauma. This conclusion was further supported by the fact that the prior subdural was small,

236 Mitchell, 437 F.3d at 887.
237 Id. at 887–89. Ironically, most defense medical witnesses and legal commentators attempt to dismiss the significance of retinal hemorrhages when they are present, suggesting that they are caused by means other than shaking or are not diagnostic of trauma. See, e.g., Barnes, Imaging of Nonaccidental Injury, supra note 110, at 218; Leestma, supra note 113, at 599–600; Miller & Miller, supra note 109, at 169; Tuerkheimer, Science-Dependent Prosecution, supra note 23, at 516.

238 The absence of retinal hemorrhages does not preclude a diagnosis of AHT/SBS because, as numerous researchers have documented, retinal hemorrhages are not present in approximately 15–20% of AHT/SBS cases. See, e.g., Duhaime et al., supra note 91, at 410; Geddes et al., supra note 226, at 1294 (reporting presence in 71% of fatal cases); Craig E. Munger et al., Ocular and Associated Neuropathologic Observations in Suspected Whiplash Shaken Infant Syndrome, 14 AM. J. FORENSIC MED. & PATHOLOGY 193 (1993); Ralph S. Riffenburg & Lakshmanan Sathyavagiswaran, The Eyes of Child Abuse Victims: Autopsy Findings, 36 J. FORENSIC SCI. 741, 742 (1991) (reporting retinal hemorrhages in forty-seven of seventy-seven suspected child abuse deaths); Brandon Togioka et al., Retinal Hemorrhages and Shaken Baby Syndrome: An Evidence-Based Review, 37 J. EMERGENCY MED. 98, 100 (2009) (providing a systematic review of the medical literature on retinal hemorrhages and noting a 53–80% incidence in AHT); Dimitra Tzioumi & R. Kim Oates, Subdural Hematomas in Children Under 2 Years. Accidental or Inflicted? A 10-Year Experience, 22 CHILD ABUSE & NEGLECT 1105, 1108 (1998).

239 Reporter’s Transcript on Appeal, supra note 142, at 1276–78, (testimony of Dr. Stephanie Erlich).
240 See id. at 1298–99.
241 Id. at 1270–71.
242 Id. at 1303–05.
243 Id.
avascular, and there was no mass effect from the prior bleeding to support a rebleed claim.244

(b) Dr. Erlich’s Opinion Regarding Etzel’s Cause of Death

Based on these findings, both Dr. Erlich and the supervising medical examiner, Dr. Carpenter, concluded that Etzel died as a result of head trauma attributed to SBS.245 Dr. Erlich indicated that the mechanism of death was the tearing of brain tissue and long nerve fibers within Etzel’s brain and that bleeding over the brain’s surface was a marker for this type of brain injury and mechanism.246 Later in her testimony, she indicated there were three mechanisms by which SBS causes death: (1) mass effects from subdural bleeding put pressure on the brainstem, (2) swelling of the brain causes herniation, or (3) direct trauma to the brainstem.247 Dr. Erlich noted that she did not submit sections of Etzel’s brainstem for microscopic examination because injury would not be evident on microscopic examination if the child died quickly, and this would not have assisted in her diagnosis because the examiners “wouldn’t have seen anything anyway.”248

244 Id. at 1319-20. For further discussion of the rebleeding claim and the lack of scientific support for this frequent “alternative theory” proposed by the defense, see generally Barbara L. Knox et al., Subdural Hematoma Rebleeding, in ABUSIVE HEAD TRAUMA: POCKET ATLAS (Kay Rauth-Farley & L. Frasier eds.) (forthcoming 2013).
246 See Reporter’s Transcript on Appeal, supra note 142, at 730–31, 1324 (testimony of Dr. Stephanie Erlich).
247 Id. at 801–02. Specifically, trauma to areas of the brain that control heartbeat and respiration leads to a quick death that would not necessarily be recognizable through microscopic examination of the brainstem. Id. at 1297–99.
248 Smith v. Mitchell, 437 F.3d 884, 887 (9th Cir. 2006), vacated sub nom. Patrick v. Smith, 550 U.S. 915 (2007); see also Reporter’s Transcript on Appeal, supra note 142, at 803–05 (testimony of Dr. Stephanie Erlich) (noting that changes to the brainstem would not have been detectable unless the baby survived for a period of time). Dr. Erlich’s conclusion is confirmed by the medical literature. See S.M. Gentleman et al., Axonal Injury: A Universal Consequence of Fatal Closed Head Injury?, 89 ACTA NEUROPTHOLOGICA 537, 537, 541 (1995) (noting that the frequency of axonal damage had been vastly underestimated using conventional techniques and newer techniques using Beta Amyloid Protein Precursor (BAPP) were revealing much more prevalent axonal damage). Dr. Chadwick’s trial testimony described the advent of new staining techniques—although he did not mention the BAPP technique by name—noting their capacity to reveal this type of brainstem damage, but that they were not yet widely available. See infra notes 283–285 and accompanying text. Had BAPP staining been more widely available at the time of Etzel’s death, the likelihood of detection of axonal damage would have been substantially greater, although far from certain due to his rapid death. See, e.g., P. Shannon et al., Axonal Injury and the Neuropathology of Shaken Baby Syndrome, 95 ACTA NEUROPATHOLOGICA 625, 630 (1998) (indicating that at the time of this study, which used BAPP staining techniques to evaluate the brains of fatal abuse victims, there had been only a single series previously reported using silver stains—findings of this study identified axonal injury in the spinal cord suggesting that flexion-extension injury to the cervical spinal column may be caused
Dr. Erlich also analogized this situation to adult patients who suffer strokes and die right away, noting that doctors will know clinically that the person had a stroke, but would not see changes in the neurons of the brain immediately, although they might have seen such changes if the person had instead survived for several hours. Today, more advanced staining techniques that might have identified the precise locations of Etzel’s microscopic nerve tears are increasingly available to medical examiners and neuropathologists. However, at the time of his autopsy, Etzel’s brain evidenced swelling indicating that some period of time elapsed between the initial trauma to his brain and his death.

The Ninth Circuit and the Smith dissenters concluded that the absence of direct evidence of microscopic tears in Etzel’s brainstem demonstrated that there was no medical evidence of traumatic injury from shaking or impact trauma from which the jury could reliably determine guilt. On this critical point, the judges and justices were wrong—both legally and medically. Legally, both the expert witness testimony and the findings of blood markers for traumatic injury provided abundant evidence supporting the medical conclusions about the cause of Etzel’s death. Moreover, this testimony was entirely consistent with the available medical evidence and literature both at the time of the trial and today.

In contrast, the six justices who joined in the Smith per curiam opinion correctly recognized that “[t]he Ninth Circuit’s assertion that these [prosecution] experts ‘reached [their] conclusion because there was no evidence in the brain itself of the cause of death’ is simply false.” These justices understood that the injury findings and medical opinions were medically consistent with AHT/SBS because “[t]he autopsy revealed indications of recent trauma to Etzel’s brain, such as subdural and subarachnoid hemorrhaging, hemorrhaging around the optic nerves, and the presence of a blood clot between the brain’s hemispheres” and these “affirmative indications of trauma formed the basis of the experts’ opinion that Etzel died from shaking so severe that his brainstem tore.” The justices also understood why direct evidence of injury to the brain tissue was not presented, noting that “the experts explained why the location of the tear was undetectable: ‘Etzel’s death happened so quickly that the effects of the trauma did not have time


249 See Reporter’s Transcript on Appeal, supra note 142, at 766–67 (testimony of Dr. Stephanie Erlich). Similarly, people who die of heart attacks do not show evidence of this in their heart tissue if they die immediately because it takes time for damage to the heart tissue to be seen. Id. at 805.

250 See sources cited supra note 248.

251 See supra note 228 and accompanying text.


253 Id.
to develop.’ According to the prosecutions’ experts, there was simply no
opportunity for swelling to occur around the brainstem before Etzel died.”
Most importantly, they understood that the lack of evidence visualizing the location of
these microscopic tears did not undermine the accuracy of the AHT/SBS diagnosis.

2. Testimony from the Supervising Medical Examiner, Dr. Carpenter

(a) Dr. Carpenter’s Qualifications and Findings

Dr. Carpenter, a medical examiner board-certified in anatomic, clinical, and
forensic pathology, who had performed 3,000 to 4,000 autopsies, agreed with Dr.
Erlich that Etzel did not die from SIDS; he died from brain trauma. Dr.
Carpenter explained that if his death had resulted from SIDS, there would have
been no internal trauma, no bruises, and no abrasions.

(b) Dr. Carpenter’s Opinion Regarding Etzel’s Cause of Death

Dr. Carpenter testified that the bleeding on top of Etzel’s brain was caused by
shaking so that “there was a whiplash action of the head on top of the body with
the back of the head slamming into the back and the front of the head slamming
into the chest repeatedly so that the vessels on the top of the brain tore.”
According to Dr. Carpenter, Etzel’s “brain was so damaged from this violent
shaking that death occurred relatively quickly, within . . . 30 minutes . . . [and] there was no significant delay between the shaking and the brain damage.” As
Dr. Carpenter explained,

[T]here are two ways that . . . shaking to the head can cause death. One
way is that the shaking itself is so severe that the brain tears in vital areas
that control the heartbeat and the breathing. And then the bleeding can be
small and won’t amount to that much. The other way that a baby’s body
can die from shaking or from head trauma is for the vessels to tear, but
for the brain not to be damaged that much. And for the blood to
accumulate on top of the head. There is not enough room in the skull for

254 Id. (citation omitted).
255 Id. at 4–5.
SIDS diagnosis is precluded by Etzel’s significant internal injuries because under medical
guidelines promulgated prior to 1996, and still used today, SIDS can only be diagnosed in a
child who had a “negative autopsy” (i.e., an autopsy that had no other findings that could
explain cause of death). See, e.g., Comm. on Child Abuse & Neglect, Distinguishing
Sudden Infant Death Syndrome from Child Abuse Fatalities, 94 PEDIATRICS 124, 125
(1994) (listing sources and noting the requirement that the autopsy show no gross or
microscopic evidence of head injury or intracranial trauma).
258 Id.
the brain and the blood. So the brain is pressed downward into the spinal canal and is crushed and th[en] dies.\textsuperscript{259}

Dr. Carpenter explained that in Etzel’s case there was not enough accumulation of blood to press the brain downward and cause death from pressure on the brain, so Etzel’s death was caused by direct trauma to vital areas of the brain.\textsuperscript{260} Dr. Carpenter indicated that “he could not rule out the possibility of other blows to the head, but that there was no clear evidence of such blows except for the one small area of bruising” on Etzel’s lower skull.\textsuperscript{261}

Dr. Carpenter could not rule out the possibility that a fall from the sofa could account for one of Etzel’s injuries, the subdural hemorrhages.\textsuperscript{262} But he also testified that Etzel’s “injuries could not have been caused by the administration of CPR.”\textsuperscript{263} Dr. Carpenter stated that retinal hemorrhages and fractures are often seen in infants who are shaken and Etzel had neither injury,\textsuperscript{264} and that it is possible for subdural hemorrhages and subarachnoid hemorrhages to occur as a result of birth trauma.\textsuperscript{265} However, the medical research on subdural and subarachnoid hemorrhages indicates that birth-related hemorrhages do sometimes occur but that they could not account for Etzel’s injuries because they generally (1) are

\textsuperscript{259} Id. at 6–7. There are several additional ways for brain injury and death to occur. See R.A. Minns, Shaken Baby Syndrome: Theoretical and Evidential Controversies, 35 J. ROYAL C. PHYSICIANS EDINBURGH 5, 10 (2005) (documenting four mechanisms for fatal trauma including the mechanism described by the prosecution’s experts at Smith’s trial, and discussed infra note 334 and accompanying text). See generally Geddes et al., supra note 226 (studying fifty-three nonaccidental head injuries in children and finding skull fractures, acute subdural bleeding, retinal hemorrhages to be the most common injuries).

\textsuperscript{260} People v. Smith, No. B118869, slip op. at 7; see also supra notes 224–226, 246–247 and infra notes 329–331 and accompanying text (discussing Dr. Erlich’s testimony describing how AHT/SBS typically does not present itself with a large volume of subdural blood causing a mass effect on the brain, but instead serves as a “marker” for rotational injury to the brain involving shaking or shaking combined with impact).

\textsuperscript{261} People v. Smith, No. B118869, slip op. at 7.

\textsuperscript{262} Id. On cross-examination Dr. Carpenter acknowledged that one of Etzel’s injuries, subdural hemorrhages, could occur from a fall. Id. at 8. The medical literature indicates that such findings are exceedingly rare, however, and are typically focal in nature, confined to the location of the impact. See, e.g., Ann-Christine Duhaime et al., Nonaccidental Head Injury in Infants—The “Shaken-Baby Syndrome,” 338 NEW ENG. J. MED. 1822, 1822 (1998) (“[M]ost investigators agree that trivial forces, such as those involving routine play, infant swings, or falls from a low height are insufficient to cause [SBS].”). Etzel’s subdural hemorrhage and subarachnoid hemorrhage injuries are diffuse and indicative of a more global injury to the brain produced from rotational trauma seen with SBS.

\textsuperscript{263} People v. Smith, No. B118869, slip op. at 7.

\textsuperscript{264} Id at 7–8. Laura K. Brennan et al., Neck Injuries in Young Pediatric Homicide Victims, 3 J. NEUROSURGERY PEDIATRICS 232, 232 (2009) (finding that retinal hemorrhages are present in 80% of cases and 30–50% have skeletal injuries of various ages).

\textsuperscript{265} People v. Smith, No. B118869, slip op. at 8.
asymptomatic, (2) resolve after one month, (3) do not result in sudden deterioration and collapse, and (4) cannot result in diffuse acute bleeding or bleeding in a different location in the brain. 266

3. Testimony from the Child Abuse Expert Pediatrician—Dr. Chadwick

(a) Dr. Chadwick’s Qualifications and Findings

In addition to Drs. Erlich and Carpenter, Dr. David Chadwick, a pediatrician specializing in abuse, also testified for the prosecution. 267 Dr. Chadwick, who had been personally involved as a treating physician in fifty cases of SBS and AHT, half of which involved fatal outcomes, provided some of the most relevant and significant medical evidence. 268 Thus, it is odd that the reviewing courts have consistently ignored his testimony. Like Drs. Erlich and Carpenter (and defense-retained witness Dr. Siegler), Dr. Chadwick ruled out the possibility that Etzel died from SIDS because, based on Etzel’s other significant injuries, he did not have the requisite “negative autopsy” for a SIDS diagnosis. 269

(b) Dr. Chadwick’s Opinion Regarding Etzel’s Cause of Death

Dr. Chadwick explained the precise mechanism of Etzel’s death by describing how shaking infants creates acceleration and deceleration of the brain, which causes nerve fibers in the brain stem to tear. 270 When these nerve fibers tear they cease to function and the brain undergoes major changes, including the interruption

266 Birth-related subdural hemorrhages typically resolve by one month of age and are distinct from both the acute hemorrhages and the older subdural hematoma found in Etzel’s brain. The scattered pattern of acute subdural and subarachnoid bleeding throughout Etzel’s brain is also inconsistent with the theory that a chronic subdural from birth has rebled. See E.H. Whitby et al., Frequency and Natural History of Subdural Hemorrhages in Babies and Relation to Obstetric Factors, 362 LANCET 846, 847–50 (2004) (controlled research concluding that birth-related subdural hemorrhages are rare, benign, have clinically insignificant sequelae, and resolve by one month); see also Surya N. Gupta et al., Intracranial Hemorrhage in Term Newborns: Management and Outcomes, 40 PEDIATRIC NEUROLOGY 1, 11 (2009) (summarizing the typical treatment for intracranial hemorrhages and suggesting possible improvements); Knox et al., supra note 244 (summarizing research on rebleeding theory and concluding that rebleeds in children do not result in the spectrum of injuries and sudden collapse seen with AHT); Christopher B. Looney et al., Intracranial Hemorrhage in Asymptomatic Neonates: Prevalence on MR Images and Relationship to Obstetric and Neonatal Risk Factors, 242 RADIOLOGY 535, 538 (2007) (finding that seventeen out of eighty-eight asymptomatic neonates born via vaginal delivery had subdural hemorrhages).


268 Reporter’s Transcript on Appeal, supra note 142, at 1436–37 (testimony of Dr. Chadwick).

269 Id. at 1440–44.

270 Id. at 1448–49.
of vital functions controlling breathing and heartbeat, which leads to death. This injury mechanism was supported by evidence of the petechiae (spots caused by minor hemorrhaging) found in Etzel’s lungs which indicated a disruption of Etzel’s breathing mechanism and that he had been struggling to get air.

With respect to the timing of the injury, a critical question for the jury, Dr. Chadwick testified that Etzel could not have sustained this type of injury, remained asymptomatic for hours, and then suddenly collapsed and died. Dr. Chadwick supported this conclusion with both his extensive clinical experience and the medical literature documenting that children with severe head trauma immediately become unconscious or comatose, although brain death may occur hours or even days later. Dr. Chadwick indicated there were many cases in which children

271 Id. at 1448–49, 1476–77, 1480–82.
272 Id. at 1495.
273 Id. at 1460–61.
274 Id. at 1450–54. In AHT/SBS cases, defendants and their medical witnesses frequently assert that infants with severe or fatal head injuries experience a “lucid interval” of significant duration between the traumatic injury and when the child becomes severely symptomatic and unconscious. Medical research and clinical experience, however, do not support these claims but instead confirm the testimony offered by Dr. Chadwick—that children with severe and fatal head injuries do not experience significant periods of lucidity and are not asymptomatic. See Adamsbaum et al., supra note 49, at 550, 554; Mary E. Case, Head Injury in Child Abuse, in CHILD MALTREATMENT: A CLINICAL GUIDE AND REFERENCE, supra note 9, at 87, 95 [hereinafter Case, Head Injury in Child Abuse]; Mary E. Case et al., Position Paper on Fatal Abusive Head Injuries in Infants and Young Children, 22 AM. J. FORENSIC MED. & PATHOLOGY 112, 118–19 (2001); Duhaime et al., supra note 262, at 1825; Geddes et al., supra note 226, at 1303; M.G.F. Gilliland, Interval Duration Between Injury and Severe Symptoms in Nonaccidental Head Trauma in Infants and Young Children, 43 J. FORENSIC SCI. 723, 724 (1998); Tom. D. Lyon et al., Medical Evidence of Physical Abuse in Infants and Young Children, 28 PAC. L.J. 93, 146, 163 (1996); Starling et al., Analysis of Perpetrator Admissions, supra note 49, at, 456–57; Starling et al., Abusive Head Trauma, supra note 49, at 261; Krista Y. Willman et al., Restricting the Time of Injury in Fatal Inflicted Head Injuries, 21 CHILD ABUSE & NEGLECT 929, 938–39 (1997). Both Dr. Starling and Dr. Adamsbaum’s research confirms that perpetrators who confess to shaking and other forms of AHT consistently report that children are immediately symptomatic. Adamsbaum, supra note 49, at 551; Starling et al., Analysis of Perpetrator Admissions, supra note 49, at 456. In fact, Dr. Gilliland’s research found that in all cases where there was a person (other than the suspect) present at the time of injury, the child was immediately symptomatic. Gilliland, supra, at 724. But see Kristy B. Arbogast et al., Initial Neurologic Presentation in Young Children Sustaining Inflicted and Unintentional Fatal Head Injuries, 116 PEDIATRICS 180, 180–84 (2005) (pointing out that the Willman study involved only two infants and arguing that lucidity does not imply that children are asymptomatic because the Glasgow Coma Scale does not measure common symptoms such as vomiting, irritability, or subtle changes in alertness which means that children with less severe injuries from AHT may exhibit symptoms that do not precipitate a medical crisis and consequently may not be brought in for medical treatment and their head trauma may be undiagnosed); Antoinette L. Laskey et al., Occult Head Trauma in Young Suspected Victims of Physical Abuse, 144 PEDIATRICS 719 (2004).
sustaining these types of injuries were dead on arrival at the hospital. Based on this evidence, he opined that Etzel’s death had occurred very rapidly. Dr. Chadwick testified that if the injury is sufficient to cause the infant to stop breathing because it affects the vital centers, then the baby will die in five to ten minutes.

Dr. Chadwick testified that the blood on Etzel’s brain provided additional evidence of a traumatic injury. He also noted that his diagnosis was not undermined by the absence of retinal hemorrhages because these findings are not present in approximately 20% of AHT/SBS cases. Dr. Chadwick testified that Etzel’s optic nerve hemorrhages were consistent with SBS and the older hemorrhage in the optic nerve likely indicated a prior occasion of similar injury. Dr. Chadwick also ruled out rebleeding from the older subdural hematoma as a cause for the acute blood in Etzel’s head and testified that a short fall from a sofa could not explain Etzel’s fatal injuries.

Dr. Chadwick explained that tears in an infant brain’s nerve fibers are microscopic and difficult to locate during autopsy. But he strongly rejected the suggestion that because doctors could not locate these microscopic tears, they could not make a definitive cause of death determination. In his view, this same flawed logic should be compared to the (equally unsupportable) assumption that doctors cannot accurately diagnose cause of death for infants who die of overheating in cars because they do not present anatomic or physiological evidence that is specific to (can only be explained by) overheating. Finally, Dr. Chadwick clarified that Etzel’s lack of neck injuries did not undermine his diagnosis because neck injuries are identified in only 10% of AHT/SBS cases.

275 Reporter’s Transcript on Appeal, supra note 142, at 1480 (testimony of Dr. Chadwick).
276 Id.
277 Id. In other circumstances, however, death might occur hours or days later as a result of brain swelling. Id. at 1477.
278 Id. at 1446–49, 1466, 1474 (explaining that this blood is not what hurts or kills the baby, rather it is the damage to the brain itself and the blood is an indicator of this injury); see also id. at 1467–68 (describing how the older bleeding on Etzel’s brain indicated an earlier less severe injury of a similar nature).
279 Id. at 1463–65, 1490.
280 Id. at 1467–68.
281 Id. at 1446–47.
282 Id. at 1455–59 (describing his own research and publications in this arena and that of others). Since this trial testimony, Dr. Chadwick has published a meta-analysis of the short-fall research. See Chadwick et al., supra note 167.
283 Reporter’s Transcript on Appeal, supra note 142, at 1448–49 (testimony of Dr. Chadwick) (describing new techniques to identify these injuries not used in this autopsy and not widely available at the time).
284 Id. at 1483–84
285 Id.
286 Id. at 1491–93.
4. Testimony from Defense-Retained Witness—Dr. Siegler

(a) Dr. Siegler’s Qualifications and Findings

The defense called two experts at trial. The first expert, Dr. Richard Siegler, a forensic pathologist, agreed that Etzel did not die from SIDS but died from brain trauma. The California Court of Appeal described Dr. Siegler’s qualifications as follows:

On voir dire, the prosecutor established that Dr. Seigler’s extensive experience was not in forensic pathology, but in general, clinical and anatomic pathology. Ninety-five percent of his publications had been on cancer. He had not authored any articles dealing with child abuse or forensic pathology in determining the causes of death in infants. Although he had been a medical examiner in Philadelphia in 1960, shaken baby syndrome was unknown at that time. He had not been involved in any autopsies where the cause of death was found to be shaken baby syndrome. He had performed 50 autopsies on children with brain trauma over the last 25 years.

(b) Dr. Siegler’s Opinion Regarding Etzel’s Cause of Death

In Dr. Siegler’s view, Etzel’s death was not caused by shaking because there were no retinal hemorrhages. As discussed above, Dr. Siegler’s testimony is not supported by the medical evidence introduced at trial or the medical literature. The absence of retinal hemorrhages does not preclude a diagnosis of shaking or AHT because, as numerous researchers have documented, retinal hemorrhages are not present in approximately 20% of AHT/SBS cases. Moreover, as Dr. Siegler subsequently acknowledged from the witness stand, the absence of retinal hemorrhages cannot exclude a diagnosis of SBS.

Dr. Siegler testified that Etzel’s brain trauma did not happen shortly before his death because he did not see evidence of a fresh trauma. The Ninth Circuit inexplicably misstated Dr. Siegler’s testimony on this vital point. According to the appellate court, Dr. Siegler opined that Etzel must have died from the lingering effects of earlier brain trauma of unknown but quite possibly innocent cause.

---

287 Id. at 1469.
288 Id.
290 Id.
291 Id.
292 See sources cited supra note 238.
293 People v. Smith, No. B118869, slip op. at 9.
294 Id. at 9–10.
This is not an accurate description of Dr. Siegler’s testimony. A review of the transcript of Dr. Siegler’s testimony reveals that his opinions pertaining to “older trauma” were based on the presence of the older subdural hemorrhages, which he (mistakenly) described as broadly present (i.e., diffuse) in Etzel’s brain.296 As noted above, the older subdural hemorrhages found in Etzel’s brain were not diffuse, but were instead focal (confined to one area of the brain). It is clear from the trial transcript that Dr. Siegler could not account for the diffuse acute (fresh) hemorrhages described in the neuropathology report and diagrammed by Dr. Erlich at trial. Thus, the Ninth Circuit’s speculation that Etzel likely died from older brain trauma of “possibly innocent” origin is unsupported by Dr. Siegler’s testimony or any other medical evidence presented at trial, or any medical literature.297

Finally, although both the Ninth Circuit and the Smith dissenters ignored this fact, Dr. Siegler conceded that his medical opinions were based solely on photographs taken by the neuropathologist.298 He also admitted that he had not even considered the evidence presented by Dr. Erlich and Dr. Carpenter—which showed extensive fresh bleeding to Etzel’s brain—in reaching his conclusion that there was no medical evidence of fresh trauma.299

5. Testimony from the Defense-Retained Witness—Dr. Goldie

(a) Dr. Goldie’s Qualifications and Findings

The defendant’s second expert, Dr. William Goldie, a pediatric neurologist, disagreed with all of the other experts and testified that Etzel’s death was due to SIDS.300 According to Dr. Goldie, jaundice, a heart murmur, and low birth weight were all factors that made Etzel predisposed to a SIDS death.301 Without support from the medical evidence or the medical literature and in contrast to the other

296 See id.
297 See supra notes 219–232, 240–244 and accompanying text.
300 Id. at 10.
301 Id. According to the Ninth Circuit, Dr. Goldie described some of the characteristics that led him to conclude that Etzel died of SIDS. Mitchell, 437 F.3d at 888, vacated sub nom. Patrick, 550 U.S. at 915. Dr. Goldie specified, “With SIDS, the infant usually would appear normal, but then he or she suddenly would die.” Id. Furthermore, Dr. Goldie stated, “SIDS occurred more frequently in babies who, like Etzel, were small for their age, who had mothers who had multiple children already or smoked or used drugs, and, most importantly, who had been placed face-down on their stomachs.” Id. Finally, Dr. Goldie indicated, “Males were more likely victims than females.” Id. But the medical research on SIDS diagnosis does not support Dr. Goldie’s conclusion that these factors, individually or in combination, support a finding that Etzel died of SIDS. See Comm. on Child Abuse & Neglect, supra note 256 (explaining that SIDS can only be diagnosed in a child who had a “negative autopsy” and therefore a SIDS diagnosis was precluded by Etzel’s significant internal injuries (i.e., the presence of old and acute subdural hemorrhage, old and acute optic nerve hemorrhage, and abrasion trauma to the neck)).
experts, Dr. Goldie opined that some of the acute bleeding on Etzel’s brain was the result of CPR\textsuperscript{302} and that, in the alternative, premature infants sometimes bleed into their heads without cause.\textsuperscript{303}

(b) Dr. Goldie’s Opinion Regarding Etzel’s Cause of Death

Dr. Goldie, who was not board certified in neuropathology, clinical pathology or forensic pathology, concluded that the pathologists had inaccurately determined the cause of Etzel’s death\textsuperscript{304} because, in his opinion, SBS can only be diagnosed based on a finding of either massive brain bleeding or massive brain swelling (at least when the brain stem does not show damage).\textsuperscript{305}

The medical evidence presented at trial and the medical literature all contradict Dr. Goldie’s opinion. First, massive brain bleeding is not a valid diagnostic criterion for AHT/SBS. The volume of blood in these cases is typically small, one to two tablespoons, just as in Etzel’s case.\textsuperscript{306} Second, Dr. Goldie’s testimony on brain swelling ignores the medical evidence indicating that Etzel died from injury to his brain stem that disrupted the respiratory center of his brain and the fact that because the brain stops swelling when a child dies, the lack of swelling is medically consistent with brain death occurring shortly after injury.\textsuperscript{307}

E. The Smith Dissenters Interpret the Evidence

1. The Dissenters’ Misguided Concern that Etzel’s Injury Could not be Located

According to Justice Ginsburg, “[w]hat is now known about SBS casts grave doubt on the charge leveled against Smith.”\textsuperscript{308} In addition to this inaccurate general assessment of the current state of medical knowledge, the dissenters reviewed the specific medical evidence presented by the five experts who testified at the Smith trial and inaccurately concluded that “[f]ew of the[] signs of SBS were present here.”\textsuperscript{309} This assumption echoes the Ninth Circuit’s mistaken finding that the jury based its verdict on the opinions of medical experts who had improperly “reached [their] conclusion because there was no evidence in the brain itself of the cause of death.”\textsuperscript{310}

\textsuperscript{302} People v. Smith, No. B118869, slip op. at 10. There is no medical evidence to support Dr. Goldie’s assertion that CPR can cause subdural or subarachnoid hemorrhage. See Evan W. Matshes, Emma O. Lew, Do Resuscitation-Related Injuries Kill Infants and Children?, 31 AM J FORENSIC MED. & PATHOLOGY 178 (2010).

\textsuperscript{303} Mitchell, 437 F.3d at 888.

\textsuperscript{304} People v. Smith, No. B118869, slip op. at 10.

\textsuperscript{305} Id.

\textsuperscript{306} See supra notes 224–226 and accompanying text.

\textsuperscript{307} See supra notes 227–228, 247–251, 258–259, 270–275 and accompanying text.


\textsuperscript{309} Id.

\textsuperscript{310} Id. at 7 (majority opinion).
The Smith dissenters begin with the complaint that the “autopsy revealed no physical evidence of such injury, either grossly or microscopically,” that “Dr. Carpenter was unable to state which particular areas of the brain were injured, and the neuropathologist found no evidence of specific brain injury,” that “[n]o doctor located any tear,” and that “the examining physicians did not cut open Etzel’s brainstem, or submit it to neuropathology.”

Justice Ginsburg’s concerns illustrate how courts can adopt and then advance a profound misunderstanding of the medical evidence presented at trial in AHT/SBS cases. First, there was no gross physical evidence of the tears to Etzel’s brain because they were too small to see and occurred at a microscopic level. Second, the pathologists who conducted Etzel’s autopsy did not section his brain for microscopic examination because this procedure would not have helped them “locate[] any tear.” As Dr. Erlich stated at trial, dissection was not necessary because, given the techniques available at the time, they “wouldn’t have seen anything anyway.” In fact, Dr. Chadwick described the advent of new staining techniques with the capacity to reveal this type of brainstem damage, but noted that at the time of Etzel’s autopsy they were not yet widely available. Etzel did not survive long enough for changes to his brain to become detectable using standard techniques, despite the fact that the bleeding around his brain clearly indicated injury to the brain itself.

311 Id. at 9 (Ginsburg, J., dissenting) (citations omitted).
312 Id. But see supra notes 245–251 and infra notes 329–332 and accompanying text.
313 Id. But see supra notes 245–251 and infra notes 329–332 and accompanying text.
314 See Smith v. Mitchell, 437 F.3d 884, 887 (9th Cir. 2006), vacated sub nom. Patrick v. Smith, 550 U.S. 915 (2007). This opinion is confirmed by the fact that in 1996 the staining techniques generally available to detect such injury—hematoxylin-eosin (H&E) stains—required the victim to survive for nearly fifteen hours. See Gentleman et al., supra note 248, at 537 (noting that the frequency of axonal damage had been vastly underestimated using conventional silver techniques and newer techniques using BAPP were revealing much more prevalent axonal damage).
315 See sources cited supra note 248 (discussing new techniques). Although Dr. Chadwick did not mention the BAPP technique by name, had BAPP staining been more widely available at the time of Etzel’s death, the likelihood of detection of axonal damage would have been substantially greater, although far from certain due to his rapid death. But see Manfred Oehmichen et al., Shaken Baby Syndrome: Re-examination of Diffuse Axonal Injury as Cause of Death, 116 ACTA NEUROPATHOLOGICA 317, 317–29 (2008) (not detecting this pattern of injury and ascribing multiple causes for the disruption of breathing mechanisms resulting from shaking). For an excellent overview of these issues, see Lucy B. Rorke-Adams, Neuropathology of Abusive Head Trauma, in CHILD ABUSE AND NEGLECT: DIAGNOSIS, TREATMENT, AND EVIDENCE, supra note 9, at 413; see also J.F. Geddes et al., Neuropathology of Inflicted Head Injury in Children: II. Microscopic Brain Injury in Infants, 124 BRAIN 1299, 1299–306 (2001).
316 Reporter’s Transcript on Appeal, supra note 142, at 1449 (testimony of Dr. David L. Chadwick).
317 See supra notes 219–221, 246–251, 278 and accompanying text.
Although the Smith dissenters appear confused on this point, the fact that the doctors who performed Etzel’s autopsy in 1996 could not visualize the microscopic tears that caused Etzel to stop breathing does not mean, as the Ninth Circuit asserted, that “there was no evidence . . . of the cause of death.” This argument is both illogical and unscientific. It is akin to asserting that the advent of a new medical technology enabling physicians to more accurately pinpoint the location of a microscopic injury undermines the validity of all previous diagnoses. Physicians have a long and widely known history of accurately diagnosing a range of diseases that only recently can be visualized using new imaging and diagnostic technologies. The argument also fails as a matter of basic science. Diseases can be diagnosed, and even cured, without a precise understanding of the entire disease process. To cite just one famous example, scurvy was first described by Hippocrates. By the mid-eighteenth century, the disease was successfully treated with fresh food (particularly citrus fruit). The discovery of vitamin C deficiency as the precise biological disease mechanism did not occur until 1932.

2. The Dissenters’ Five Specific Concerns Regarding the Medical Evidence

In addition to their specific concern that Etzel’s brain tears could not be visualized, the Smith dissenters list five general problems with the medical evidence that, in their view, “cast grave doubt” on the accuracy of the jury verdict. Because these general concerns will likely be repeated by defense-retained medical witnesses, counsel, and judges in future AHT/SBS cases, each will be evaluated below.

(a) “Etzel’s Subdural Hemorrhage and Subarachnoid Hemorrhage Were Minimal”

This statement, although technically accurate, cannot undermine the validity of the AHT/SBS diagnosis and reflects a lay misunderstanding of the mechanism of AHT injuries. The autopsy revealed that Etzel had subdural hemorrhages and subarachnoid hemorrhages located between the two halves of his brain in the posterior fossa (i.e., at the base of his skull) and on the under surface of his brain. The autopsy also revealed older blood, which indicated an earlier brain injury that could have occurred days or weeks prior to his death.
hemorrhages were also found in both optic nerve sheaths, as well as older blood in this region, again suggesting an earlier injury.327

Etzel’s subdural and subarachnoid hemorrhages were described as being small in size, measuring between one and two tablespoons in volume.328 But as Dr. Erlich explained, subdural bleeding in AHT cases need not manifest as large, mass effect collections of blood that put pressure on the brain and lead to death.329 Instead, the presence of such blood acts as a “marker” for a specific mechanism of traumatic brain injury—the rotational acceleration-deceleration of the head and brain.330 This injury mechanism tears bridging veins causing the bleed and concurrently causing direct injury to the brain tissue but does not result in significant bleeding.331 This is consistent with the extensive medical literature and Dr. Carpenter’s testimony explaining that Etzel’s death was caused by “direct trauma to vital areas of the brain.”332

(b) “There Was No Brain Swelling”333

This statement reflects a fundamental misunderstanding of how shaking can cause infant death. At trial, Dr. Carpenter specifically testified that shaking can cause death in different ways.334 First, shaking can create a large bleed that creates

327 Id.
328 Mitchell, 437 F.3d at 887.
330 See Case, Inflicted Traumatic Brain Injury, supra note 221, at 576.
331 See, e.g., id. at 573 (noting that subdurals may involve less than 5–10 mL of blood); Geddes et al., supra note 226, at 1291–94 (noting out of fifty-three cases of inflicted fatal head trauma only four of the older children had large enough hematomas to act as space occupying lesions while thirty-four had only a “trivial” quantity of subdural blood invariably described as “thin film”).
332 People v. Smith, No. B118869, slip op. at 7. It is also consistent with Dr. Chadwick’s testimony. See supra notes 270–272, 278 and accompanying text.
334 People v. Smith, No. B118869, slip op. at 7. The doctors at Etzel’s trial primarily discussed the two mechanisms as listed above. Another mechanism, which was not described, involves a primary injury to the brain (e.g., torn or damaged brain tissue) causing the brain to swell in response to this injury. A second mechanism is via hypoxic-ischemic injury to the brain. In these cases, trauma causes a disruption in breathing that produces a lack of oxygen and hypoxic insult. The medical evidence presented by all three prosecution experts indicated that Etzel died quickly as a result of this second injury mechanism. But this hypoxic insult can also occur over a protracted time frame with hypoxia being caused by ongoing swelling. Other medical literature describes additional mechanisms of injury and ongoing research continues to elucidate additional primary and secondary injury mechanisms. See Minns, supra note 259, at 11–12 (describing four principle patterns of presentation). Ongoing research into the brain’s biochemical responses to traumatic injury is providing additional information regarding injury pathways. See Rachel P. Berger & Noel Zuckerbraun, Biochemical Markers, in Mild Traumatic Brain
pressure on the brain causing the brain to herniate. But another way is that “the shaking is so severe that the brain tears in vital areas that control the heartbeat and the breathing.”³³⁵ As noted above, Dr. Carpenter, Dr. Erlich, and Dr. Chadwick all agreed that the mechanism of Etzel’s death was the tearing of brain tissue and long nerve fibers within his brain and that bleeding over the brain’s surface was a marker for this type of brain injury.³³⁶ Thus, Etzel died from injury to the brain stem that disrupted the respiratory center of his brain. Because extensive medical evidence supports the finding that the brain stops swelling when the child dies, the lack of brain swelling is medically consistent with brain death occurring shortly after injury³³⁷ and Dr. Chadwick testified that Etzel’s death had occurred very rapidly.³³⁸ According to Dr. Chadwick, if the injury is sufficient to cause the infant to stop breathing because it affects the vital centers, the baby can die in five to ten minutes.³³⁹

(c) “There Was . . . No Retinal Hemorrhage in Either Eye”³⁴⁰

This concern is likely derived from Dr. Siegler’s testimony that Etzel’s death was not caused by shaking because there were no retinal hemorrhages.³⁴¹ The lack of retinal hemorrhages in one of every five AHT/SBS cases itself establishes that this finding cannot undermine the abuse diagnosis.³⁴² Because this fact was well known at the time, presented at trial, and even conceded by Dr. Siegler,³⁴³ it is difficult to understand why both the Ninth Circuit and the Smith dissenters ignored it.

³³⁵ People v. Smith, No. B118869, slip op. at 6–7. Brain swelling takes time to occur and does not continue after death, so if the child dies quickly following trauma then the swelling would be minimal and would not be the direct cause of brain herniation or death. Dr. Carpenter’s opinions are confirmed in subsequent research, including two articles published in 2001 by Dr. Jennian Geddes describing the pathophysiology of brain trauma and concluding that shaking trauma can disrupt respiratory centers in the brain stem and cause hypoxic injury to babies which is precisely the cause of death described by the prosecution experts here. See Geddes et al., supra note 226; Geddes et al., supra note 315; Rorke-Adams, supra note 315.

³³⁶ See supra notes 245–251, 259, 270–272 and accompanying text.


³³⁸ Reporter’s Transcript on Appeal, supra note 142, at 1480 (testimony of Dr. Chadwick).

³³⁹ Id. In other circumstances, however, death might occur hours or days later as a result of brain swelling. Id. at 1477.


³⁴¹ See id.

³⁴² See supra note 238 and accompanying text.

³⁴³ Smith, 132 S. Ct. 2 at 5.
(d) “[A]bsent Were Any Fractures, Sprains, Bleeding in the Joints, or Displacement of the Joints”\textsuperscript{344}

This statement cannot undermine the validity of any AHT/SBS diagnosis. None of these different types of injuries are diagnostic criteria for AHT/SBS. The medical evidence further reveals that it is not uncommon for infants suffering from AHT/SBS to present without other musculoskeletal injuries.\textsuperscript{345}

(e) “A ‘Tiny’ Abrasion on the Skin and a Corresponding Bruise Under the Scalp Did Not Produce Brain Trauma”\textsuperscript{346}

This concern is likely derived from Ninth Circuit dicta opining that Etzel’s “scalp abrasion was minimal and was not even discovered until well into the autopsy”\textsuperscript{347} and, in their view, insufficient to have caused his death. This comment inaccurately understates the diagnostic importance of Etzel’s scalp abrasion and underlying bruise. A child of Etzel’s age should not have any scalp injuries. Any abrasions or contusions (both were present here) are significant findings because they provide evidence indicating that the child was not merely shaken but also impacted.\textsuperscript{348} Any evidence of impact injuries is especially important in AHT/SBS cases because medical research shows that impact magnifies the forces to the infant brain anywhere from ten to fifty fold, making the injury that much greater.\textsuperscript{349}

The medical evidence contradicts the \textit{Smith} dissenters’ suggestion that small abrasions and contusions cannot contribute to brain trauma. In this case, Dr. Carpenter specifically ruled out the possibility that a fall from the sofa could account for the traumatic injuries to Etzel’s head.\textsuperscript{350} He also testified that Etzel’s injuries could not have been caused by the administration of CPR.\textsuperscript{351} On this point,
the Ninth Circuit’s concern that these findings were “discovered well into the autopsy” is equally unpersuasive. Impact trauma to the head is frequently difficult to detect on external visual inspection. In fact, this evidence can be easily missed in children who do not die, because these lesions are best visualized during autopsy because the scalp is revealed.352

Thus, there is no evidence-based reason for the Smith dissenters to have concluded that all or any of these five medical questions cast doubt on the accuracy of the jury verdict.

IV. CONCLUSION: HOW THE “TAIL WAGS THE DOG” ON POST-CONVICTION REVIEW

A. The Legal Standard

1. The Smith Per Curiam Opinion and the Legal Standard

The legal standard for appellate review of state court convictions is abundantly clear. It is, of course, the responsibility of the jury and not the appellate court to draw conclusions based on the evidence presented at trial.353 A reviewing court “may set aside the jury’s verdict on the ground of insufficient evidence only if no rational trier of fact could have agreed with the jury.”354 Moreover, when reviewing a decision from a state court rejecting a sufficiency of the evidence challenge, the federal court may only overturn if the state court decision was “objectively unreasonable.”355 Under the review standard from the Antiterrorism and Effective Death Penalty Act of 1996,

a writ of habeas corpus on behalf of a person in custody pursuant to the judgment of a State court shall not be granted with respect to any claim that was adjudicated on the merits in State court proceedings unless the adjudication of the claim . . . 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.356

The Ninth Circuit had no power to afford habeas relief unless Smith could demonstrate that the California Court of Appeals decision affirming her conviction

352 See, e.g., Duhaime et al., supra note 91, at 411; Gill et al., supra note 235, at 621.
353 Cavazos v. Smith, 132 S. Ct. 2, 3–4 (2011) (per curiam) (“The opinion of the Court in Jackson v. Virginia, 443 U.S. 307 (1979), makes clear that it is the responsibility of the jury—not the court—to decide what conclusions should be drawn from evidence admitted at trial.”).
354 Id. at 4.
355 Id. (quoting Renico v. Lett, 130 S. Ct. 1855, 1862 (2010)).
“was contrary to, or involved an unreasonable application of,’ clearly established federal law as reflected in the holdings of th[e] [Supreme] Court’s cases.”

The six justices who joined the Smith per curiam decision correctly held that in granting habeas relief, the Ninth Circuit improperly ignored “the plentitude of expert testimony in the trial record concluding that sudden shearing or tearing of the brainstem was the cause of Etzel’s death.”

The Ninth Circuit’s assertion that these [prosecution] experts “reached [their] conclusion because there is no evidence in the brain itself of the cause of death” is simply false. There was “evidence in the brain itself.” The autopsy revealed indications of recent trauma to Etzel’s brain, such as subdural and subarachnoid hemorrhaging, hemorrhaging around the optic nerves, and the presence of a blood clot between the brain hemispheres. The autopsy also revealed a bruise and abrasion on the lower back of Etzel’s head. These affirmative indications of trauma formed the basis of the experts’ opinion that Etzel died from shaking so severe that his brainstem tore.

Thus, Smith’s conviction was not objectively unreasonable.

2. The Smith Dissent and the Legal Standard

Justice Ginsburg, writing for the three dissenters, did not focus on the governing legal standard when she opined that the Supreme Court erroneously granted California’s petition for review. As shown above, she also did not focus on the extensive medical and nonmedical evidence presented at trial.

Thus, Smith’s conviction was not objectively unreasonable.

357 Smith, 132 S. Ct. at 6 (quoting 28 U.S.C. § 2254(d)(1)).
358 Id. at 5–6.
359 Id. at 7.
360 See id.
361 Id. at 8 (Ginsburg, J., dissenting) (quoting Smith v. Mitchell, No. CV 01-4484-ABC, slip op. at 10 (C.D. Cal. Mar. 22, 2004) (report and recommendation)). The Ninth Circuit opinion not only echoed these comments, but took them a step further stating,

Nothing significant in the background suggests guilt, therefore, and many factors suggest innocence. Indeed, not only was there no evidence of any ‘precipitating event that might have caused [Smith] to snap,’ but it is extremely unlikely that even a very troublesome act by seven-week-old Etzel would cause Smith to shake Etzel to death when his mother lay but a few feet away and easily available. A constitutionally permissible finding of guilt in this case therefore depends on the expert evidence of the cause of death.

Smith v. Mitchell, 437 F.3d 884, 889 (9th Cir. 2006), vacated sub nom. Patrick v. Smith, 550 U.S. 915 (2007). This inappropriate and speculative commentary from the appellate court does not reflect the proper postconviction review standard, which requires that the evidence be viewed in the light most favorable to supporting the conviction.
The dissenters apparently based their decision, at least in part, on a series of assumptions excerpted from the 2004 opinion of the Federal Magistrate Judge who had originally denied Smith’s habeas corpus petition. These included the following: (1) “Grandmothers, especially those not serving as the primary caretakers, are not the typical perpetrators [in shaken baby cases];”\(^{362}\) (2) Smith “was helping her daughter raise her other children;”\(^{363}\) (3) there was “no hint of [Smith] abusing or neglecting these other children;”\(^{364}\) (4) there was “no evidence of any precipitating event that might have caused [Smith] to snap and assault her grandson;”\(^{365}\) (5) Smith “was not trapped in a hopeless situation with a child she did not want or love[;] [n]or was she forced to single-handedly care for a baby that had been crying all day and all night;”\(^{366}\) (6) there was “no evidence that Etzel was doing anything other than sleeping the night he died;”\(^{367}\) (7) “Etzel’s mother[] was in the room next door when Etzel died;”\(^{368}\) (8) “[T]he medical evidence was not typical . . . .”\(^{369}\) These emotionally appealing, but speculative and even irrelevant, assumptions provide significant insight into the dissenters’ skewed jurisprudential approach. There are lingering questions after any case is decided. However, the federal magistrate who created this list, the Ninth Circuit that parroted it, and the three Supreme Court Justices who featured it in their decision, cannot plausibly argue that any or all of these assumptions demonstrate that Smith’s conviction was objectively unreasonable.

However, Justice Ginsburg relies on these assumptions to make a strange and very different argument. In the dissenters’ view, Smith is a “tragic” case.\(^{370}\) The death of a seven week-old child is always tragic. But Etzel’s death does not appear to be the focus of the Justices’ concern. In the next sentence, they reveal their belief that the only thing the Court has achieved by reviewing this decision from the Ninth Circuit is “to prolong Smith’s suffering and her separation from her family.”\(^{371}\) According to Justice Ginsburg, not only has Smith suffered enough, but she is a loving grandmother who “poses no danger whatever to her family or anyone else in society.”\(^{372}\) Returning Smith, who “the evidence indicate[s] . . . [is] warm hearted, sensitive, and gentle,” to prison will be “depriving Smith of the liberty she currently enjoys, and her family of her care.”\(^{373}\) Given the extensive inculpatory medical and nonmedical evidence presented at trial, the dissenters’

\(^{362}\) Smith, 132 S. Ct. at 8 (Ginsburg, J., dissenting).
\(^{363}\) Id.
\(^{364}\) Id.
\(^{365}\) Id.
\(^{366}\) Id.
\(^{367}\) Id.
\(^{368}\) Id.
\(^{369}\) Id.
\(^{370}\) Id. at 9.
\(^{371}\) Id.
\(^{372}\) Id.
\(^{373}\) Id. at 11–12.
unusual, extensive, and unsupported commentary regarding Smith’s good character during postconviction review merits careful consideration.

(a) Postconviction Distortion of the Medical and Nonmedical Evidence

An unfortunate reality of the postconviction review process is that pertinent facts from the trial record are often lost, ignored, distorted, taken out of context, or omitted as the case is presented in later hearings. In some cases, decreased judicial reliance on the facts may be appropriate, especially where the facts have little bearing on the legal analysis. However, in child abuse and child homicide cases, like Smith, problems arise as multiple layers of appeals move the courts further from the relevant facts.

At every level of appeal, the critical issue in Smith was the defendant’s claim that there was insufficient evidence of her guilt. This claim pertains to both the sufficiency of the medical evidence regarding the victim’s cause of death and the nonmedical evidence implicating the defendant’s acts as the cause of the victim’s fatal injuries. Because child abuse and child homicide cases invariably require judges and juries to examine these different types of evidence, we offer four corrections to the approach adopted in this case. These corrections elucidate how the dissenters’ distorted review of the record resulted in their legally and factually inaccurate conclusions. They also anticipate and may help prevent similar inaccurate evidentiary evaluations in future AHT/SBS cases.

(i) Empathy

Empathy for the defendant based on a reviewing court’s selective reading of the evidence is improper and antithetical to the deference owed the trial jury. The Smith dissenters’ speculations about the defendant (e.g., that she was generously helping to raise Etzel) or about grandmothers in general simply have no bearing on the only relevant legal question—whether it was unreasonable for the jury to have concluded (based on the medical and nonmedical evidence) that Smith had assaulted Etzel “by means of force that to a reasonable person would be likely to produce great bodily injury.”

The Justices’ unusual approach may help explain the omission of the bulk of the inculpatory evidence contained in the trial record and their argument (despite the clear habeas standard) that the Court should have simply refused to review the evidence.

---

375 See Mitchell, 437 F.3d at 888–90. The California Court of Appeals has already rejected defendant’s challenge to the sufficiency of the evidence during her direct appeal. Id. at 885. In her federal habeas petition, Smith claimed that her due process rights were violated because the evidence was “constitutionally insufficient” particularly with respect to “one element of the crime—the cause of the child’s death.” Id.
376 CAL. PENAL CODE § 273ab (West 2008).
Ninth Circuit decision.\textsuperscript{377} As the Court correctly—and presciently—recognized, the dissenter may believe “that Smith, who already has served years in prison, has been punished enough, and that she poses no danger to society[,] [t]hese or other considerations perhaps would be grounds to seek clemency,” but are not grounds for overturning the jury verdict in this case.\textsuperscript{378}

(ii) Preferencing the Nonmedical Evidence

As the Smith Court observed, the jury heard extensive medical evidence that “Etzel died from shaking so severe that his brain stem tore”\textsuperscript{379} and that this must have occurred while he was in the defendant’s care. In the face of this evidence, the dissenter concluded that Smith is a “warm-hearted,” “sensitive,” “gentle,” and “loving” grandmother. This suggests that the dissenter discounted the complex medical evidence and focused instead on the simple (although far from entirely) exculpatory nonmedical testimony from Smith’s relatives and the fact that the defendant was never previously charged with (or apparently even suspected of) physically abusing Etzel. Not only are these assumptions contradicted by the trial record, but it is an obvious logical fallacy to assume that a suspect was wrongfully convicted simply because she does not represent the stereotypical “majority” of offenders for a particular crime or because she has never been prosecuted or suspected of this crime in the past.

Although there have been some studies of the statistical profile of AHT/SBS offenders,\textsuperscript{380} the data tells us nothing about whether Smith shook her grandson hard enough on the night of November 29, 1996, to critically injure his brain. Moreover, as the dissenting Justices must be aware, child abuse is almost invariably committed by parents and caregivers who have displayed gentle and

\textsuperscript{377} According to Justice Ginsburg:

In sum, this is a notably fact-bound case in which the Court of Appeals unquestionably stated the correct rule of law. It is thus “the type of case in which we are most inclined to deny certiorari.” Nevertheless, the Court is bent on rebuking the Ninth Circuit for what it conceives to be defiance of our prior remands. I would not ignore Smith’s plight and choose her case as a fit opportunity to teach the Ninth Circuit a lesson.

Smith, 132 S. Ct. at 12 (Ginsburg, J., dissenting) (citations omitted) (quoting Kyles v. Whitley, 514 U.S. 419, 460 (1995) (Scalia, J., dissenting)). Of course, the Smith majority concluded that the Ninth Circuit had erroneously overturned Smith’s conviction, which means that any “liberty” Smith enjoyed was erroneously granted in the first place.

\textsuperscript{378} Id. at 7 (majority opinion).

\textsuperscript{379} Id.

\textsuperscript{380} See, e.g., Lazoritz et al., supra note 15, at 1011 tbl.1; Suzanne Starling et al., Abusive Head Trauma, supra note 49, at 260–61.
loving behavior towards the same children at other times. The assumption that a grandmother could not harm a child simply because she is a grandmother or because she has previously behaved gently is unsupported. Finally, Smith’s clean record also tells us nothing about her culpability. Child abuse is a crime frequently committed by caretakers who lack any prior criminal history, but who lose control during single or isolated episodes of violent conduct. As all of the judges who have reviewed this case on appeal were well aware, evidence of Smith’s violent temperament or prior bad conduct, if such evidence existed, would have been excluded from trial under the state rules barring propensity evidence.

(iii) Ignoring the Obvious and Relying on the Straw Man Argument

During their review of the medical and nonmedical evidence, the dissenters ignored the fact that Etzel was only seven weeks old and spent his brief life in a home surrounded by other relatives. The close proximity of other relatives may have insulated Etzel from other abusive acts because child abuse typically occurs

381 See Frasier et al., supra note 9, at 406 (noting that those who abuse children, “[w]ith rare exceptions . . . do so in private . . . and maintain a completely normal appearance in every other way”).

382 In child abuse cases, jurors and judges frequently have a difficult time conceptualizing parents as abusers of their children. They are reluctant to believe that an apparently loving caretaker would purposely injure their child. Many SBS cases do not involve caretakers who deliberately set out to kill or injure the child by violently shaking them. Rather these consequences often result from the caretaker’s loss of control and momentary violent behavior directed toward the child. This explanation does not mean that SBS cases are “unintentional” crimes within the legal definition of “intent,” or that the offender is less culpable because they did not specifically set out to kill or maim the baby. Shaking a vulnerable infant is still a deliberate act of extreme violence which justifies holding the caretaker accountable for the dire consequences which ensue.


384 The then applicable rule of evidence barred admission of bad character or prior conduct evidence for propensity purposes, but allowed it to show motive, intent, knowledge, or the absence of mistake or accident. Cal. Evid. Code § 1101(b) (West 2009). The current California evidentiary rule permits the evidence for propensity purposes.
when a caregiver is left alone with a child.\textsuperscript{385} While the dissenters use the fact that Etzel did not have any fractures or sprains as a straw man to suggest that Smith could not have shaken Etzel because she did not inflict other types of injuries, this argument cannot be sustained.\textsuperscript{386}

The lack of physical evidence of different types of abusive injuries is both unsurprising, in this case, and well documented in the child abuse literature.\textsuperscript{387} The dissenters’ concern about the lack of other types of abusive injuries also ignores the fact that the medical evidence did show that Etzel had suffered similar injuries in the past. Etzel’s autopsy revealed an earlier brain injury and older optic nerve bleeding, which clearly indicated prior head trauma. These findings were not disputed by the defendant’s experts.\textsuperscript{388} No evidence was presented at trial to establish that Smith was not in a position to have caused this prior trauma.

(iv) Discounting the Inculpatroy Nonmedical Evidence

According to Justice Ginsburg, the nonmedical evidence offered by the prosecution was “meager” because (1) “[t]here was no evidence whatever that Smith abused her grandchildren in the past or acted with any malicious intent on the night in question”\textsuperscript{389}; (2) there was no “motive or precipitating event that might have led Smith to shake Etzel violently”;\textsuperscript{390} (3) there was no “evidence [that] showed that Etzel was crying in the hours before he died”;\textsuperscript{391} and (4) “[a]ny loud crying likely would have woken Etzel’s siblings, Yondale, age 14 months, and Yolanda, age 4, asleep only feet away, even Etzel’s mother, Tomeka, asleep in the neighboring room.”\textsuperscript{392} As shown above, these speculations are not just a distortion of the record, they are legally irrelevant because the defendant was convicted

\textsuperscript{385} See ABUSIVE HEAD TRAUMA IN INFANTS AND CHILDREN: A MEDICAL, LEGAL, AND FORENSIC REFERENCE, supra, note 9 at 358; THE SHAKEN BABY SYNDROME: A MULTIDISCIPLINARY APPROACH, supra note 9, at 238, 284.

\textsuperscript{386} See supra note 345 and accompanying text.

\textsuperscript{387} See supra note 345 and accompanying text.


\textsuperscript{389} Id. at 11 (Ginsburg, J., dissenting).

\textsuperscript{390} Id.

\textsuperscript{391} Id.

\textsuperscript{392} Id.
under a California child abuse statute that specifically does not require proof of motive or malicious intent.\(^{393}\)

The dissenters’ skewed jurisprudential approach to the nonmedical evidence is best illustrated by their treatment of the defendant’s multiple admissions that she shook or “jostled” Etzel immediately before he died. According to Justice Ginsburg, Smith’s admission that she had given Etzel “a little shake, a jostle to awaken him” and then asked, “Oh, my God. Did I do it? Did I do it? Oh, my God,” “cannot be equated to a confession of guilt.” While Smith’s admission is strongly suggestive of her guilt, and likely was understood as inculpatory by the jury, Justice Ginsburg opines that “[g]iving a baby ‘a little shake, a jostle to wake him’” cannot be “an admission to shaking a child violently, causing his brainstem to tear.”\(^{394}\)

This conclusion is remarkable for several reasons. First, the dissenters claim the almost supernatural ability to determine (without citation to a single source) that what the perpetrator has described as “a little shake” cannot possibly result in traumatic head injury. Second, the dissenters ignore the legal standard, which requires that they take the evidence in the light most favorable to supporting the jury’s conviction, that is, that the jurors could have found that the defendant had admitted the specific mechanism of traumatic injury while simultaneously attempting to minimize the severity of her actions. Third, the dissenters ignore readily available medical research establishing that perpetrators often initially provide false information regarding how abusive injuries occurred.\(^{395}\) As the three dissenting Justices must know, perpetrators who make admissions frequently minimize the force or violence associated with their conduct.

The dissenters take a similarly dismissive approach, or simply ignore, the following additional inculpatory nonmedical evidence: (1) the undisputed fact that the defendant was the only adult in the living room with Etzel from midnight until he stopped breathing; (2) the evidence that Etzel was asymptomatic prior to the time the defendant was left alone with him; (3) the defendant’s statement that Etzel fell off the couch, a claim that is developmentally unlikely for a seven-week-old infant; (4) the defendant’s conflicting statements to the social worker and to the police; (5) testimony from the social worker that the defendant demonstrated shaking Etzel; and (6) evidence that Tomeka, Etzel’s mother, accused the defendant of causing Etzel’s death during the interview with the social worker.\(^{396}\)

\(^{393}\) See Cal. Penal Code § 273ab (West 2008) (“Any person who, having the care or custody of a child who is under eight years of age, assaults the child by means of force that to a reasonable person would be likely to produce great bodily injury, resulting in the child’s death, shall be punished by imprisonment . . . .”).

\(^{394}\) Smith, 132 S. Ct. at 11 (Ginsburg, J., dissenting).

\(^{395}\) See, e.g., Duhaime, supra note 91, at 410 tbl.2; Carole Jenny et al., Analysis of Missed Cases of Abusive Head Trauma, 282 JAMA 621, 622 (1999); James A. O’Neill et al., Patterns of Injury in the Battered Child Syndrome, 13 J. Trauma 332, 332 (1973). See also supra notes 168–169 and accompanying text.

\(^{396}\) Smith, 132 S. Ct. at 11 (Ginsburg, J., dissenting). According to the dissenters, “[t]he social worker also testified that Etzel’s mother, Tomeka, told Smith: ‘If it wasn’t for
(b) Preventing Future Courts from Making Similar Mistakes

Justices Ginsburg, Breyer, and Sotomayor are apparently convinced that “[w]hat is now known about SBS casts grave doubt on the charge leveled against Smith”397 in this case and more generally that “[d]oubt has increased in the medical community ‘over whether infants can be fatally injured through shaking alone.””398 As the foregoing review of the medical and nonmedical evidence demonstrates, appellate court conclusions in child abuse and child homicide cases may be based on a distorted analysis of the record and, as we discuss in a companion article,399 an equally skewed and inaccurate review of the medical literature.

The Smith dissenters’ myopic view of the evidence muddies their analysis of the legal and scientific questions and raises real concerns about the message sent to future courts, the media, and the public. Viewed most charitably, the opinion may reflect an empathetic rejection of the horrifying notion that anyone capable of being a loving grandmother could lose control once and shake her infant grandson hard enough to cause his death. However, the Smith dissent is dangerous because it appears to add legitimacy to the false AHT/SBS controversy, which continues to gain traction despite its lack of empirical or clinical support.400 Left uncorrected and misunderstood, these problems will make it harder for judges and juries to focus on the medical evidence and reach accurate verdicts. Outside the courthouse, we hope to prevent these decisions from promulgating misinformation about how infants can be injured or killed by shaking that directly threatens vital public health and child abuse prevention, investigation, and prosecution efforts.401

You this wouldn’t have happened.”” Id. at n.3. On this point, Justice Ginsburg simply ignored the requirement that the reviewing court take this evidence as true adding only that “Tomeka denied making any statement to that effect.” Id.

397 Id. at 10.
398 Id.
399 Moreno & Holmgren, supra note 17.
400 See sources cited supra notes 8–9.
401 See Christian & Block, supra note 15, at 1409–11 (setting forth the American Academy of Pediatrics position paper on AHT and discussing the public health and prevention efforts encouraged by the Academy to educate parents about the dangers of shaking).
MAPPING, MODELING, AND THE FRAGMENTATION OF ENVIRONMENTAL LAW

Dave Owen*

Abstract

In the past forty years, environmental researchers have achieved major advances in electronic mapping and spatially explicit, computer-based simulation modeling. Those advances have turned quantitative spatial analysis—that is, quantitative analysis of data coded to specific geographic locations—into one of the primary modes of environmental research. Researchers now routinely use spatial analysis to explore environmental trends, diagnose problems, discover causal relationships, predict possible futures, and test policy options. At a more fundamental level, these technologies and an associated field of theory are transforming how researchers conceptualize environmental systems.

Advances in spatial analysis have had modest impacts upon the practice of environmental law, little impact on environmental law’s structure or theory, and minimal impact on environmental law research. However, the potential legal implications of these advances are profound. By focusing on several of environmental law’s traditional core debates and by using urban development as a central example, this Article explores those implications. It shows that spatial analysis can change the problems environmental law addresses, the regulatory instruments environmental law uses, the entities law empowers to address those problems, and the methodologies of environmental law research.

INTRODUCTION

Imagine a proposed housing development—call it “Greenacres”—at the fringe of a metropolitan area. Greenacres will contain several dozen new homes,

---

* © 2013 Dave Owen. Professor, University of Maine School of Law. Research for this Article was supported by National Science Foundation award EPS-0904155 to Maine EPSCoR at the University of Maine and by the University of Maine School of Law. I thank Todd Aagaard, Colin Apse, Dmitry Bam, Eric Biber, William Boyd, Robin Craig, Lincoln Davies, Dan Farber, Eric Freyfogle, Kelley Hart, Rita Heimes, Blake Hudson, Michelle Johnson, Sanne Knudsen, Aga Pinette, Rebecca Purdom, J.B. Ruhl, Sarah Schindler, Sean Smith, Sarah Tran, Jennifer Wriggins, and participants in conferences and workshops at the University of Washington, Georgetown Law School, Vermont Law School, and the University of Maine for comments on earlier drafts, and Daniel D’Alessandro, Robin Campbell, John O’Hara, and Lauren Parker for research assistance.
all constructed on one-to five-acre lots. The developer plans to clear forests, fill wetlands, and replace undeveloped wildlife habitat with buildings, pavement, and landscaped yards. Stormwater runoff from Greenacres will pollute local streams and increase flooding risk. The new roads, buildings, and driveways will limit groundwater recharge, reducing local water supplies even as the houses increase water demand. The houses will also consume energy, most likely from fossil fuels, and the residents will burn gasoline while they drive, generating conventional air pollutant and greenhouse gas emissions. Greenacres will also bring benefits: profits for the current landowner and the developer, construction jobs, new housing options, an increased tax base, and potential customers and employees for area businesses. But those benefits come with an environmental price.

If viewed in isolation, each of these impacts might seem like a drop in a bucket, not worthy of regulatory oversight or response. When viewed in combination, however, and when combined with the impacts of other similar developments, Greenacres’ consequences might seem problematic. A regulator taking a holistic view might conclude that development should occur elsewhere, or that it need not occur at all. More plausibly, the regulator might negotiate changes that reduce or compensate for some of Greenacres’ impacts. That broader

1 This type of exurban development has rapidly expanded in recent decades. See Andrew J. Hansen et al., Effects of Exurban Development on Biodiversity: Patterns, Mechanisms, and Research Needs, 15 ECOLOGICAL APPLICATIONS 1893, 1893–94 (2005).


4 See William E. Odum, Environmental Degradation and the Tyranny of Small Decisions, 32 BIO SCIENCE 728, 728 (1982); Kevin M. Stack & Michael P. Vandenbergh, The One Percent Problem, 111 COLUM. L. REV. 1385, 1398–402 (2011); David M. Theobald et al., Ecological Support for Rural Land-Use Planning, 15 ECOLOGICAL APPLICATIONS 1906, 1908 (2005) (“The aggregate effect of land-use change is the result of many, relatively small individual decisions that are diffuse in space and time, made by a diverse array of planners and policymakers . . . .”).


6 This sort of negotiation is more common than flat prohibitions. See, e.g., Daniel A. Farber, Taking Slippage Seriously: NonCompliance and Creative Compliance in
perspective might also spur the adoption of legal measures—perhaps new administrative rules, or even legislation—to address the larger environmental impacts to which Greenacres is contributing. Not surprisingly, environmental commentators have spent decades arguing for such holistic review. Calls are legion for policymakers to consider their decisions’ impacts upon a wider variety of environmental, social, and economic outcomes; to consider broader spatial and temporal trends when making those decisions; and to involve more entities, both public and private, in decisionmaking processes.

Unfortunately, those aspirations have been difficult to fulfill. Environmental problems are notoriously complex, and considering the impacts of a range of activities, all dispersed across space and time, upon a variety of environmental media can be exceedingly challenging. The challenges become even greater

---


11 See, e.g., Jody Freeman & Daniel A. Farber, Modular Environmental Regulation, 54 DUKE L.J. 795, 797–98 (2005) (“There is rarely a single tool, or a lone agency at either the federal or state level, that is capable of producing the desired environmental benefit by itself . . . .”); Jody Freeman, Collaborative Governance in the Administrative State, 45 UCLA L. REV. 1, 28–29, 31–33 (1997).

12 See generally LAZARUS, supra note 7, at 6–19.

13 See, e.g., JOSEPH L. SAX, DEFENDING THE ENVIRONMENT: A STRATEGY FOR CITIZEN ACTION 56 (1970) (“The greatest problems are often the outcome of the smallest-scale decisions precisely because the ultimate, aggregate impacts of those decisions are so difficult to see and the pressures so difficult to cope with . . . .”); William W. Buzzbee, Urban Sprawl, Federalism, and the Problem of Institutional Complexity, 68 FORDHAM L. REV. 57, 58–60 (1999); Eric T. Freyfogle, Better Ways to Work Together, in THE EVOLUTION OF NATURAL RESOURCES LAW AND POLICY 98, 98 (Lawrence J. McDonnell & Sarah F. Bates eds., 2010) (“Other stresses stem from the difficulties of shifting resources to higher and better uses, and coordinating activities at landscape or watershed scales.”).
when, as is often the case, many different entities have information about an activity’s environmental impacts, opinions about the importance of those impacts, and partial capacity to respond. 

A central conflict of environmental law therefore has pitted the desire for holistic decisionmaking against the obvious need to keep decisionmakers’ tasks manageably discrete. For decades, legal commentators have debated how this tension should be resolved. 

Meanwhile, other environmental disciplines have transformed, and their transformation has significant but unappreciated implications for these debates. Over the past four decades, increased data availability, new software systems, and exponentially greater computing power have combined to turn spatial analysis—that is, quantitative analysis of data coded to specific geographic coordinates—into the coin of the environmental realm. At federal, state, and local government offices, in the private sector, and throughout nonlegal academia, thousands of analysts in dozens of fields now spend their days gathering and crunching spatial data. Their efforts serve a wide variety of purposes and, more fundamentally, 

---

14 See Freeman & Farber, supra note 11, at 797–98.


17 See infra Part II. The transformation extends well beyond environmental management and research. See, e.g., About NGA, NAT’L GEOSPATIAL-INTELLIGENCE AGENCY, https://www1.nga.mil/About/Pages/default.aspx (last visited Jan. 28, 2013) (describing activities of an agency that uses spatial data and analysis to promote national security); see also Keith Harrries, Mapping Crime: Principle and Practice (1999) (describing the science of crime mapping); Laxmi Ramasubramanian, Geographic Information Science and Public Participation 14–16 (2010) (describing the use of geographic data to prove racial redlining in insurance policy sales).

18 See sources cited supra note 17; see also Jacek Malczewski, GIS-Based Land-Use Suitability Analysis: A Critical Overview, 62 PROGRESS PLANNING 3, 5 (2004) (“Over the last forty years or so GIS-based land-use suitability techniques have increasingly become integral components of urban, regional and environmental planning activities.”). One author defines “spatial analysis” as representing “a collection of techniques and models that explicitly use the spatial referencing associated with each data value or object that is specified within the system under study.” Robert Haining, Spatial Data Analysis: Theory and Practice 4 (2004).


20 See, e.g., Tenley M. Conway & Richard G. Lathrop, Alternative Land Use Regulations and Environmental Impacts: Assessing Future Land Use in an Urbanizing
are leading to new ways of conceptualizing ecological systems and environmental change.\(^{21}\)

The emergence of spatial analysis merits revisiting environmental law’s traditional debates about integrative, holistic decisionmaking. Spatial analysis can facilitate better assessments of the cumulative environmental consequences of activities dispersed across space and time. By enabling analysts to simultaneously evaluate a variety of environmental impacts, spatial tools and models can allow concurrent pursuit of multiple environmental goals.\(^{22}\) And by producing maps, which are a compelling and accessible means of conveying information, spatial analysis can improve communication among the many entities involved in environmental policymaking.\(^{23}\) In short, spatial analysis can facilitate more integrative approaches to environmental law. Spatial analysis technologies are by no means perfect tools, and they cannot turn environmental regulators into omniscient seers.\(^{24}\) But they still can change our approaches to environmental protection.

Despite that potential, legal thinkers have devoted little attention to spatial analysis. Legal-academic literature does contain abundant references to geographic information systems (GIS), and most practicing environmental lawyers are at least vaguely aware of the increasing pervasiveness of spatial analysis tools.\(^{25}\) Environmental law researchers also increasingly draw on nonlegal literature, and many of the scientific and economic articles they cite draw on spatial analysis. But very few legal authors have considered whether emerging spatial analysis techniques hold transformative potential for either the practice or theory of

---

\(^{21}\) See, e.g., MARINA ALBERTI, ADVANCES IN URBAN ECOLOGY: INTEGRATING HUMANS AND ECOLOGICAL PROCESSES IN URBAN ECOSYSTEMS (2008); see also NAT’L RESEARCH COUNCIL, GRAND CHALLENGES IN ENVIRONMENTAL SCIENCES 4 (2001) (identifying spatial modeling of land use change as one of the greatest future challenges for environmental science).

\(^{22}\) See, e.g., Theodore C. Weber & William L. Allen, Beyond On-Site Mitigation: An Integrated, Multi-Scale Approach to Environmental Mitigation and Stewardship for Transportation Projects, 96 LANDSCAPE & URB. PLAN. 240 (2010). For more discussion, see infra Part III.

\(^{23}\) See infra Part III.

\(^{24}\) See infra notes 169–174 and accompanying text.

\(^{25}\) As of June 27, 2013, a search of Westlaw’s journals and law reviews database for the phrase “geographic information system” produced 692 hits. Most of these articles contained passing references to GIS systems, and the articles that discuss GIS in more depth generally focus on evidentiary issues and privacy questions.
environmental law. Nor do legal researchers typically use spatial analysis tools. Even as other research fields move toward quantitative analysis based on spatial data, environmental law research remains largely the domain of qualitative argument, often grounded in intuition and anecdote and delivered exclusively in prose.

This Article argues for bridging the divide between spatial analysis and environmental law. Part I summarizes some of the classic fragmentation challenges of environmental law, and thus maps problems that spatial analysis might help law address. It first discusses fragmentation of different environmental regulatory programs, then fragmentation across space and time, and then federalism-based debates about decisionmaking authority. Part II turns from traditional legal debates to the technological and theoretical evolution of spatial analysis. Part III explores some of spatial analysis’s implications for environmental law. Using land use as a central example, it explains how spatial analysis can change which environmental

---


28 With the emergence of empirical legal studies, the primacy of qualitative arguments is fading. And some empirical legal research does draw upon data coded to geographically defined jurisdictions, like counties or congressional districts. See, e.g., Lisa R. Pruitt & Beth A. Colgan, *Justice Deserts: Spatial Inequality and Local Funding of Indigent Defense*, 52 Ariz. L. Rev. 219 (2010).
problems we find cognitively tractable, what tools we use to address those problems, and to whom we allocate authority to respond. Finally, Part IV focuses on legal research, explaining how spatial analysis could generate more empirically grounded and practically useful academic inquiries about environmental law.

Throughout, the Article also discusses limitations of spatial analysis, which can suffer from the opacity, manipulability, and false certainty that plague any complex and quantitative mode of analysis. It does not claim that spatial analysis will readily or easily solve environmental law’s challenges, and in some circumstances spatial analysis tools will be too reductionist or too cumbersome to improve environmental regulation and research. Nor does this Article argue that better information will always lead to better decisions. As the politics of climate change have thoroughly demonstrated, such an expectation is unduly optimistic. But despite these limitations, the emergence of spatial analysis is an important, and potentially quite positive, development for environmental law.

I. THE PERSISTENT FRAGMENTATION OF ENVIRONMENTAL LAW

The years 2008 and 2009 brought the Atlanta metropolitan region too much sun, too little rain, and a big legal scare. For decades, the metropolitan area had grown rapidly, piling one Greenacres-style development upon another and becoming a poster child for suburban sprawl. Greater Atlanta’s growth led to massive increases in water use, and when drought struck, the Atlanta region stretched the limits of its water supply. Then the United States Court of Appeals for the D.C. Circuit and a federal district court held that greater Atlanta was using water to which it had no legal entitlement, raising the specter of legal limits atop

---


30 See generally Irene Lorenzoni & Mike Hulme, Believing Is Seeing: Laypeople’s Views of Future Socio-Economic and Climate Change in England and Italy, 18 PUB. UNDERSTANDING SCIENCE 383, 393–94 (2009) (finding that prior beliefs influence people’s willingness to accept new information).


The drought has since ended, and an Eleventh Circuit decision placed Atlanta’s water use on a less tenuous footing—providing a respite, at least, from the apparent water supply disaster.\footnote{See In re MDL-1824, 644 F.3d at 1166 (holding that the Army Corps of Engineers had authority to deliver water to Atlanta).} But water conflict continues, and many other growth problems persist.\footnote{See Carole Rutland, No Way to Run a River—After More than Two Decades of Water War, We Need a Truce—and a Fresh Approach, LEDGER-ENQUIRER (Columbus, Ga.), Nov. 27, 2011 (“The case now lingers as Alabama and Florida think about their next move.”).} Atlanta has not attained federal air quality standards, and its nonattainment status is partly caused by a sprawling, automobile-dependent growth pattern.\footnote{See Michael Lewyn, How City Hall Causes Sprawl: A Case Study, 30 ECOLOGY L.Q. 189, 191–92 (2003) (reviewing LARRY KEATING, ATLANTA: RACE, CLASS, AND URBAN EXPANSION (2001)); Larry Hartstein, Atlanta’s Air Quality: Better, but Still Bad, ATLANTA J.-CONST., Apr. 29, 2010, available at http://www.ajc.com/news/news/local/atlantas-air-quality-better-but-still-bad/nQfcf/.} At the region’s urban fringe, development has clashed with the protective mandates of the Endangered Species Act (ESA).\footnote{See Seth J. Wenger et al., Runoff Limits: An Ecologically-Based Stormwater Management Program, 9 STORMWATER 1 (2008) (describing impacts on protected aquatic species in the Etowah watershed).} Other social and environmental problems associated with sprawl—traffic congestion, for example, and isolation of people (particularly the socially and economically disadvantaged) from workplaces, services, and each other—continue to plague the region.\footnote{See Robert D. Bullard et al., The Costs and Consequences of Suburban Sprawl: The Case of Metro Atlanta, 17 GA. ST. U. L. REV. 935 (2001).} While Atlanta may present an extreme case, it is not unique. Similar tensions between development and environmental quality recur across the country.\footnote{See, e.g., Lincoln Davies, Just a Big, “Hot Fuss”? Assessing the Value of Connecting Suburban Sprawl, Land Use, and Water Rights Through Assured Supply Laws, 34 ECOLOGY L.Q. 1217, 1219–25 (2007) (discussing tensions between growth and water supplies); Ben Giles, Chesapeake Bay Cleanup Could Cost Prince George’s $800 Million, WASH. EXAMINER, Nov. 27, 2011, at 4 (describing the costs of water pollution, partly derived from urbanization, in the Chesapeake Bay).}

These tensions did not arise in a legal void. Many growth areas boomed after the early 1970s, when a series of federal and state statutes created legal standards that these areas now fail to attain.\footnote{See U.S. ENVTL. PROT. AGENCY, OUR BUILT AND NATURAL ENVIRONMENTS 4–8 (2001) (describing rapid growth in recent decades).} Nor did these problems emerge because growing populations inevitably require our current pace of environmental degradation. By regulating the configuration, layout, and landscaping of...
developments, communities can minimize or mitigate many environmental impacts, sometimes while imposing relatively small costs on developers and creating more livable communities. Instead, one important reason why greater Atlanta and its sprawling brethren have grown problematically is that it is exceedingly difficult for any single entity to grasp, let alone respond to, the full range of impacts of sprawl.

That challenge exemplifies an often-criticized feature of the United States’ system of environmental law. Too often, critics argue, environmental law depends upon regulatory agencies addressing one environmental goal and one project at a time, and doing so with insufficient involvement from other agencies, levels of government, affected firms, or members of the public. While alternative approaches exist, their informational demands can strain the cognitive capacities of the human mind. This section explores three prominent examples of that fragmentation—specifically, fragmentation across environmental media, space and time, and governmental jurisdictions—and the continuing debates about an optimal response.

A. Fragmentation Across Environmental Media

Environmentalists often cite the so-called First Law of Ecology: that everything is connected to everything else. That law captures the widely shared view that human actions have far-reaching consequences, which are not confined to air, water, or any other single environmental medium. Greenacres, for example, would likely impact air quality, water quality, wildlife habitats, energy

---


42 See generally Buzbee, supra note 13, at 63–74 (describing causes and effects of sprawl). This is not the only reason: consumer preferences, racial biases, poor urban schools, and the economic influence of development interests all also play substantial roles in sprawling development patterns. See generally ANDRES DUANY ET AL., SUBURBAN NATION: THE RISE OF SPRAWL AND THE DECLINE OF THE AMERICAN DREAM (2000).

43 See, e.g., Stewart, supra note 9, at 21.


45 Jonathan Cannon, Environmentalism and the Supreme Court: A Cultural Analysis, 33 ECOLOGY L.Q. 363, 369–70 (2006) (“Environmentalists share a belief that . . . human intervention affecting one part of a human-natural system can be expected to have deleterious effects elsewhere in the system.”).
use, and aesthetics. Regulatory initiatives designed to control these impacts will create their own collateral effects. The consequences will not merely be environmental, for environmental protection is inextricably intertwined with economics and health. These interconnections inevitably inspire calls for holistic regulatory approaches that take into account the full range of consequences of any action.

Despite these calls, much of our environmental regulatory system is divided into media-specific compartments. Many (though not all) of the major federal environmental statutes focus on a single type of pollution or on protecting a single kind of environmental resource. The Clean Air Act, Clean Water Act, and ESA provide obvious examples. Land use regulation is typically addressed not just by separate laws, but also by different levels of government. Consequently, environmental regulation is often highly compartmentalized, with distinct agency offices applying separate statutes to address different environmental consequences of the same underlying action.

This fragmentation is problematic in several ways. First, it can lead to counterproductive regulation. Constraints designed to protect one environmental medium can encourage alternative activities with even worse environmental effects. Second, fragmentation could generate economically inadvisable

---


49 The National Environmental Policy Act (NEPA) provides a significant exception to this generalization, as do its state-law counterparts. 42 U.S.C. §§ 4331–4332 (2006); see infra notes 262–271 and accompanying text (discussing NEPA); see also 16 U.S.C. § 1455(d) (attempting to provide a framework for using planning to address multiple environmental issues in coastal zones).

50 See 16 U.S.C. §§ 1536, 1538 (providing protection only to threatened or endangered species); John Charles Kunich, *Preserving the Womb of Unknown Species with Hotspots Legislation*, 52 HASTINGS L.J. 1149, 1150 (2001) (“The ESA focuses on species, and moves to protect only one species at a time.”).


52 See Stewart, *supra* note 9, at 21.

regulation, as agencies unwittingly impose controls that create economic costs outweighing environmental benefits.\textsuperscript{54} Both of these potential problems have been exhaustively discussed in legal-academic literature, and avoiding them has been a recurrent justification for cross-media integration.\textsuperscript{55} However, underregulation may be a greater problem. Many environmental statutes and regulations establish thresholds, and if an activity’s impacts do not rise to those thresholds, more lenient regulatory controls apply.\textsuperscript{56} There are obvious reasons for adopting such thresholds,\textsuperscript{57} but sometimes an activity with media-specific effects that fall below those thresholds might seem inadvisable if all of its consequences are considered together.\textsuperscript{58} Greenacres, for example, might not strike a water quality regulator, an air quality regulator, a wetlands regulator, or a land use planner as problematic if each impact is considered separately. The collective effects of the project, however, might justify major changes, and perhaps even an outright regulatory denial.

For years, environmental policymakers have been aware of these problems, and they have tried to respond in many ways. One category of responses seeks to expand the analytical scope of environmental decisionmaking. NEPA, for example, attempts to compel more integrative thinking by requiring a single study of a broad range of environmental impacts.\textsuperscript{59} Mandates for cost-benefit and regulatory impact exemplify a similar impulse toward broadening analytical frames, albeit toward consideration of economic rather than environmental impacts.\textsuperscript{60} Concepts like

\textsuperscript{54} See Breyer, supra note 53, at 11 (criticizing administrative “[t]unnel vision”); Sunstein, supra note 8, at 1010, 1027–28.


\textsuperscript{57} Regulating small sources of environmental degradation can be difficult for administrators, costly for regulated entities, and at odds with a widely shared ideological commitment to regulatory minimalism. See Exec. Order No. 13563, 76 Fed Reg. 3821, 3821 (2011) (asserting that our regulatory system “must identify and use the best, most innovative, and least burdensome tools”); Stack & Vandenbergh, supra note 4, at 1395–98.

\textsuperscript{58} See, e.g., Fontaine, supra note 53, at 38–46 (describing a facility that for too long escaped vigorous enforcement, largely because different regulators did not realize that violations were part of a larger trend).


\textsuperscript{60} See Sunstein, supra note 55, at 1656–63 (summarizing arguments in favor of cost-benefit analysis). For a summary of requirements for federal administrative rulemaking, see Mark Seidenfeld, A Table of Requirements for Federal Administrative Rulemaking, 27 Fla. St. U. L. Rev. 533, 536–37 (2000). A less-cited but still important motivation for these requirements is to place procedural hurdles before agencies likely to take undesired
“sustainable development” reflect the same underlying goal, for sustainable development’s basic precept is that economic, social, and environmental systems should be viewed as integrated parts of a larger whole. The concept of “ecosystem management,” which now pervades the rhetoric of natural resource law, embodies similar ambitions.

Putting these ambitions into practice has not been easy. The Environmental Protection Agency (EPA) has been the focus of several reform movements, each with the primary goal of addressing multiple pollutants and impacts through consolidated permitting processes. Those efforts produced a few limited pilot programs and high-profile initiatives, but multimedia permitting processes remain rare. Cost-benefit analysis is now entrenched in administrative decisionmaking processes. But finding enough information to do a good cost-benefit analysis can be very difficult, and observers disagree vehemently about whether those analyses improve or worsen regulatory decisionmaking. Sustainable development is now one of the most pervasive buzz-phrases in the environmental field, but giving the concept a meaningfully precise definition, let alone transforming it into legal mandates, has not been easy.

The ecosystem management concept has helped...
produce some concrete results, but these, too, have a mixed track record of success, and critics have questioned the attainability of ecosystem management almost since the concept's emergence. In the decades since multimedia integration became a widely shared aspiration, Congress has done little to reorient environmental law or to create major integrating institutions. Instead, the fragmented statutory system commentators have been criticizing since the 1970s remains largely unchanged.

The persistence of fragmentation should not be entirely surprising, for any integrative initiative raises significant informational challenges. To understand the impacts of a project or regulatory action upon just a single environmental medium can be difficult. To understand the impacts of a single project or regulatory action across a range of media—and to understand all of the economic and social consequences of that action—may be much more than a single person or even agency office can accomplish. That problem may be addressed by pulling more people and offices into the project, but then a coordination challenge partially replaces the initial informational challenge. Fragmentation, for all its dysfunctions, can be administratively efficient, and the continued compartmentalization of environmental law reflects a tacit recognition of this reality.

---

68 The most notable examples are regional, multispecies habitat conservation plans developed under sections 9 and 10 of the ESA. See Matthew E. Rahn et al., *Species Coverage in Multispecies Habitat Conservation Plans: Where's the Science?*, 56 BIOSCIENCE 613, 613–14 (2006) (describing the increasing prevalence, and agency promotion, of this approach).


71 See Peter A. Buchsbaum, *Permit Coordination Study by the Lincoln Institute of Land Policy*, 36 URB. LAW. 191, 193 (2004) (documenting “a general consensus that environmental land use regulation continues to suffer from lack of coordination”).


73 Some environmental studies still address an impressive range of environmental consequences. *See, e.g.*, MINERALS MGMT. SERV., U.S. DEP’T OF INTERIOR, OCS PUBLICATION NO. 2008-040, CAPE WIND ENERGY PROJECT FINAL ENVIRONMENTAL IMPACT STATEMENT (2009) (providing detailed analysis of a broad range of impacts and alternatives). But that sort of comprehensive analysis is generally very expensive and time consuming to prepare.

74 See Krier & Brownstein, supra note 15, at 126 (“[D]isjointed incrementalism is necessarily the actual method of policy making in the real world.”).
That reality could change, however. In a complex world, some informational and coordination challenges will always accompany environmental decisionmaking. But if technology can effectively bolster the human mind’s capacity to process information, then adjustment of current fragmentary approaches still would be appropriate. A key question for the future of environmental law is whether such tools are beginning to emerge.

B. Fragmentation Across Space and Time

Compartmentalization along media-specific lines may be a central challenge for environmental law, but it is by no means the only fragmentation problem. Environmental regulation also routinely confronts decisionmakers with the need to think across spatial and temporal scales.75

Few environmental problems arise solely from the consequences of a single event, project, or decision. Instead, environmental degradation is often the consequence of many different actions spread across space and time.76 Greenacres, for example, might be just one of many developments in its watershed and air basin, and over time, the combined effects of those developments for water quality, water supply, and air quality might become significant.77 With climate change, the relevant impact could even be global in scale.78

The incremental causes of environmental challenges create an obvious need for integrated responses. If policymakers focus only on one event or location, they may not recognize an important threat.79 They also may respond inefficiently or inequitably. Some causes might be more cost-effectively redressed than others, but if regulators deal only with one activity at a time, they will miss those


76 See Buzbee, supra note 13, at 86 (discussing sprawl and the cumulative impact of individual decision); Theobald et al., supra note 4, at 1908–09.


78 See Stack & Vandenbergh, supra note 4, at 1402–12 (discussing the cumulative impact of greenhouse gas emissions).

efficiencies. Conversely, if regulators deal with each contributing source in isolation, they may fail to establish consistent standards or equitable distinctions.

For all of these reasons, environmental policymakers for years have sought to broaden the geographic and temporal scope of environmental analysis. They have done so through several techniques. First, many statutes and regulations call for “cumulative impact analyses.” Such analyses strive to place the potential impacts of a proposed activity in a broader context by considering the effects of other related projects and trends. Second, many environmental statutes, as well as most states’ land use laws, call for planning processes, which generally are designed to provide frameworks for decisions on individual projects or regulatory initiatives. Often planning and cumulative impact analysis are tightly coupled, with planning initiatives providing opportunities for more programmatic environmental analyses.

More recently, environmental policymakers have sought spatial and temporal integration through trading schemes. In their earliest and simplest form, these trading schemes expanded the geographic focus of regulation from individual smokestacks to facilities as a whole, and allowed regulated plants to compensate for emissions increases in one location through reductions elsewhere. The appeal of this approach was straightforward: regulators would still obtain their desired emission limitations, and regulated entities could find the cheapest place to put

---

80 See Ackerman & Stewart, supra note 16, at 1335 (stressing these disparities in cost, though in an argument for market-based schemes).
82 See U.S. ENVTL. PROT. AGENCY, supra note 79, at 2.
83 See, e.g., 16 U.S.C. § 1533(f) (2006) (requiring recovery planning for threatened and endangered species); id. § 1604 (requiring forest planning); 33 U.S.C. § 1313(e) (requiring water quality planning); 42 U.S.C. § 7410 (requiring ambient air quality planning); 43 U.S.C. § 1712 (requiring “land use plans”); see also Patricia E. Salkin & Amy Lavine, Regional Foodsheds: Are Our Local Zoning and Land Use Regulations Healthy?, 22 FORDHAM ENVTL. L. REV. 599, 611 (2011) (“Most state statutes require that zoning regulations be developed and implemented in accordance with a comprehensive land use plan . . . .”).

Like efforts at multimedia integration, these temporal and spatial integration efforts have faced challenges. For planning and cumulative impact analysis, the core problem is simple: doing either well requires gathering and processing a tremendous amount of information.\footnote{See OLIVER A. HOUCK, \textit{THE CLEAN WATER ACT TMDL PROGRAM: LAW, POLICY, AND IMPLEMENTATION} 63 (2d ed. 2002).} If many different governmental and private actors contribute to an environmental problem, even identifying all the activities that create that environmental problem can require a significant effort.\footnote{See Fine & Owen, \textit{supra} note 29, at 953–55 (describing the information gathering necessary to support air quality modeling).} Predicting the collective consequences of those many activities can be even more difficult. Environmental systems are often complex and dynamic, with synergistic effects and feedback loops complicating efforts at prediction.\footnote{See Ruhl & Salzman, \textit{supra} note 75, at 88–92 (describing these dynamics).} Consequently, cumulative impact analyses and comprehensive plans, while easy to call for, are often difficult to complete. The environmental law literature is filled with accounts of plans gone wrong,\footnote{See Fine & Owen, \textit{supra} note 29, at 962–64 (describing unsuccessful air quality planning); Arnold W. Reitze, Jr., \textit{Air Quality Protection Using State Implementation Plans—Thirty-Seven Years of Increasing Complexity}, 15 VILL. ENVTL. L.J. 209, 357–65 (2004) (calling state implementation planning a “failure”).} and many critics have argued that environmental planning’s unrealistic information demands doom it to failure.\footnote{See, e.g., HOUCK, \textit{supra} note 91, at 257 (describing planning-based approaches as “chronically difficult in their science and their political science”).} Similarly, cumulative impact analyses often appear to be neglected afterthoughts within environmental impact statements,
as though the agency drafting the statement viewed the cumulative analysis requirement as an inconvenient and unrealistic demand, to be complied with as cursorily as possible. As one EPA study succinctly observed, “Cumulative impacts . . . are not often fully addressed in NEPA documents.”

Trading schemes might seem to obviate some of these informational problems, for a market-based system theoretically can succeed without any single entity possessing synoptic knowledge of the activities at issue. In practice, however, trading schemes raise their own informational challenges. The traded things rarely are fungible. For example, a natural wetland that will be destroyed to allow development may be far more ecologically valuable than a replacement wetland constructed elsewhere, and a ton of emissions at the upwind side of an air basin—or adjacent to a low-income, minority community—may be far more problematic than a ton of emissions at the basin’s downwind edge. A recurring concern about environmental trading schemes therefore is that the trades will be chronically uneven, with the environment and, perhaps, the disadvantaged on the losing end, unless regulators review each trade. Providing that oversight, however, can be a substantial task, particularly if, as is often the case, the trading

---

96 See generally Michael D. Smith, Cumulative Impact Assessment Under the National Environmental Policy Act: An Analysis of Recent Case Law, 8 ENVTL. PRAC. 228 (2006) (finding that cumulative impact analyses were often found inadequate by courts, and concluding that “inadequate cumulative impact analyses continue to be major shortcomings in many NEPA documents.”). Smith’s observations are entirely consistent with my own experiences litigating NEPA cases.

97 U.S. ENVTL. PROT. AGENCY, supra note 79, at 1; see also Courtney Schultz, Challenges in Connecting Cumulative Effects Analysis to Effective Wildlife Conservation Planning, 60 BIOSCIENCE 545, 546 (2010) (“Past studies have found that [cumulative effects analysis] was absent or inadequate in many NEPA documents, and that the requirement has not been implemented to its full potential for numerous reasons, including a lack of monitoring data, funding, and adequate training.”). Studies of cumulative impact analyses under state environmental laws have found similar problems. See Ma et al., supra note, 81, at 397.

98 See Hahn & Hester, supra note 89, at 361–62 (identifying markets and marketable permits as an antidote to informational challenges).


scheme involves many actors and actions. Consequently, administering an environmentally protective trading scheme is often an information-intensive exercise, which raises transaction costs and limits efficiency. Partly for these reasons, many commentators remain skeptical about the utility of environmental trading schemes.

As with debates over multimedia integration, these debates over spatial and temporal integration remain unresolved. In practice, environmental law retains a mix of all of these approaches, with technology-based systems, trading systems, and planning systems often overlapping in ways that defy easy categorization, and with the proper balance among those approaches still subject to vigorous discussion. That balance also could change. Our capacity for spatially and temporally integrative decisionmaking is limited largely by our capability for processing information. If technology is enhancing that capacity, then integrative regulatory approaches should be increasingly viable.

C. Institutional Fragmentation

These challenges of spatial, temporal, and media-based fragmentation are intertwined with challenges of institutional fragmentation. Most major environmental problems implicate federal, state, and local regulatory authority. They also affect the interests of private businesses, advocacy groups, and individuals. Often the knowledge necessary to understand environmental

102 See James Salzman & J.B. Ruhl, “No Net Loss”: Instrument Choice in Wetlands Protection, in MOVING TO MARKETS IN ENVIRONMENTAL REGULATION, supra note 90, at 323, 338–39 (discussing how this tension affects wetlands trading). For air pollution trading, some studies have concluded that informational burdens are more manageable. See Winston Harrington & Richard D. Morganstan, International Experience with Competing Approaches to Environmental Policy: Results from Six Paired Cases, in MOVING TO MARKETS IN ENVIRONMENTAL REGULATION, supra note 90, at 95, 117–18.
104 See generally MOVING TO MARKETS IN ENVIRONMENTAL REGULATION, supra note 90 (containing multiple views of markets, some complementary and others less so). For disparate views on environmental planning, compare Houck, supra note 91 (criticizing planning), with THE WHITE HOUSE COUNCIL ON ENVTL. QUALITY, FINAL RECOMMENDATIONS OF THE INTERAGENCY OCEAN POLICY TASK FORCE (2010) (calling for a massive new planning initiative).
problems is dispersed throughout these complex institutional landscapes, and solving environmental problems is impossible without coordination across both jurisdictional and public-private boundaries. 107

This dispersal of knowledge and authority frustrates environmental law at multiple levels. If multiple agencies hold responsibility over different aspects of the same activity, they may act at cross-purposes. Local land use regulators, for example, might pass large-lot zoning requirements designed to preserve aesthetic qualities (or, more insidiously, socioeconomic segregation), 108 yet those requirements can spread development across more of the landscape, creating perverse outcomes for water quality protection, habitat protection, air quality, and energy use. 109 Energy regulators might try to promote energy-efficient power plant cooling systems even as water quality and fishery regulators complain of impacts upon aquatic systems. 110 Regulators also may not act at all. An upstream or upwind state, for example, may have little incentive to control pollution emissions. 111 Even where multiple jurisdictions share the burden of an environmental problem, a “regulatory commons” dynamic, in which no agency has enough incentive to act, can preclude effective responses. 112 Combinations of inaction and conflicting action also may arise. When they do, as the bungled response to Hurricane Katrina illustrates, the result can be costly. 113 Addressing these problems of institutional complexity therefore remains another central challenge of environmental law.

These problems are centrally important to debates about environmental federalism. By design, our government is a system of divided authority, with federal, state, and local authorities and a robust private sector all theoretically playing important roles. 114 But when these different institutions come into conflict, questions arise about who holds decisionmaking authority and where jurisdictional boundaries lie. 115 For years, those questions have formed one of environmental law’s key battlegrounds, with jurists and commentators asserting dramatically

107 Freeman & Farber, supra note 11, at 797–98.
111 See Daniel C. Esty, Toward Optimal Environmental Governance, 74 N.Y.U. L. REV. 1495, 1543 (1999) (“[P]rogress on acid rain would likely never have been made as long as the issue were left to state level initiative . . . .”).
112 See Buzbee, supra note 75, at 6.
114 See Bond v. United States, 131 S. Ct. 2355, 2364–65 (2011) (arguing that federalism protects political liberty).
different views about how our federalist system should face the challenges of jurisdictional fragmentation.\footnote{See infra notes 309–319 and accompanying text.}

These challenges also raise important questions about how to encourage effective coordination among multiple institutions.\footnote{See Ruhl & Salzman, supra note 75, at 64–65 (explaining the prevalence of these challenges).} In response to these questions, many legal scholars have explored what makes interjurisdictional coordination succeed.\footnote{See, e.g., Holly Doremus, CALFED and the Quest for Optimal Institutional Fragmentation, 12 ENVTL. SCI. & POL’Y 729 (2009); Freeman & Farber, supra note 11; Ruhl & Salzman, supra note 75, at 109–19.} They have focused primarily on bureaucratic structures, divisions of authority, and measures for public participation.\footnote{See, e.g., Freeman & Farber, supra note 11, at 798 (offering “modular regulation” as a solution to coordination challenges); Ruhl & Salzman, supra note 75, at 109–19 (promoting solutions based on “weak ties networks”).} Such questions of power and procedure obviously are very important. But an equally consequential, and largely unexamined, set of questions involves the substance of interjurisdictional communication. Different agencies have different goals and cultures, rely on different data, and use different methods and terminology for communication.\footnote{See generally Eric Biber, Which Science? Whose Science? How Scientific Disciplines Can Shape Environmental Law, 79 U. CHI. L. REV. 471 (2012) (exploring these differences).} Private firms and public participants often bring their own divergent perspectives and knowledge to the table. Finding common languages for these participants to pool information, develop shared understanding, and identify areas where their goals coincide or conflict therefore is crucially important. A key question for environmental law, then, is whether new mechanisms for communication should change approaches to environmental regulation.

D. Intertwining Systems of Fragmentation

These problems of fragmentation among media, within space and time, and across jurisdictions often occur in combination. With Greenacres, for example, impacts on different environmental media would be regulated not just by different statutes but also by different agencies—some local, some state, and some federal.\footnote{See Buzbee, supra note 13, at 91 (describing the dispersion of authority over sprawl’s causes and effects).} The impacts also would likely spill across municipal, state, and sometimes even national boundaries.\footnote{See, e.g., U.S. ENVTL. PROT. AGENCY REGION 3 ET AL., CHESAPEAKE BAY TOTAL MAXIMUM DAILY LOAD FOR NITROGEN, PHOSPHORUS AND SEDIMENT ES-3 (2010) (describing sources of impairment of Chesapeake Bay); W.R. Stockwell et al., Ozone Formation, Destruction and Exposure in Europe and the United States, in FOREST DECLINE AND OZONE 1, 1 (Heinrich Sandermann et al. eds, Ecological Studies, vol. 127, 1997) (describing regional ozone transport); Hari M. Osofsky, Is Climate Change}
differences will create barriers to coordination—a volunteer local planning board will likely frame issues quite differently from a federal agency dominated by wildlife biologists or air quality scientists and engineers. The common result is what policy analysts refer to as a “wicked” problem, in which simply defining the scope of the regulatory challenge, let alone resolving it, is very difficult.

In practice, those difficulties often seem insurmountable. Rather than coordinate effectively, state and federal regulators may initially leave oversight of Greenacres almost entirely under local control. Local regulators, though perhaps generally aware that development affects habitat protection, water supply, water quality, air quality, and a variety of other environmental outcomes, may have little idea how to translate those broad concerns into site-specific regulatory controls. Often, it is only when development patterns clearly become incompatible with state or federal environmental quality mandates that local, state, and federal entities attempt to coordinate—or resign themselves to do battle. By that time, proactive solutions are unlikely to be available. The remedies instead will be expensive, if they are implemented at all, and both local autonomy and environmental quality will suffer. This dysfunctional dynamic creates an acute need to find a better way.

II. THE EMERGENT GEOCODED AGE

Forty years ago, when environmental law began developing its current responses to these challenges of fragmentation, the term “geographic information systems” was hardly ever used. Computer-based modeling was in its infancy,


I base this claim on experience with local boards and with agency scientists and engineers. The differences in expertise and perspective are often profound.


See Buzbee, supra note 13, at 91 (noting presumptions favoring local control).

Their inaction also may be motivated by the local political influence of prodevelopment entities and by collective action problems. See id. at 77–91 (exploring the political dynamics of sprawl). In working with local governments, however, I have often found genuine interest in protecting environmental quality but little understanding about how to connect those overall goals to specific land use decisions.

See, e.g., Owen, supra note 2, at 480–83, 502–03 (describing innovative but belated water quality protection efforts); Nat’l Ass’n of Home Builders v. San Joaquin Valley Unified Air Pollution Control Dist., 627 F.3d 730, 731–32 (9th Cir. 2010) (discussing a rule, adopted only after years in nonattainment status, designed to control ozone precursor emissions from development), cert denied, 132 S. Ct. 369 (2011); supra notes 31–39 and accompanying text (discussing the Atlanta region).

Prototypes of modern GISs were emerging but not in widespread use. See KEITH C. CLARKE, GETTING STARTED WITH GEOGRAPHIC INFORMATION SYSTEMS 8–9 (1999) (describing the evolution of GIS).
and the processing capacity of computer systems was orders of magnitude lower than it is today. Researchers in many fields used statistics, but regression analyses of large, multivariable data sets were enormously time-consuming. Consequently, many environmental sciences were very different than they are now. Ecologists may have believed that “everything was connected to everything else,” but they had limited tools to understand how.

In the past four decades, those tools have evolved dramatically, and this Part turns from the dysfunctions of fragmented regulation to the coevolution of spatial analysis and environmental research. The discussion is necessarily quite general, and it covers only a small subset of the ways in which spatial analysis is now used. Nevertheless, even that subset illustrates a critical point: spatial analysis has important implications for any field, like environmental law, that depends on information about the physical or human environment.

A. The Emergence of Quantitative Spatial Analysis

In the mid-nineteenth century, London suffered a series of cholera outbreaks. The cause of cholera then was unknown; a leading theory postulated that the primary disease vector was a “miasma” emanating from an infected person’s body. But John Snow, a young doctor, suspected that the disease instead was transferred through fecal-oral contact. To test his hypothesis, Snow gathered and mapped data on cholera deaths during a particularly virulent outbreak

129 Environmental researchers understand the term “model” in a variety of ways, and one recent report defines a model as “a simplification of reality that is constructed to gain insights into select attributes of a particular physical, biological, economic, or social system.” NAT’L RESEARCH COUNCIL, MODELS IN ENVIRONMENTAL REGULATORY DECISION MAKING 31 (2007). If the term is used in that broad sense, a map or even a single equation qualifies as a model. In this Article, the term “model” holds a narrower meaning. I use the term to refer to simulation models, which take a series of data inputs, run numeric calculations according to prespecified rules, and produce numeric outputs. This narrower usage distinguishes models from maps. As discussed in more detail below, however, maps can be both inputs for and outputs from models.


131 See John O. McGinnis, Age of the Empirical, 137 POL’Y REV. 47, 49 (2006) (“One University of Chicago social scientist is said to have taken the entire summer to run a regression on a mainframe computer 40 years ago. Now researchers can run scores of regressions on their laptops in a few hours.”).


133 S.W.B. Newsom, Pioneers in Infection Control: John Snow, Henry Whitehead, the Broad Street Pump, and the Beginnings of Geographical Epidemiology, 64 J. HOSP. INFECTION 210, 211 (2006).

134 Id.

135 Id.
in London’s Broad Street neighborhood.\textsuperscript{136} His research showed, and subsequent studies confirmed, that the outbreak could be traced back to the contamination of a single well.\textsuperscript{137} Snow persuaded the city to close the well, ending the epidemic, and he eventually earned recognition as one of the founding figures of epidemiology.\textsuperscript{138} By coupling a database with a map—in other words, by using spatial analysis—Snow was able to solve a problem that previously had defied understanding, to save hundreds of lives, and to help redefine the methodologies of an emerging research field.

In the decades since, and particularly in the last forty years, researchers have turned Snow’s methodology into one of the predominant research and planning practices in many fields.\textsuperscript{139} That transition has been supported by several technological shifts, each of which has helped transform spatial analysis into a tool more powerful than John Snow ever could have imagined.

One shift was an enormous expansion in societal capacity to gather spatially coded data. Remote sensing, which uses satellites to document landscape features like elevation or infrared radiation, now allows researchers to quickly gather information about the distribution of landscape features and to track changes in land cover.\textsuperscript{140} Satellite photography serves similar purposes, and is particularly useful for mapping land use.\textsuperscript{141} With the widespread availability of global positioning systems (GPS), data gathered through more traditional technologies also can be geographically coded.\textsuperscript{142} Census data, for example, now are linked to specific geographic locations,\textsuperscript{143} as are many of the datasets gathered through hundreds of ongoing environmental monitoring programs.\textsuperscript{144} Funding has both spurred and followed these technological advances. The National Science
Foundation and other major supporters of scientific research continue to call for, and support, major initiatives focused on gathering spatially coded data.145

Computing advances also improved the availability of those data.146 Database programs now allow enormous datasets to be stored, searched, and transferred. In the 1980s, the emergence of personal computers allowed broader access to these databases; no longer did users need access to a large and expensive mainframe.147 In the 1990s and 2000s, with the enormous growth in Internet use, many databases became available online.148 These advances have not eliminated data shortages; despite improved technologies, data gaps remain persistent.149 But the amount of spatial environmental data available for analysis has vastly increased.150

Improved technology also has offered new ways to convey the information stored in these growing databases. Using maps to convey data is nothing new; traditional maps convey information about where roads, landforms, and jurisdictional boundaries are located.151 But traditional maps are unwieldy in some ways. Put too much data on the map and it becomes cluttered; put too little, and the map does not provide the information a user needs. Scale can be problematic; traditional paper maps do not allow a user to zoom in or out. The static nature of paper maps also presents difficulties. Landscapes evolve, but paper maps, once printed, do not.

---

145 See, e.g., NEON, http://www.neoninc.org/ (last visited May 25, 2013) (detailing NEON’s mission “to gather and provide 30 years of ecological data on the impacts of climate change, land use change and invasive species on natural resources and biodiversity”).

146 See Malczewski, supra note 18, at 9.

147 See CLARKE, supra note 128, at 9; Malczewski, supra note 18, at 10 (emphasizing the importance of “low-cost mini and [personal computer] platforms”). This evolution continues to unfold. See John D. Landis, A Brave and Better World? The iPad and the Future of Planning, PLANETIZEN (Feb. 7, 2012, 2:00 PM), http://www.planetizen.com/node/54337.

148 See Malczewski, supra note 18, at 12 (“All major GIS vendors are developing procedures for WWW-based access to data and models developed with their software.”). Government entities often play a major role in disseminating spatial data. See Peter M. Flannery, How to Pry with Maps: The Fourth Amendment Privacy Implications of Governmental Wetland Geographic Information Systems (GIS), 29 RUTGERS COMPUTER & TECH. L.J. 447, 454–55 (2003) (listing governmental sources). Datasets have value, however, and both private and some governmental entities therefore have incentives to make data available only for a fee. See Amy Wilson Morris & Adena R. Rissman, Public Access to Information on Private Land Conservation: Tracking Conservation Easements, 2009 WIS. L. REV. 1237, 1242; Allyson Phillips, A Portal to Reliable Real Estate Data or a Door to Nowhere?—A Look at How State and Local Dissemination Policies Have Impacted the Development of the National Spatial Data Infrastructure and Geospatial One-Stop Portal, 34 REAL ESTAT. L.J. 9, 9–10, 26 (2005).


150 See Malczewski, supra note 18, at 12 (describing “an explosion of digital data”).

151 See CLARKE, supra note 128, at 7.
Newer technologies address these problems in a variety of ways. Perhaps the most prevalent is to integrate graphical representations of many individual “datalayers,” with each layer describing a particular set of geographic features. To build a map, a user can select whatever series of datalayers serves her present purpose while leaving out any extraneous information, and can do so at her preferred scale. Once the datalayers have been created, the process is fast and increasingly user-friendly; with just a few mouse clicks, the planner can create a
customized map. Indeed, many GIS platforms allow Internet users—even users with very little technological sophistication—to create and customize their own maps. Datalayers also can be updated, so a map reflects the most recent version of the underlying databases, and maps themselves can be animated to show changes over time. In short, in a variety of ways, computers can turn maps into more versatile, dynamic, and accessible tools for conveying information.

Technological advances also facilitate more sophisticated quantitative analyses of that information. For generations, economists, environmental scientists, and other researchers have used statistical analysis as an important research tool. Their statistical work, however, was once limited by the storage and processing capacity of the human brain. Advances in computer processing capacity have dramatically changed the game. Computers now are capable of running millions of calculations in relatively short periods of time. Concurrent with those advances, GIS programmers have developed multiple ways to mathematically represent the geography of features in their databases, thus allowing statistical analyses of the spatial relationships among data points. With these advances, analysts now can analyze enormous spatially coded datasets to detect trends, correlations, and causal factors that even a few decades ago would have eluded discovery.

152 See Malczewski, supra note 18, at 12 (“The common interface tools like on-screen ‘buttons’ and drop-down menus . . . can be understood quickly and easily with the result that GIS can tap into the growing market of untrained users.”). I have watched planners do this, with results projected on a screen and participants suggesting changes. The ease with which maps can be customized is remarkable.


155 See Malczewski, supra note 18, at 11 (explaining how computing power spurred GIS development); Robert A. Pietrowsky, Foreword to CONVERGING WATERS: INTEGRATING COLLABORATIVE MODELING WITH PARTICIPATORY PROCESSES TO MAKE WATER RESOURCES DECISIONS, at vi (Lisa Bourget ed., 2011) [hereinafter CONVERGING WATERS] (“[T]echnology has transformed what is possible”).

156 See McGinniss, supra note 131, at 49.

157 See CLARKE, supra note 128, at 9 (describing the emergence of these techniques).

Partly because of these increases in computing capacity, spatial analysis has become inextricably linked with environmental managers’ large and growing dependence upon computer-based simulation modeling.159 From climate change to wildlife management, models now pervade almost every sub-field of environmental decisionmaking.160 Many of those models draw upon spatial data, and many produce spatially explicit outputs—which then can be used as input data by other models.161 Consequently, spatially explicit modeling has become a pervasive, and often indispensable, part of environmental management and research. At their best, these models add a whole new power to spatial analysis.162 Rather than just delineating the location of current landscape features, or, like John Snow’s research, teasing out causal relationships based on data about past events, they allow environmental managers to offer spatially explicit representations of possible futures.163

These changes represent more than just the emergence of a new set of technological tools. Improvements in hardware, software, and remote sensing technology would have only modest utility if not accompanied by an associated body of theory. Researchers now refer to this field as “geographic information science,” and it has its own professors, journals, conferences, blogs, and even subfields.164 Many researchers who would not define themselves as geographic information scientists also consider spatial analysis techniques to be centrally important to their discipline.165 Geography and planning, for example, are as reliant on spatial analysis as lawyers are upon web-based research systems like Lexis and Westlaw. In many other academic disciplines, spatial analysis also plays a growing role.166 Most universities offer courses in spatial analysis, and many

159 For a brief discussion of computer-based modeling, and how modeling differs from mapping, see supra note 129.
162 See Fine & Owen, supra note 29, at 904, 913 (discussing benefits of modeling).
163 See infra Part III (describing applications).
offer GIS certificates. Their graduates use those skills for city planning departments, consulting firms, and state and federal agency offices throughout the country. A new academic and professional discipline has emerged, with capabilities largely unheard of when our system of environmental law was formed, and with influence extending throughout—indeed, well beyond—the environmental field.

B. Spatial Analysis and Public Participation

While technological advances allow spatial analysts to do remarkable things, those advances are not an unqualified good. Any increase in the technological sophistication of decisionmaking creates the threat of overreliance on technology at the expense of common sense. Quantitative modes of decisionmaking almost invariably involve oversimplifications, subjective judgments, and data of uneven quality, but the apparent precision of the numeric outputs can conceal these limitations behind veils of false certitude. By combining visually compelling graphics with the deceptive precision of numbers, maps heighten these risks, and spatial analysis tools can easily be misperceived as “truth machines,” with their truths all dressed up in seductively pretty colors. Similarly, increased reliance on quantitative decisionmaking methodologies threatens to exacerbate digital divides, with highly educated or well-funded elites enjoying preferential access to powerful technologies, and poorer or less savvy stakeholders further excluded. Those fears have reverberated through the spatial analysis literature, with several articles warning, as one put it, that GIS “has effectively raised barriers to empowerment by creating exclusive, sophisticated user-communities beyond the reach of less powerful, resource poor citizens.”

---

169 See generally Wagner et al., supra note 29 (arguing that this threat is often realized).
171 See Farber, supra note 160, at 1674 (warning that “accessibility and clarity . . . may cause users to underestimate the amount of uncertainty associated with projections”). See generally Denis Wood, The Power of Maps (1992) (discussing the interests that maps serve).
172 See Wagner et al., supra note 29, at 296 (“[M]odels are often misunderstood as ‘truth machines’ . . . .”).
174 Id.
These threats of false confidence and digital divides are very real, but technological advances also can promote participation and inclusion. In part, that potential derives from increased access to computers and to data, and both general operating software and GIS-specific programs are far more intuitive than they were in decades past. The communicative power of maps also increases information availability, for maps can convey large amounts of information in an easily searchable and visually accessible format. Spatial analysts also can generate animations, computer-generated views of future land use patterns, and a variety of other visualizations, all capable of making future environmental change more cognitively “available.” Those advances, in combination with a massive effort by private nonprofit groups and by local, state, and particularly federal governmental entities to distribute maps and underlying data, have made extraordinary quantities of information not just accessible, but also potentially comprehensible, to millions of people.

Spatial analysts also have developed multiple ways to actively facilitate participation. With increased computing speed, GIS technicians can sometimes project alternative scenarios within a single meeting, and thus can map several alternative futures as participants watch. Environmental modelers also can involve lay people in building models. Land conservation modelers, for example, now often build models that allow community members to adjust the importance of different environmental goals—for example, habitat protection, groundwater protection, or aesthetics—and to produce multiple model runs showing how conservation strategies would be affected as preferences change. Similarly,

175 See CLARKE, supra note 128, at 9–10 (noting the importance of low cost, efficient personal computers in GIS development); Malczewski, supra note 18, at 18 (describing increasingly user-friendly GIS systems).


178 See, e.g., Kurt Stephenson & Leonard Shabman, Executing CADRe: Integration of Models with Negotiation Processes, in CONVERGING WATERS, supra note 155, at 23, 33 (describing “[g]aming or what-if exercises”); Greenprinting, supra note 20 (explaining TPL’s “greenprinting” process); see also Arnab Chakroborty, Enhancing the role of participatory scenario planning processes: Lessons from Reality Check exercises, 43 FUTURES 387, 391 (2011) (describing Reality Check Washington, a participatory GIS-based “one-day visioning exercise”).

180 See, e.g., Greenprinting, supra note 20.
modelers can ask expert panels to help develop algorithms for selecting parcels for conservation or development, rather than simply relying on the modelers’ intuitions or on historic data.\textsuperscript{181} Lay participants also can run some models, perhaps playing roles different from those they occupy in real life, and can thereby better understand the consequences of their actions and the perspectives of competing resource users.\textsuperscript{182} These methods for facilitating participation can be combined. Emerging processes known as “participatory modeling” or “computer-aided dispute resolution” involve stakeholders in iterative processes of building, running, critiquing, and rerunning models designed to resolve complex environmental management challenges.\textsuperscript{183}


\textsuperscript{183} Leonard Shabman & Kurt Stephenson, The Purpose and Goal for CADRe, in Converging Waters, supra note 155, at 9; see also Hulse et al., supra note 177, at 325 (describing citizen input for land and water use models).
This figure shows steps in the process of creating a Bayesian model of the suitability of land in Maine’s lower Penobscot River watershed for conservation. The modelers brought together a group of conservation planners and advocates, who identified and assigned relative values to various features they believe make land desirable for conservation. The modelers then combined the resulting algorithm with land use and land cover data to produce a map (shown here in draft form) identifying high-value conservation areas. They will later combine that map with maps of lands suitable for development, forestry, and agriculture to identify areas of potential conflict and compatibility. See infra Figure 4. Images courtesy of Robert Lilieholm, University of Maine Alternative Futures research project.
These efforts generally require substantial time investment, but they also can produce significant benefits.\(^{184}\) Perhaps most importantly, they can bring multiple perspectives into the model-development process, helping balance policy preferences and unexamined assumptions held by the modelers.\(^{185}\) Actively involving nonmodelers can be a more effective way of conveying information, allowing participants to be active rather than passive learners.\(^{186}\) Involvement also can promote both realism and trust. A participant who has helped develop a model—or, perhaps, helped run it—is likely to have a more realistic understanding of the limitations and uncertainties of the model, yet also may have more confidence that the modeling outputs represent a good faith effort toward objectivity.\(^{187}\) In an era when environmental debates are routinely undermined by both overconfidence in and distrust of scientific information, building that sort of realism can be crucially important.\(^{188}\)

Despite its promise, spatial analysis will never offer perfectly transparent, objective decisionmaking tools. Uncertainties, concealed subjective choices, and false precision are likely to remain ubiquitous.\(^{189}\) Modelers also face difficult tradeoffs between simple models, which are often faster and more transparent, and more complex models, which often, though not always, provide outputs that better correspond to the complexity of the real world.\(^{190}\) Even where models perform well, spatial information, like most information, can serve as an instrument of

---


Despite positive case studies, the empirical literature on the benefits of scenario evaluation and participatory modeling remains thin. Christian Albert, *On the Influence of Scenario-Based Landscape Planning—A Comparison of Two Alternative Futures Projects*, 28 PROBS. LANDSCAPE ECOLOGY 33, 33 (2010) (“[L]ittle light has so far been shed on [scenario planning’s] effectiveness to change local governance.”).

\(^{185}\) See Fine & Owen, *supra* note 29, at 926–30 (explaining the prevalence of judgment in modeling, and the associated importance of public input).

\(^{186}\) See Joan P. Baker et al., *Alternative Futures for the Willamette River Basin, Oregon*, 14 ECOLOGICAL APPLICATIONS 313, 321 (2004) (describing “greater stakeholder understanding, a feeling of ownership in the final product, and increased likelihood that the results will be used”).

\(^{187}\) See Stephenson & Shahman, *supra* note 179, at 28 (explaining the value of “model architecture that is transparent”).

\(^{188}\) See Laura K. Schmitt Olabisi et al., *Using Scenario Visioning and Participatory System Dynamics Modeling to Investigate the Future: Lessons from Minnesota 2050*, 2 SUSTAINABILITY 2686, 2700 (2010) (discussing this potential to “mediate the extremes of distrust or blind acceptance”).

\(^{189}\) See Fine & Owen, *supra* note 29, at 921–38 (discussing limitations of environmental modeling).

\(^{190}\) See *id.* at 924–26 (discussing advantages and disadvantages of model complexity).
power or exploitation, and increasing the availability of such information does not guarantee that it will be used in just or fair ways.191

Nevertheless, if carefully used, these tools have much to offer. If they are to act at all, environmental managers cannot avoid the necessity of trying to understand and, often, predict the behavior of complicated systems. Any set of tools that even incrementally improves those abilities therefore holds value. With spatial analysis, that value can be substantial.

III. SPATIAL ANALYSIS, ENVIRONMENTAL LAW, AND CHANGING LAND USE

Despite all the promise of spatial analysis, its emergence has done little to change environmental law. Law generates many lines on maps; between critical habitat designations, flood insurance mapping, and traditional zoning, to name just a few examples, environmental law and mapping are in some ways closely integrated.192 But while environmental lawyers often confront the products of spatial analysis, that does not mean they understand the tools their colleagues in other environmental fields now routinely use.193 Instead, lawyers, whether practicing or academic, rarely analyze spatial data.194 And while past changes in environmental science have raised basic questions about environmental law, the evolution of spatial analysis technology has generated only a limited academic reaction and minimal legislative or regulatory change.195 Among environmental fields, law stands alone in its diffident reaction to the emergent geocoded world.

If environmental law already dealt effectively with all the problems it faced, that diffident reaction would not be problematic. But clearly that is not the case. Environmental scholars routinely identify information management as a central


193 See Salkin & Nolon, supra note 26, at 526 (“While planning schools offer hands-on courses in the use of these new technologies, they are not typically part of the curriculum in law school.”).

194 See supra note 26 and accompanying text.

challenge, and that challenge severely constrains environmental law’s ability to turn the theoretical appeal of integrated regulation into a practical reality.\textsuperscript{196} For environmental law, then, the emergence of quantitative spatial analysis could represent a crucially important development. This section discusses that potential. The discussion is by no means exhaustive. Nevertheless, even a few examples from one subfield of environmental law should illustrate the potential breadth of spatial analysis’s implications and the ways it can facilitate more integrative regulatory approaches.

\textbf{A. What We Understand}

Perhaps the most important way spatial analysis can change environmental law is by improving our understanding of environmental problems. Almost any regulatory response to an environmental problem requires a demonstration that the problem exists, some explanation of its causes, and a reasonably robust grasp of how individual activities create problems manifested at broader temporal and spatial scales and across jurisdictional boundaries.\textsuperscript{197} If the regulatory response is to be effective, it requires some understanding of the negative tradeoffs and positive synergies likely to arise from regulatory intervention, including tradeoffs and synergies that span the compartmental boundaries of individual regulatory programs.\textsuperscript{198} Because of the complex, multiscale, and intertwined nature of environmental problems, achieving that understanding can be difficult, and those difficulties can limit the problems we respond to—or even recognize.\textsuperscript{199} As the following examples illustrate, however, advances in spatial analysis can expand that realm of understanding.

\textbf{1. Diagnosing Environmental Problems}

Sometimes an environmental problem and its source are obvious. Pollution belching from a factory or untreated sewage discharging from a pipe can demand attention, and the impacts can be difficult to miss. But many present-day environmental challenges—including many challenges associated with land use change—involves multiple stressors, some acting through geographically or temporally attenuated chains of causation.\textsuperscript{200} Understanding those causal

\textsuperscript{196} See, e.g., Holly Doremus, \textit{Data Gaps in Natural Resource Management: Sniffing for Leaks Along the Information Pipeline}, 83 IND. L.J. 407, 408 (2008); Wagner, supra note 16.

\textsuperscript{197} See supra notes 75–78 and accompanying text (discussing the tendency for large environmental problems to arise from an accumulation of smaller impacts).

\textsuperscript{198} See supra notes 44–58 and accompanying text.

\textsuperscript{199} See generally Boyd, supra note 26, at 847 (developing this argument).

\textsuperscript{200} See Carol M. Rose, \textit{The Story of Lucas: Environmental Land Use Regulation Between Developers and the Deep Blue Sea}, in \textit{ENVIRONMENTAL LAW STORIES} 239 (Richard J. Lazarus & Oliver A. Houck eds., 2005) (exploring how incremental change
relationships can be nearly impossible without a large, spatially coded data set and a computational model. Consequently, spatial analysis now plays a central role in monitoring environmental changes and in diagnosing environmental problems, particularly when those problems manifest themselves at broad spatial or temporal scales. The applications are far too numerous to list, but of many possible examples, the relationship between urbanization and water quality illustrates particularly well how spatial analysis can transform our capacity to understand environmental problems.

For decades, environmental scientists have understood, at least at a general level, that urbanization degrades water quality. As development progresses, both dry spells and floods become more extreme, aquatic and riparian habitats are degraded or disappear, water temperatures become warmer and more variable, and pollutant loads increase. Usually no single lot or even development project is the cause, and the degradation instead arises from the cumulative effect of dozens—perhaps hundreds—of land use decisions. At even sparse suburban densities, the result is usually an impaired waterway, particularly if the watershed is small.

Environmental scientists have understood these general dynamics for years. But building a regulatory regime has been difficult, largely because of the informational difficulties associated with translating a general understanding of watershed degradation into site-specific regulatory controls, and because of jurisdictional divides between federal environmental regulation and local land use control. The primary focus of water quality regulation instead has largely been the sort of large, discrete sources amenable to regulatory coverage under a temporally and geographically focused, media-specific regulatory regime. Consequently, many metropolitan areas developed with minimal regard to water

---

201 See Oreskes, supra note 158, at 70–71 (“One of the driving forces behind the increased use of computer models in the earth sciences is their applicability to systems that are too large, too complex, or too far away to study by other means.”).

202 See, e.g., NAT’L RESEARCH COUNCIL, supra note 129, at ix (stating that modeling is necessitated by “[t]he spatial and temporal scales on which environmental controls and environmental quality are linked”); Oreskes, supra note 158, at 71 (describing several applications).


204 See Christopher J. Walsh et al., The Urban Stream Syndrome: Current Knowledge and the Search for a Cure, 24 J. N. AM. BENTHOCOLOGICAL SOC’Y 706 (2005).

205 Owen, supra note 2, at 460.

206 See CTR. FOR WATERSHED PROT., IMPACTS OF IMPERVIOUS COVER ON AQUATIC SYSTEMS 2 (2003); NAT’L RESEARCH COUNCIL, URBAN STORMWATER MANAGEMENT IN THE UNITED STATES 20–26 (2009); Walsh et al., supra note 204, at 710.

207 See Owen, supra note 2, at 445–50 (explaining these challenges).

208 See id. at 446.
quality, and impaired waterways are now pervasive features of American suburban and urban landscapes.\(^{209}\)

In recent years, spatial analysis has helped researchers build a more robust conceptual foundation for addressing these problems. By comparing spatially coded water quality data with land use and land cover data, environmental scientists have refined their understanding of the dynamics of urban water quality impairment. A series of studies has identified a close relationship between impervious cover—primarily roads, roofs, and pavement—and water quality impairment, and has identified rough impervious cover thresholds above which water quality impairment almost invariably occurs.\(^{210}\) Those findings have transformative implications for water quality regulation. They connect individual increments of land development—an activity already subject to regulation, but through local land use regulations and building codes rather than federal water quality law—with a broader-scale environmental problem that previously seemed intractable.\(^{211}\) Some recent research initiatives have gone a step further and can bring spatial data on water supply withdrawals into the analysis.\(^{212}\) The result of this work is an improved understanding of the relationships between urban development and aquatic ecosystems.

In addition to helping researchers understand the dynamics of present problems, spatial analysis also can help identify areas where future water quality problems are likely to arise. Some studies have used “build-out” analyses, which assume future development to the full extent allowed by current zoning, to predict the extent and geographic location of future water quality problems.\(^{213}\) Those build-out analyses also can be adapted to predict impairment under alternative zoning regimes, providing a community with some sense of the implications of

\(^{209}\) See id. at 443–44 (describing the pervasiveness of the problem).

\(^{210}\) See CTR. FOR WATERSHED PROT., supra note 206, at 1–2; NAT’L RESEARCH COUNCIL, supra note 206, at 226–30. Some recent studies have found that degradation begins at even lower levels of development, while others have raised questions about whether widely used thresholds are adequately supported. See, e.g., Thomas F. Cuffney et al., Responses of Benthic Macroinvertebrates to Environmental Changes Associated with Urbanization in Nine Metropolitan Areas, 20 ECOLOGICAL APPLICATIONS 1384, 1398 (2010) (finding an onset of degradation at lower levels); Glenn E. Moglen & Sunghee Kim, Limiting Imperviousness: Are Threshold-Based Policies a Good Idea?, 73 J. AM. PLANNING ASS’N. 161, 168–69 (2007) (arguing that the existing studies are based on very different methods of calculating impervious area, which limits their ability to support consistent thresholds). Nevertheless, the literature documents widespread agreement that increasing impervious area correlates with decreasing water quality.

\(^{211}\) See Owen, supra note 2, at 462–63 (describing how a focus on impervious cover can facilitate local responses).


\(^{213}\) See, e.g., Conway & Lathrop, supra note 20 (assessing build-out scenarios for coastal New Jersey).
alternative growth strategies.\textsuperscript{214} They also can expand the analytical focus from the community to the regional scale, examining, for example, the aggregate water quality impacts if individual communities use large-lot zoning or other growth-spreading regulatory regimes.\textsuperscript{215} By integrating economic growth models into future land use projections, modelers now can go beyond these build-out analyses and create more nuanced projections of the extent and location of future development.\textsuperscript{216} That could help water quality regulators and land use planners not just understand, but also improve, the relationship between developments like Greenacres and regional water quality trends.\textsuperscript{217}

By providing visually accessible outputs, spatial analysis also can help modelers and scientists explain water quality problems to lay audiences, and thus can expand the realm of understanding beyond technically sophisticated regulatory agency staff. Even a relatively simple concept, like the relationship between large-lot zoning and aggregate impervious cover levels, is far easier to explain with a conceptual map than with a series of sentences.\textsuperscript{218} Sophisticated models can add even more explanatory power. One intriguing example is the UVa Bay Game, a model that allows users to act as farmers, municipalities, environmental regulators, and other players in the complex dynamics of protecting Chesapeake Bay.\textsuperscript{219} Researchers report that the game already has proven a valuable educational tool, helping both students and actual stakeholders understand the interconnections and complexities involved in Chesapeake Bay management.\textsuperscript{220}

\begin{flushleft}
\textsuperscript{214} See, e.g., \textit{id.} at 9.


\textsuperscript{216} Several collaborators and I are currently developing a water quality model that uses this approach, which is particularly appropriate where zoning classifications are either aspirational, negotiable, or both, and a buildout scenario therefore is implausible. But buildout analyses have the advantage of using relatively simple, transparent rules. See Robert Gilmore Pontius Jr. & Neeti Neeti, \textit{Uncertainty in the Difference Between Maps of Future Land Change Scenarios}, 5 SUSTAINABILITY SCI. 39, 46–48 (2010) (stressing these advantages).

\textsuperscript{217} See, e.g., Baker et al., \textit{supra} note 186, at 316 (describing a modeling process that contrasted river and stream conditions under alternative future regulatory regimes).

\textsuperscript{218} See, e.g., U.S. ENVTL. PROT. AGENCY, \textit{supra} note 215, at 17.


\end{flushleft}
In sum, these advances are helping watershed scientists and regulators overcome the temporal, spatial, and jurisdictional fragmentation that often hamstrings efforts to address the environmental consequences of urbanization. And watershed scientists are by no means the only ones using spatial analysis to understand the environmental impacts of urbanization. For years, ecologists have used spatial models to track the impacts of regional development on habitat and to identify areas particularly under threat, and their analytical tools continue to evolve. See, e.g., Jeffrey A. Hepinstall et al., Predicting Land Cover Change and Avian Community Responses in Rapidly Urbanizing Environments, 23 LANDSCAPE ECOLOGY 1257 (2008) (using an “integrated modeling approach to simulate future land cover and predict the effects of future urban development and land cover on avian diversity in the Central Puget Sound region”); David M. Theobald, Targeting Conservation Action Through Assessment of Protection and Exurban Threats, 17 CONSERVATION BIOLOGY 1624 (2003) (using “socioeconomic indicators of risk” to measure “the proportion of conservation lands affected by developed areas”).

Figure 3: Large lot zoning and impervious area. From U.S. ENVTL. PROT. AGENCY, supra note 215, at 17.
traditionally have been attenuated, numerous studies now model the air quality implications of alternative development patterns, making quantification of the aggregate air quality impacts of developments like Greenacres increasingly feasible. Consequently, some new regulatory initiatives now integrate development permits and air quality control. Similar analytical tools could help link developments like Greenacres with specific increments of greenhouse gas emissions. All of these efforts increase our capacity to understand the relationships between developments like Greenacres and the goals reflected in a wide variety of environmental laws.

These advances have their limits, and water quality modeling exemplifies the continuing challenges as much as the positive potential. No water quality model can precisely and accurately quantify the impacts of each individual development, or even the aggregate consequences of a growth program or regulatory regime. Instead, data gaps and errors are ubiquitous, and some relationships remain poorly understood. Even where understanding is robust, models still must reduce the complexity of real-world ecological interactions to a relatively simple set of algorithms and equations. But if modeling produces a general understanding of the dynamics of environmental impairment and provides some basis for comparing alternative regulatory approaches, and if candid disclosures accompany the model results, even an imperfect effort can provide analytical traction for problems that regulators previously grasped only at the most conceptual of levels.

---

223 See TRANSP. RESEARCH Bd., supra note 3, at 3 (summarizing studies, and concluding that “[d]eveloping more compactly . . . is likely to reduce VMT”).
225 See TRANSP. RESEARCH Bd., supra note 3, at 2 (noting that greenhouse gas emissions associated with land use change are largely due to fossil fuel consumption, which also generates conventional air pollutants).
226 See Christer Nilsson et al., Ecological Forecasting and the Urbanization of Stream Ecosystems: Challenges for Economists, Hydrologists, Geomorphologists, and Ecologists, 6 ECOSYSTEMS 659, 660 (2003) (“[O]ur ecological forecasts were crude, largely qualitative in nature, and essentially based on expert knowledge . . . and correlative evidence . . . .”).
228 See ALBERTI, supra note 21, at 225–26 (noting that planners must evaluate the future implications of policies); Sierra Club v. Costle, 657 F.2d 298, 334–35 (D.C. Cir. 1981) (“The safety valves in the use of such sophisticated methodology are the requirement of public exposure of the assumptions and data incorporated into the analysis and the acceptance and consideration of public comment, the admission of uncertainties where they exist, and the insistence that ultimate responsibility for the policy decision remains with the agency rather than the computer.” (citations omitted)).
2. Recognizing Tradeoffs and Synergies

Helping researchers understand and explain a particular environmental problem, like the relationship between urbanization and water quality, is a significant achievement. But Greenacres’ water resource impacts will likely overlap with impacts on a variety of other environmental media and nonenvironmental outcomes. Similarly, some regulatory responses to Greenacres’ water impacts may create counterproductive results for other environmental goals, or for economic outcomes, while others may produce synergistic benefits. Understanding those tradeoffs and opportunities also has been a central challenge, often unmet, for environmental law. But in several ways, spatial analysis already is improving responses to multifaceted environmental problems.

One of the best illustrations of this potential involves mapping environmental constraints when siting development projects. Most development projects are subject to multiple regulatory constraints that can be depicted spatially. For example, wetlands, floodplains, zoning controls, conservation lands, and protected habitat areas all can be mapped, and all provide important signals about where a project might face a difficult regulatory process. Similarly, many landscape features desired by developers, like favorable soils and slopes, low taxes, quality school districts, proximity to complementary businesses, and access to roads and other preexisting infrastructure, also can be mapped. If the map layers are publicly available—and in many communities, they are—GIS technology allows developers to find areas that maximize positive features and minimize negative ones and also allows local officials to steer development projects toward particularly promising sites.

---

229 See supra notes 53–58 and accompanying text.
230 See supra Part I.A.
231 See generally Malczewski, supra note 18, at 4 (describing applications).
233 See, e.g., McCloskey et al., supra note 181, at 192–94.
The same approach can target conservation efforts. A land trust might overlay data layers showing wetland resources, aquifer recharge zones, rare plant and wildlife habitats, and potential habitat corridors to determine where to purchase conservation easements. At broader scales, conservation organizations often use

a process called “gap analysis,” which involves using GIS to identify regionally underprotected habitat types. They also can use economic development models to identify parcels where development potential is high, and therefore the threat to resources is larger, or, conversely, where development potential is lower, reducing purchase prices and potential conflict with community economic development goals. Likewise, economic models can explore whether purchase- or zoning-based strategies would more effectively accomplish conservation and economic goals.

Like other applications of spatial analysis, these uses have significant limitations. Turning land use and land cover databases, which usually contain gaps and inaccuracies, into suitability maps inevitably involves some distortion and oversimplification. Features like a community’s receptivity to development, the efficiency of a town’s regulatory approval processes, and the willingness of state or local governments to grant variances or leave laws unenforced all have important implications for development patterns but are not easily mapped. Consequently, spatial analysis is not a perfect tool for identifying all the implications of development proposals or regulatory initiatives. But it can facilitate simultaneous consideration of a variety of different environmental and nonenvironmental opportunities and constraints and exploration of some of the implications of planned activities. In these capacities, it now sees widespread use, with new innovations continuing to emerge.

3. Modeling Complicated Systems

The emergence of tools for conceptualizing individual environmental problems or siting individual projects is quite important in its own right, but these capabilities also form building blocks toward a larger goal. The complex relationships among multiple environmental challenges, human activities, and

---

236 See Theobald, supra note 221, at 1625 (describing U.S. Geological Survey’s Gap Analysis Program, which uses GIS technology to detect conservation “gaps”).
237 See, e.g., Kathleen A. Lohse et al., Forecasting Relative Impacts of Land Use on Anadromous Fish Habitat to Guide Conservation Planning, 18 ECOLOGICAL APPLICATIONS 467, 479 (2008) (describing this tradeoff).
239 See generally WOOD, supra note 171, at 25 (noting that maps often draw sharp lines where natural conditions form more of a continuum).
240 See Malczewski, supra note 18, at 4 (“GIS-based land-use suitability analysis has been applied in a wide variety of situations . . . .”). The Obama Administration has called for a major increase in spatial mapping in ocean and coastal areas. See THE WHITE HOUSE COUNCIL ON ENVTL. QUALITY, supra note 104, at 51–59.
regulatory policies create a need for ways to understand larger human-ecological systems. Again, innovations in spatial analysis hold promise.

In recent years, spatial analysts have begun creating comprehensive urban simulation models. Rather than modeling a single component of the urban system, these models would integrate multiple systems, including transportation, land uses, employment and economic outcomes, municipal finances, and a range of environmental outcomes, into systemic models. These models would have the capacity to simulate possible futures, testing, for example, the potential effects of different regulatory regimes upon a range of environmental and economic outcomes. More conceptually, the models should be able to explore and predict ways in which environmental outcomes both derive from and drive other urban dynamics. Ambitious though it may sound, the modelers’ goal is a quantitative, spatially explicit representation of the coupled human and natural ecology of urban systems.

---

244 See, e.g., Baker et al., supra note 186 (describing a multifaceted modeling effort for Oregon’s Willamette Basin).
246 See generally ALBERTI, supra note 21.
Modelers are pursuing this goal in a variety of ways. Some use “cellular automata” models, which divide the study area into grid cells, each assigned a particular set of values, and then project landscape change through an iterative process of updating each cell’s values based on the values of neighboring cells. “Agent-based models” simulate land use change by modeling the behavior of multiple actors, each trying to achieve some set of economic, social, or environmental goals, and each reacting to other actors and to the changing landscape. “Bayesian belief network” models use expert opinions to develop a set of model preferences and then combine those preferences with land use and land cover data to simulate future changes. Multiple other techniques exist, and

249 See, e.g., McCloskey et al., supra note 181, at 191.

Figure 5: Conceptual diagram of an urban growth model. Image courtesy of Charles Colgan, University of Southern Maine, and adapted from urbanism.org.
many of these approaches can be combined.\textsuperscript{250} Indeed, some of the most complex modeling systems are really aggregations of multiple models.\textsuperscript{251}

Despite this explosion of methodologies, building an urban model that provides reliable predictions remains an aspiration. The data needs of highly complex urban models are extraordinary.\textsuperscript{252} While some spatial data sets are now widely available, mismatches between the modelers’ needs and available data are almost certain to occur.\textsuperscript{253} Other modeling techniques, like Bayesian belief network modeling, create lesser (though still significant) data demands, but their reliance on expert opinion heightens the risk that the model will incorporate rather than challenge flaws in conventional wisdom.\textsuperscript{254} Additionally, any model’s internal logic is effectively a set of simplified assumptions about the behavior of human and environmental systems, and there are many aspects of that behavior that we poorly understand or that cannot be predicted with any real confidence.\textsuperscript{255} As the complexity of the model grows, the combined effect of uncertainties and judgments can grow as well, leading to poor simulations and inhibiting error detection.\textsuperscript{256}

For all of these reasons, the day is a long way off when urban modelers can take a municipality’s proposed general plan, plug it into a model, and quickly produce reliable predictions about air quality, water quality, ESA compliance, economic growth, governance costs, transportation efficiency, and the social equity implications of both regulation and development.\textsuperscript{257} Even further away, if it ever comes, is the day when modelers can provide those predictions for a single

\begin{itemize}
  \item \textsuperscript{250} See Peter Gomben et al., \textit{Impact of Demographic Trends on Future Development Patterns and the Loss of Open Space in the California Mojave Desert}, 49 ENVTL. MGMT. 305, 311–16 (2012) (explaining multiple techniques).
  \item \textsuperscript{251} See, e.g., Hulse et al., supra note 177, at 331 (diagramming a set of interlocking models); Waddell, supra note 242, at 303 (describing UrbanSim as “a software architecture for implementing models and a family of models implemented and interacting within this environment”).
  \item \textsuperscript{252} See Hulse et al., supra note 177, at 327 (describing a two-year data gathering effort).
  \item \textsuperscript{253} See Pontius & Neeti, supra note 216, at 41, 46–48 (describing sources of uncertainty).
  \item \textsuperscript{254} See McCloskey et al., supra note 181, at 191 (“[Bayesian belief network’s (BBN)] models are particularly useful when empirical data are limited and decisions are based largely on expert knowledge . . . .”), BBN modeling relies on people’s opinions, which may be informed and accurate, but people are often ignorant of biases driving their own decisionmaking. \textit{See generally} Daniel Kahneman, \textit{Thinking, Fast and Slow} (2011).
  \item \textsuperscript{255} See Pickett et al., supra note 245, at 352 (acknowledging that “identifying causal effects of land use change is extremely challenging”); Pontius & Neeti, supra note 216, at 41 (noting the potential for modeled algorithms to poorly reflect actual processes).
  \item \textsuperscript{256} See Nat’l Research Council, supra note 129, at 10 (advocating “model parsimony”).
  \item \textsuperscript{257} In comments on an earlier draft, Kelley Hart of the Trust for Public Land pointed out that the lack of specificity in many general plans also would limit such analyses.
\end{itemize}
development project. The current generation of highly integrative urban models is better viewed as a preliminary set of exploratory tools, usually best applied at broad geographic scales, rather than as comprehensively accurate and detailed representations of urban systems.

Nevertheless, even in their present state, urban growth models can help policymakers understand some of the likely environmental consequences of development trends and explore the sensitivity of those trends to different economic scenarios and regulatory interventions. Because of these capabilities, complex urban models already are important to environmental planners, and the push to develop more sophisticated tools shows no sign of abating. The time may yet arrive when many environmental and nonenvironmental regulatory processes can be linked through computer-based simulations, allowing policymakers to think about multiple environmental goals, at multiple scales, and across jurisdictional boundaries all at once.

B. How We Regulate: Changing Legal Instrument Selection

The evolution of spatial analysis has important implications not just for our conceptualization of the dynamics of environmental change, but also for the legal tools we use to address these dynamics. This section discusses two ways in which spatial analysis can support instruments that attempt to integrate regulation across spatial and temporal scales and media-specific boundaries, and thus can shift environmental law’s core debates about regulatory instrument choice.

1. Muddling Toward Synoptic Regulation

One of environmental law’s most venerable debates concerns regulatory approaches that demand information-intensive studies or plans. Environmental assessment laws like NEPA, which requires far-ranging environmental impact statements in advance of governmental actions, exemplify one version of this approach. Comprehensive planning statutes are another common example. In

---

258 See, e.g., R.D. Swetnam et al., Mapping Socio-Economic Scenarios of Land Cover Change: A GIS Method to Enable Ecosystem Service Modeling, 92 J. ENVT. MGMT. 563, 573 (2011) (noting that modeled scenarios can have greater value at larger geographic scales).
259 See id. at 564 (“[E]ach scenario should be thought of as a description of a possible future, albeit one which is plausible . . . .”).
261 See Gomben et al., supra note 250, at 311–16 (noting many existing uses); User List Sorted by Location, URBANSIM, http://www.urbansim.org/Main/UserListByLocation (last visited May 25, 2013) (listing dozens of countries where versions of UrbanSim have been used).
theory, these laws should facilitate the sort of broad, integrative thinking that almost everyone agrees is desirable. The practical utility of these laws has been hotly debated for decades, however, with one representative critique asserting that “NEPA ambitiously, and naively, demands the impossible: comprehensive, synoptic rationality, in the form of an exhaustive, one-shot set of ex ante predictions of expected environmental impacts. In the ordinary course of events, that task proves insuperable.” Planning’s skeptics often say much the same thing, and argue that alternative regulatory approaches should receive more emphasis.

These debates have important practical implications. Both environmental assessment laws and planning statutes occupy influential roles in federal and state environmental law. But they are not ubiquitous. Most states lack laws like NEPA, and many states and localities have a rather uneven commitment to planning. Consequently, many sources of environmental impact, including the kind of urban development exemplified by Greenacres, are only partially covered. Even where such laws do exist, their application is often controversial. NEPA continues to generate legislative and administrative complaints, with federal agencies often seeking opportunities to “streamline” its processes. Initiatives to limit the applicability of state environmental assessment requirements are also

263 See, e.g., 33 U.S.C. § 1313(e) (requiring water quality planning); 42 U.S.C. § 7410 (requiring air quality planning). For further examples of comprehensive planning statutes, see supra note 83.

264 See Calvert Cliffs’ Coordinating Comm., Inc. v. U.S. Atomic Energy Comm’n, 449 F.2d 1109, 1114 (D.C. Cir. 1971) (arguing that only through such broad analyses is it “likely that the most intelligent, optimally beneficial decision will ultimately be made”).


267 See FARBER ET AL., supra note 56, at 522 (“NEPA has remained an important pillar of environmental law.”); Dave Owen, Probabilities, Planning Failures, and Environmental Law, 84 TUL. L. REV. 265, 266–67 (2009) (describing environmental law’s extensive planning requirements).


269 NEPA still applies to some development projects, but only to the extent the projects trigger discretionary federal review. See 42 U.S.C. § 4332(2)(C) (2006) (establishing requirements for “major Federal actions”).

quite common. Federal, state, and local governments therefore often debate how much, if at all, to pursue comprehensive planning and environmental review requirements.

Advances in spatial analysis have important implications for these debates. While spatial analysis cannot make the challenges of synoptic analysis disappear, it can expand the realm of the possible, making spatially and temporally broad analyses viable where they previously were unrealistically ambitious. The ability of models to simulate the combined effect of many different emission sources on regional air pollution, or of a variety of land use changes upon water quality, both exemplify that capacity. Similarly, spatial analysis can facilitate the multimedia inquiries that environmental laws like NEPA are supposed to promote. Environmental impact assessment laws provide a rare obligation to address, in a single process leading to a unified document, the impacts of a project on a variety of environmental media, and to consider alternatives to that project. That obligation would mean little if environmental scientists lacked the tools to produce the required predictions, and instead were providing encyclopedic compilations of semi-informed guesswork. But as modeling capabilities increase, retaining a procedural obligation for crossmedia, multi-scalar analysis should become increasingly important.

Those advances do not imply that we should abandon the hedge strategies we have adopted to compensate for pervasive informational shortfalls. Spatial analysis will never make those shortfalls completely disappear, and technology-based standards, adaptive management programs, and all the other ways we now address the informational challenges of environmental regulation will continue to play important roles. But increases in our analytical capacity clearly do mean that the debate over these competing approaches needs to evolve. Comprehensive planning and analysis have produced uneven results, but as we choose among regulatory instruments, past limitations should not prevent us from asking whether technological advances are closing the gaps between our synoptic ambitions and

272 See, e.g., THE WHITE HOUSE COUNCIL ON ENVTL. QUALITY, supra note 104 (proposing ambitious new planning initiatives for marine resource management).
273 See supra notes 210–221 and accompanying text.
274 See supra notes 49–74 and accompanying text.
276 See generally Karkkainen, supra note 265, at 907–08 (promoting adaptive management in contexts of uncertainty); Wendy E. Wagner, The Triumph of Technology-Based Standards, 2000 U. ILL. L. REV. 83 (arguing that technology standards are an important hedge against information deficits).
our practical realities. The role of planning and assessment in environmental law should at least remain stable, and there are many opportunities—particularly at state and local levels—for their greater use.

2. Making Environmental Trading Systems Work

A second recurring challenge of environmental law involves turning the theoretical appeal of environmental trading systems into practical results. Here, as well, spatial analysis offers the potential for supporting more economically efficient and environmentally protective regulatory approaches.

In theory, the appeal of trading systems is elegantly simple: by allowing exchanges across large geographic areas and through time, trading systems should allow regulated actors to efficiently allocate the burdens of compliance while still attaining the desired environmental result. But that theory works best when relatively large actors trade fungible things, and in practice, such simplicity and fungibility are rare. They are particularly rare for the kinds of impacts typically generated by development. Development often involves large numbers of relatively small actors and actions, and for most of the impacts that a project like Greenacres generates—filling wetlands, increasing water withdrawals, or generating air pollution, to provide just a few examples—context matters. Consequently, each trade creates a risk that the balance of benefits and burdens will somehow be skewed, with the imbalance operating to the detriment of environmental protection, and perhaps also creating objectionable distributional impacts.

Market designers can respond to those complications in several ways, but each traditional option is problematic. One option, which provides little assurance of environmental benefit, is to simply live with skewed outcomes. Alternatively, regulators can impose offset ratios, which will compensate for nonfungibility by

---

277 See, e.g., Owen, supra note 267, at 282 n.93 (quoting EPA employees describing successful air quality planning processes).
278 See MANDELKER, supra note 268, § 1:7, at 1-14 (noting that fifteen states have statutes like NEPA).
279 See E. Donald Elliott & Gail Charnley, Toward Bigger Bubbles, 13 F. APPLIED RES. & PUB. POL’Y 4, 48 (1998); Robert Stavins, Market-Based Environmental Policies: What Can We Learn from U.S. Experience (and Related Research)?, in MOVING TO MARKETS IN ENVIRONMENTAL REGULATION, supra note 90, at 19, 20.
280 See Salzman & Ruhl, supra note 90, at 622–30 (describing examples of nonfungibility).
281 See Freyfogle, supra note 100, at 31–33 (describing the importance of real world context to water rights trading); Salzman & Ruhl, supra note 102, at 342 (discussing the challenges of using trading systems to address habitat).
282 See Tietenberg, supra note 90, at 87 (describing the risk of environmental imbalances). See generally Drury et al., supra note 99 (describing the potential for environmental justice problems arising from environmental trading systems).
283 See Salzman & Ruhl, supra note 90, at 612 (arguing that wetlands trading historically involved an excessive tolerance for nonequivalent trades).
imposing a sort of tax on transactions. Thus, for example, the developer who destroys one acre of wetlands might be required to construct four new acres. That approach may provide better environmental protection, but the tax undermines the economic appeal of the market and may deter participation. A third alternative is for regulators to review each trade, making sure it provides sufficient environmental value. That approach may ensure equivalence, but regulators will have much more work to do, and the concomitant unpredictability and higher transaction costs may deter market participation. Incremental review also may squander the efficiency that might come from integrating individual trades into a broader plan. Despite these potential problems, regulators use all of these approaches, and many trading systems still function. But because of the resulting complications, some commentators question whether environmental trading systems offer desirable options outside of a few exceptional circumstances.

Advances in spatial analysis can support a promising alternative approach. Rather than relying on regulated entities to identify their preferred mitigation option, regulators can preapprove a set of mitigation options. Thus, for example, wetlands regulators can identify areas with high restoration potential, or where high-value wetlands are under threat, and can specify those areas as preapproved mitigation zones. Developers then would receive expedited approval for trades involving mitigation in those areas. In theory, all participants benefit. Regulators and the public receive better assurance that individual trades will protect environmental values and will fit into a coherent larger plan, and developers avoid the uncertainty and delay associated with protracted review of each individual transaction.

284 See, e.g., 42 U.S.C. § 7511a(e)(1) (2006) (establishing ratios for ozone emission offsets). This approach also is commonly used where destruction of habitat is being allowed in return for preservation of existing habitat elsewhere. The ratio then compensates—hopefully—for the possibility that the preserved area might have remained intact even without legal protection and that the trade may therefore involve real destruction and superfluous preservation.

285 See Stavins, supra note 279, at 26 (noting this disincentive).

286 Id. at 25 (“[R]equiring prior government approval of individual trades may increase uncertainty and transaction costs, thereby discouraging trading . . . .”). Commentators often cite water rights trading as an area where such transaction costs have resulted in suboptimal trading levels. See, e.g., Stephen N. Bretsen & Peter J. Hill, Transaction Costs and Water Markets: An Anticommons Perspective, in AQUANOMICS: WATER MARKETS AND THE ENVIRONMENT 143, 143–81 (B. Delworth Gardner & Randy T. Simmons eds., 2012).


288 See generally Tietenberg, supra note 90, at 63–64 (describing many uses of environmental trading systems).

289 See Salzman & Ruhl, supra note 102, at 342.
This basic concept is not new,\footnote{See, e.g., Emmons & Olivier Res., Inc., Lino Lakes Special Area Management Plan (SAMP) 3–4 (2010), available at http://www.ricecreekwatershed.govoffice2.com/vertical/Sites/%7BF68A5205-A996-4208-96B5-2C7263C03A9%7D/uploads/Lino_SAMP_Edited_Oct_10.pdf (describing several “Special Area Management Plans,” which the Army Corps of Engineers sometimes uses to implement a similar regulatory approach); Buchsbaum, supra note 71, at 194–95 (describing habitat conservation planning under sections 9 and 10 of the ESA); Royal C. Gardner, Banking on Entrepreneurs: Wetlands, Mitigation Banking, and Takings, 81 Iowa L. Rev. 527, 550–76 (1996) (describing wetlands mitigation banking and its benefits to private landowners and the environment); Julian Conrad Juergensmeyer et al., Transferable Development Rights and Alternatives After Suitum, 30 URB. LAW. 441, 443–54 (1998) (describing transferable development rights, which local governments use to direct development to growth areas and compensate landowners in areas where growth is to be limited); Gregory M. Parkhurst & Jason F. Shogren, Incentive Mechanisms, in 1 The Endangered Species Act at Thirty: Renewing the Conservation Promise 247, 249 (Dale D. Goble et al. eds., 2006) (describing habitat conservation banking); K. Shawn Smallwood et al., Environmental Auditing: Indicators Assessment for Habitat Conservation Plan of Yolo County, California, USA, 22 ENVTL. MGMT. 947 (1998) (using a similar approach to identify target areas for a multispecies habitat conservation plan).} It does not necessarily require spatial analysis.\footnote{In working on one such program, I observed that experienced planners and scientists were comfortable working off paper maps to develop this sort of approach, at least at a single-municipality scale. Interestingly, a generational difference also seemed to be present, with older participants preferring the paper maps and younger participants leaning toward computer-based systems.} But in several ways, spatial analysis can make this alternative approach much more effective. Initially, spatial analysis can inform decisions about the scale and construction of the trading system. By using spatial models, analysts can approximately predict how much development is likely, where it may occur, and what kinds of impacts that development might create.\footnote{See, e.g., Thorne et al., supra note 287, at 939–40 (describing the use of a GIS for this initial step); see also Patrick R. Huber et al., Regional Advance Mitigation Planning: A Pilot Study Integrating Multi-Agency Mitigation Needs and Actions Within a Comprehensive Ecological Framework, 2009 INT’L CONF. ON ECOLOGY & TRANSP. PROC. 221, http://www.icoet.net/ICOET_2009/downloads/proceedings/ICOET09-Proceedings-Session143.pdf (describing a similar process).} That information can help the analysts assess what sort of mitigation is needed, and also how much.\footnote{See, e.g., Thorne et al., supra note 287, at 941–45 (summarizing regional mitigation needs).} Similarly, spatial models can predict the economic value generated by a trading program, which can help regulators set fee or offset schedules and determine how much mitigation work they will be able to do\footnote{I am currently a minor participant in an effort to use this approach to allow locally led vernal pools regulation. See generally ME. VERNAL POOLS, http://www.umaine.edu/vernalpools/ (last visited May 25, 2013) (describing the community-based conservation project, though focusing primarily on earlier phases).} All of this information can help
regulators decide on the scale and mechanics of the trading program. It also can help them determine whether a trading program will be viable at all.

Additionally, spatial analysis can help maximize the return on mitigation purchases. Analysts routinely use spatial analysis to identify areas with multiple features desirable for some use, whether that use is a housing development or a conservation purchase, and that same approach can be adapted to target mitigation efforts. Regulators also can use development models to identify areas where, absent mitigation purchases, development would be likely to occur. That identification could reduce the “additionality” problems that result when public money is expended to protect resources not under any realistic threat. Spatial analysis also can link individual purchases into a coherent larger plan. By analyzing not just the individual value of each protection or restoration project, but also the potential interconnections between different mitigation areas, analysts can create synergy among separate transactions. Similarly, if a central goal of the trading scheme is to keep mitigation in relatively close geographic proximity to the impacted site, modelers can add a geographic-preference criterion to the site selection algorithm.

295 See supra notes 231–240 and accompanying text (discussing land use suitability analyses).
296 See, e.g., Lohse et al., supra note 237, at 469 (describing the combined use of ecological, land use, and hedonic models to identify priority sites); Theobald, supra note 221 (using biological and socioeconomic data to identify ecologically valuable areas likely to be developed).
298 See Thorne et al., supra note 287, at 937. See generally Mikel Gurrutxaga et al., GIS-Based Approach for Incorporating the Connectivity of Ecological Networks into Regional Planning, 18 J. NATURE CONS. 318 (2010) (explaining how GIS can be used to maintain productive connections between conservation areas that otherwise might be isolated).
All of these advances have value within a single-purpose trading system, where the goal is simply to protect wetlands, farmland, or habitat for one particular endangered species. But spatial modeling also raises the possibility of coordinating...
multiple trading systems. The core concept is similar: instead of selecting sites based on one environmental value, the model could select sites based on their value for multiple species, wetlands protection, water supply protection, and recreation. By prioritizing multiple targets, the program could facilitate an integrated response to mitigation requirements set by several different laws. Indeed, as integrated urban modeling becomes increasingly advanced, it may become possible to integrate vehicle miles traveled, air quality, greenhouse gas emissions, and infrastructure costs into the trading exercise. The end result could be a trading system based on preapproved receiving zones, all selected to maximize a broad set of environmental and nonenvironmental goals.

These kinds of integrative trading systems are still quite new, and schemes integrating a full suite of environmental goals do not yet exist. If and when they do, they will suffer from all the standard problems with any approach dependent upon quantitative environmental modeling. Assigning relative weights to the various goals sought by the model also will likely be a challenge. But the potential benefits are significant. Despite their flaws, trading systems already pervade environmental regulation, and improvements that ameliorate some of the externalities and inefficiencies of those existing systems therefore could have immense value, even if the reforms are only partial. These advances also could make trading feasible in many areas where it is not currently used, and that also is a potentially significant gain. Often the practical alternative to the existence of a trading scheme is widespread tolerance of small environmental impacts, as regulators decline to impose prohibitory regulatory approaches they perceive as overly stringent. In these circumstances, the flexibility afforded by a trading

---

300 See, e.g., Huber et al., supra note 292 (describing a multi-agency mitigation strategy); Weber & Allen, supra note 22, at 252 (describing the use of multiple criteria).
301 See, e.g., Weber & Allen, supra note 22, at 240.
302 See, e.g., Huber et al., supra note 292.
304 Systems that attempt to protect multiple species are not so new, and there is evidence that they provide significant benefits over single-species approaches. See, e.g., Jared G. Underwood, Combining Landscape-Level Conservation Planning and Biodiversity Offset Programs: A Case Study, 47 ENVTL. MGMT. 121, 126 (2010) (“Our results show that significantly more conservation has occurred for almost all species of concern in the area with a combined conservation-offset plan.”).
305 See supra notes 169–174 and accompanying text (discussing modeling’s limits).
306 See generally Salzman & Ruhl, supra note 102, at 334 (discussing how the necessity of making apples-to-oranges comparisons creates difficulties for environmental trading systems). One intriguing way of resolving those problems is to allow community participants to vote on weights. See, e.g., TRUST FOR PUBLIC LAND, supra note 235, at 8 (describing the use of public outreach to identify features that would make land relatively important to protect).
307 See, e.g., Owen, supra note 6, at 192–94 (describing potential benefits of trading schemes for critical habitat protection).
program can provide a practicable way of securing some compensation for impacts that otherwise would go unregulated.\textsuperscript{308}

These changes therefore should transform how environmental lawyers evaluate trading systems. Any time such systems are proposed, a key question is whether informational challenges will be manageable or, alternatively, will create a Hobson’s choice between an efficient system that provides no environmental benefit and a protective system that is dysfunctionally cumbersome. In the future, answers to that question will often depend in part on the ability of spatial analysis to support the trading scheme. Where spatial tools allow better planning and oversight, environmental trading schemes should present viable options, sometimes, even, in circumstances where a trading scheme would have been infeasible or unwise twenty years ago.

C. Who Regulates: Toward a More Functional Federalism

Just as advances in spatial analysis will affect which environmental problems we attempt to address and how we address them, they also have implications for some of environmental law’s traditional \textit{who} questions: which entities, within or outside government, should address environmental problems, and how, if at all, should those entities coordinate their efforts? Those questions will likely emerge in a variety of contexts, but one in particular implicates foundational questions about our systems of environmental law. Spatial analysis can help complex systems of overlapping federalism work.

For decades, environmental policymakers and scholars have been obsessed with federalism. The subject looms large in the Supreme Court’s environmental jurisprudence,\textsuperscript{309} pervades political rhetoric,\textsuperscript{310} and generates reams of academic articles. Broadly speaking, the debaters can be divided into two camps. On one side are the “dual federalists,”\textsuperscript{311} whose thinking is most clearly exemplified by several recent decisions of the Supreme Court.\textsuperscript{312} In their view, federalism functions best as a system of boundary rules protecting state and local governments from federal interference and insulating federal prerogatives from state and local

\textsuperscript{308} See, e.g., id. at 193–94.


\textsuperscript{310} See Erin Ryan, \textit{Negotiating Federalism}, 52 B.C. L. REV. 1, 1, 6–7, 28 (2011) (describing recent debates in which federalism assumed prominence).

\textsuperscript{311} This term comes from Robert A. Schapiro, \textit{From Dualist Federalism to Interactive Federalism}, 56 EMORY L.J. 1, 4 (2006).

intermeddling. Importantly for Greenacres, land use planning implicates a particularly important divide. The Supreme Court has often asserted that even as the federal government takes the lead in environmental regulation, land use should remain “a quintessential state and local power.” On the other side are the “interactive” or “dynamic” federalists. Like the dual federalists, they see value in a system with federal, state, and local authority. But they argue that jurisdictional overlap can promote collaboration and communication, leading to more effective use of the “laboratories of democracy” that federalism is supposed to promote.

Many differences of opinion divide these camps, but one potentially important—albeit largely implicit and unexamined—divide involves assessments of the capacity for effective communication among different levels of government. If that capacity is limited, then a system of rigid spheres of authority may make sense. Different levels of government otherwise would stumble across each other’s efforts, with that interference often culminating in federal displacement of state or local discretion, and the isolating boundaries envisioned by the dual federalists could be necessary to preserve meaningful state and local governance. Conversely, if the potential for effective intergovernmental dialogue is high, different levels of governments should be able to communicate their needs and priorities, isolate areas of disagreement, and find common ground. Lawmakers and judges then would not need to worry quite so much

---

313 See Bond, 131 S. Ct. at 2364 (“The federal balance is, in part, an end in itself, to ensure that States function as political entities in their own right.”); United States v. Morrison, 529 U.S. 598, 599 (2000) (“The Constitution requires a distinction between what is truly national and what is truly local . . . .”); Alden, 527 U.S. at 751 (rejecting authority that “would blur . . . the distinct responsibilities of the State and National Governments”).


315 See, e.g., Kirsten H. Engel, Harnessing the Benefits of Dynamic Federalism in Environmental Law, 56 EMORY L.J. 159 (2006); Heather K. Gerken, The Supreme Court 2009 Term, Foreword: Federalism all the Way Down, 124 HARV. L. REV. 4 (2009); Hari M. Osofsky, Diagonal Federalism and Climate Change: Implications for the Obama Administration, 62 ALA. L. REV. 237, 268 (2011); Schapiro, supra note 311. These authors also use several other terms to describe variations on this general theory.

316 Schapiro, supra note 311, at 8–9.

317 Alternatively, consolidating authority within a single, unitary government might make sense, but that possibility is so politically unthinkable that this Article does not consider it.

318 See generally Bond, 131 S. Ct. at 2364–65 (arguing that federalism protects liberty by protecting spheres of state and local primacy).

319 To put the point slightly differently, if we are choosing between a federalism of voice and a federalism of exit, we need to know about the effectiveness of the means of communication. Compare Gerken, supra note 315 (promoting a federalism of voice), with
about allowing federal or state environmental programs to affect land use authority and other state and local prerogatives.

In several ways, spatial analysis can help answer that question, and a particularly illustrative set of examples involves processes sometimes called “alternative futures modeling” or “scenario planning.” These processes use spatial analysis to model future land use scenarios. Modelers can develop the scenarios in a variety of ways, including working with people in the affected area to define scenarios they think are plausible or desirable. The modelers then explore the implications of those scenarios for a variety of potential outputs, including development patterns, water quality, biodiversity, and, potentially, economic impacts like costs to state and local government and private property values. Based on the initial results, they also can develop new scenarios or work backward from desired future outcomes to recommended present policy approaches. The end result is generally a series of detailed maps that depict plausible alternative futures for the modeled area, as well as charts and graphs explaining differences between the alternatives.

In several ways, these processes can facilitate the kind of interjurisdictional coordination upon which dynamic federalism theories implicitly rely. Initially, they allow evaluation of the combined implications of a variety of current trends


Land use suitability analysis, which Part IIIB describes at length, also has important implications for federalism, for it provides an effective mechanism for communicating local, state, and federal regulatory constraints.

See, e.g., CARL STEINITZ ET AL., ALTERNATIVE FUTURES FOR CHANGING LANDSCAPES: THE UPPER SAN PEDRO RIVER BASIN IN ARIZONA AND SONORA (2003); Baker et al., supra note 186; Tony Prato et al., Evaluating Alternative Economic Growth Rates and Land Use Policies for Flathead County, Montana, 83 LANDSCAPE & URB. PLAN 327 (2007). Researchers also sometimes use terms like “scenario planning” to describe these types of analyses. For warnings about limitations of alternative futures analysis, see Pontius & Neeti, supra note 216, at 39.

See Pontius & Neeti, supra note 216, at 39.

See, e.g., Baker et al., supra note 186, at 315 (“The future landscapes are designed with stakeholder input to illustrate major strategic choices.”).

See, e.g., id. at 316 (evaluating implications for water availability, riparian habitats, and terrestrial wildlife); Chakroborty, supra note 179, at 396–97 (describing a scenario-modeling exercise for Maryland).

See Baker et al., supra note 186, at 315 (“As stakeholders see results for the initial set of alternative futures, it may lead to new ideas or compromise positions that warrant design of additional future scenarios or analysis of additional endpoints.”).

and policies. That review may reveal future conflicts or opportunities that might never have become apparent through individual plan-by-plan or project-by-project studies. Futures modeling also allows participants to explore the potential implications of—and perhaps, to reconsider—their assumptions. By visually depicting tradeoffs between competing goals, modeling creates an opportunity to consider whether rigid adherence to their prior preferences will produce a landscape participants want to live in. For similar reasons, the process of developing a model and a set of maps provides an opportunity to move discussions out of the realm of ideological abstraction. Scenario maps grab attention and convey information in a language readily understood by many people, making a possible future seem much more real. Sometimes, of course, the maps can make the future seem too real, and participants may forget that the maps are explorations of possible futures, not reliable predictions. But if the modelers are attentive to that possibility, the possibility of new insight can significantly outweigh the risk of new misconceptions.

Facilitating a more inclusive and constructive discussion may be valuable, but if that discussion only reveals intractable conflict, the exercise ultimately will have modest value. Often, however, modeling reveals otherwise unseen options that can

---

327 See, e.g., Baker et al., supra note 186, at 319 (describing a scenario that “provided a unique opportunity to examine [the] joint implications” of a variety of land use plans).
328 See, e.g., id. (describing surprising results).
329 See, e.g., Schmitt Olabisi et al., supra note 188, at 2693–94 (discussing an exercise that confronted participants with unexpected implications of a commitment to local energy production).
330 See Hulse et al., supra note 177, at 339 (explaining how mapping exercises can facilitate constructive dialogue).
332 See generally Wagner et al., supra note 29 (warning of this risk).
333 See Jonathan R. Thompson et al., Scenario Studies as a Synthetic and Integrative Research Activity for Long-Term Ecological Research, 62 BIOSCIENCE 367, 374 (2012) (explaining that a review of multiple scenario planning studies demonstrates “how more interactive engagement can enhance the interest and ownership for the challenges as well as potential solutions across different stakeholder groups”). Indeed, a similar risk exists even without computerized simulation models. Most people likely have prior assumptions about the future, and those future assumptions may be even more erroneous, and significantly less examined, than the predictions produced by a model. Discussions about a model can provide a valuable opportunity to expose and examine those assumptions. See generally ADAPTIVE ENVIRONMENTAL ASSESSMENT AND MANAGEMENT 95–106 (C.S. Holling ed., 1978) (explaining the benefits of ongoing discussions about iterative processes of building and running models).
ameliorate interjurisdictional tensions. 334 Even environmentally sensitive regions often have many areas where development can occur while causing relatively modest environmental impacts. 335 Similarly, if policymakers act proactively, many mechanisms, including protecting habitat corridors, preserving riparian buffers, and promoting cluster development and other compact patterns of growth, can help integrate development into a landscape while retaining important ecological functions. 336 Prioritizing development in some areas and limiting it in others obviously has potential implications for property values, but advance planning gives communities opportunities to set up financial mechanisms, like environmental trading systems or transferable development rights, to ameliorate those impacts. 337 Alternatively, confronting the future implications of unrestrained development may lead people to conclude that some uncompensated limitation on property use is an appropriate contribution toward maintaining a community’s identity and quality of life. 338 In short, in a variety of ways, modeling alternative futures can help people achieve the goals of environmental law while still preserving ample local discretion and community autonomy.

To be clear, this Article’s claim is not that the emergence of spatial analysis will generate universal acceptance of dynamic federalism. There are many other reasons—including generalized hostility to regulatory governance and ideological opposition to federal authority—for continued interest in federalism’s more boundary-based forms, and the prospect of effective intergovernmental collaboration will not make that interest disappear. 339 Similarly, spatial analysis will not always reveal options that meet everyone’s needs. Some conflicts really are intractable. 340 But spatial analysis can communicate federal, state, and local goals, explore compatibilities between those goals, and cabin conflict to more manageable and discrete zones. That capacity should give at least a moment’s pause to lawmakers and judges who assume that rigid limits on federal or state

334 See, e.g., Prato et al., supra note 321, at 336–37 (concluding that a “moderately restrictive” land use policy could accommodate future growth).

335 See, e.g., McCloskey et al., supra note 181, at 198 (finding abundant opportunities for conflict-free development); John Van Sickle et al., Projecting the Biological Condition of Streams Under Alternative Scenarios of Human Land Use, 14 ECOLOGICAL APPLICATIONS 368, 378 (2004) (concluding that a “Conservation scenario” would allow environmental improvements even as the Willamette Valley’s population doubles).

336 See U.S. ENVTL. PROT. AGENCY, supra note 40, at 35–79 (describing smart growth mechanisms).

337 See supra Part III.B.2. (discussing trading systems); supra note 290 and accompanying text (discussing transferable development rights).

338 See generally ERIC FREYFOGLE, ON PRIVATE PROPERTY: FINDING COMMON GROUND ON THE OWNERSHIP OF LAND (2007) (arguing that because property derives its value from human communities, those communities should be able to use democratic processes to adjust property rights).

339 It is very difficult to imagine, for example, that the venom currently directed at federal greenhouse gas regulation would disappear if those controls derived from the states.

340 See, e.g., Albert, supra note 184, at 36–39, 41–42 (describing the minimal influence of one alternative futures modeling project).
environmental regulation are necessary to protect spheres of state or local autonomy.

IV. SPATIAL ANALYSIS AND ENVIRONMENTAL LAW RESEARCH METHODOLOGIES

So far, this Article has focused on implications of spatial analysis for our understanding of environmental problems and for the design and implementation of legal solutions. Obviously that discussion should inform environmental law research, for the structure and application of environmental law are central research subjects for many academic inquiries. But the implications of spatial analysis extend beyond the subjects of environmental law research and also implicate its methodologies. This Part explores how. The discussion here is illustrative and preliminary, but even a few examples demonstrate how quantitative spatial analysis can change environmental law research. In turn, those research advances could facilitate improvements in the structure and application of environmental law.

For decades, assessing how environmental law changes real-world outcomes has often been difficult. Many environmental laws generate only partial implementation, and determining the extent of the gaps between the law on the books and the law in practice is not always easy. Even if something approaching full compliance exists, the consequences of that compliance can be difficult to discern. Determining the environmental benefits of NEPA, for example, is complicated by the attenuated causal chain between required actions and actual environmental outcomes. Environmental laws also often generate unintended consequences, and the nature and extent of those consequences is similarly difficult to predict. For all of these reasons, debate is still robust about whether and how some of our most familiar environmental laws provide environmental protection.

Spatial analysis gives environmental law researchers new tools to address these questions. For example, by comparing development patterns in areas subject to a particular law to development patterns in exempted areas, researchers can assess how that law actually affects outcomes. Similarly, longitudinal studies,
which examine development patterns before and after the imposition of some environmental constraint, could help assess what on-the-ground impact laws actually have. Each type of study involves complications; most importantly, the complexity of human and environmental systems assures a potential overabundance of confounding variables. But with the increased availability of spatial data sets and the possibility of using linear regressions to minimize statistical noise, opportunities for new insight exist.

This sort of research is not exactly new. For decades, economists have been using both theoretical models and actual datasets to test the implications of environmental laws.346 But such work rarely appears in legal journals, and even when it does, the authors usually are not lawyers.347 That is a significant absence. While an economist’s perspective has obvious value, there are ways in which lawyers could contribute to this sort of work. Environmental lawyers may not be trained in quantitative analysis or GIS, but they are taught to understand, at least at a qualitative level, how particular regulatory provisions fit within broader environmental law systems, how environmental law evolves and changes, what roles environmental law assigns to different actors, and how different institutions tend to respond to their roles. That legal perspective could help interdisciplinary research teams identify important research questions, develop hypotheses, flag potentially confounding variables, and interpret results.

The rise of environmental modeling creates similar opportunities for engagement. One primary goal of many environmental modelers is to understand and simulate the feedback loops between human and natural systems.348 Those feedback loops are partly mediated by economics, for economic incentives play a significant role in determining human actions. Consequently, and not surprisingly, economists have engaged the process of modeling land use change.349 But the relationships between human and environmental systems are also heavily mediated by law. Though they have rarely played this role, lawyers could offer important insights about how legal rules might generate environmental consequences and about how environmental change generates legal responses.350 In some circumstances, that legal perspective should help modelers build better and more useful models. In others, environmental lawyers’ insight may be that the dynamics are too complicated and unpredictable to model. But that also can be an important contribution, for it can send the modelers on to more useful endeavors.

These possibilities support a broader point about legal research. In recent years, legal research in many fields has moved toward greater reliance on quantitative analysis of empirical data and broader integration with other academic

346 See supra note 26 (citing studies).
347 See id. (citing some exceptions).
348 See, e.g., Alberti, supra note 21, at 4.
349 See, e.g., Irwin & Geoghegan, supra note 248, at 8 (explaining economic land use change models).
350 See generally Lazarus, supra note 7 (exploring the dynamics that spurred the creation of American environmental law).
fields. In some sense, a move toward quantitative spatial analysis would simply represent a continuation of that trend. But the shift has been gradual and sometimes controversial, with critics arguing, among other complaints, that both changes threaten to sidetrack legal research into a realm of impractical abstraction. Clearly a movement toward quantitative spatial analysis could create that same threat, because sometimes models are so data intensive and complex that they are unworkable or so abstract that they are meaningless. But the present uses of spatial analysis suggest that here, at least, the critiques of quantitative and interdisciplinary legal research will often miss the mark. Helping federal, state, and local governments balance economics, environmental protection, and autonomy is a highly practical goal. If, in working toward that goal, legal researchers can help achieve a better understanding of some of the core challenges of environmental law, the effort will be well worthwhile.

**CONCLUSION**

Since the 1960s, when ecology emerged as a scientific discipline, law has never been the same. The core concepts of ecology—its focus on interconnectedness, interdependence, and environmental fragility—energized an environmental movement and led to a generation of environmental laws. Similarly, the emergences of law and economics, quantitative risk analysis, and complexity theory all have had profound implications for the practice and theory of environmental law. Environmental law is inextricably, if sometimes...
uncomfortably, intertwined with environmental science, and when environmental science evolves, legal thinkers usually ask whether law should evolve too.356

The emergence of quantitative spatial analysis has just begun to spur a similar reaction.357 The products of spatial analysis often form the evidentiary basis for decisions required by environmental laws, and spatial analysts often work to fulfill environmental law’s informational demands. But while environmental law has influenced spatial analysis, the feedback loop has not closed. Advances in spatial analysis have not led to any significant revisions to the structure, practice, or theory of environmental law. The time for greater engagement has come.

---

356 See, e.g., Ruhl, supra note 355; Tarlock, supra note 195, at 1134–44 (considering potential reactions to the decline of ecology’s equilibrium paradigm).

357 See, e.g., Boyd, supra note 26 (considering the implications of spatial data for global climate change policy and forest management).
DEFINING PERSECUTION

Scott Rempell*

Abstract

Persecution is the core concept of asylum and refugee protection. Although thousands—if not tens of thousands—of decisions hinge on its meaning, a consistent definition is yet to emerge. Unmoored to any unified understanding of the term, immigration agencies and federal courts of appeals continue to articulate many different conceptions of persecution—conceptions that lack internal consistency and a coherent analytical foundation. Legal scholars have not attempted to aid adjudicators’ understanding of persecution because, by and large, scholars do not believe that a unified definition is possible. Meanwhile, the divergent definitions and understandings of persecution continue to produce unfair results for those seeking asylum, as asylum applicants receive disparate outcomes despite presenting claims based on similar situations. This Article challenges the conventional wisdom that persecution defies unified meaning. It provides a comprehensive assessment of persecution’s central underpinnings to isolate the three pillars that represent persecution’s fundamental core: harm, severity, and legitimacy. At the same time, this Article critiques a number of false dichotomies and shaky definitions that have troubled and obscured the persecution definition up to this point. Based on the analyzed core aspects of persecution and the elimination of erroneously included definitional components, this Article proposes that decisionmakers define persecution as “the illegitimate infliction of sufficiently severe harm.” Because it is grounded in an examination of persecution’s true underpinnings, the proposed definition will aid courts in their review of asylum claims and help administrators render consistent decisions. The stakes are simply too high, and the issue too prevalent, to let decades of abdication continue in any effort to form a unified definition.

I. INTRODUCTION

Persecution is the “fundamental concept at the core of the refugee definition,”¹ yet its meaning remains largely undefined. The vagueness is at least partially intentional. Both the Immigration and Nationality Act (INA) and the

* © 2013 Scott Rempell. Associate Professor of Law, South Texas College of Law. For their helpful comments and suggestions, the author wishes to thank Dru Stevenson, Susan Green, and Amanda Harmon Cooley.

immigration regulations purposefully omit any explanation of the meaning of persecution, thus leaving the task to the Board of Immigration Appeals (Board) and the federal courts of appeals. At the agency level, the Board has essentially sidestepped the question for decades. The federal courts have recognized persecution as “ill-defined,” but have shied away from formulating any unified definition. The absence of a unified definition is partially based on principles of agency deference, but it is also the result of the way courts perceive the concept of persecution. Indeed, a persecution definition has been described as everything from “elusive” to “protean.” When the courts have attempted to provide a guiding framework for assessing persecution, the result has been an amalgamation of vague principles and general statements that are neither consistent nor entirely accurate. At times, certain courts have described the threshold for persecution as merely offensive conduct, while other courts require “extreme” conduct. Some adjudicators believe one instance of harm is sufficient to establish persecution, while others require systematic abuse. The inconsistencies go on, and so does the confusion.

Refugee scholarship is similarly devoid of major efforts to define persecution, absent very limited—but notable—exceptions. Two main reasons likely account

---

3 See 8 U.S.C. § 1252 (providing the federal courts of appeals with jurisdiction to review decisions rendered by the Board); 8 C.F.R. § 1003.1 (discussing the Board’s scope of authority).
4 See Sahi v. Gonzales, 416 F.3d 587, 588 (7th Cir. 2005) (stating that the court could not find “a case in which the [Board] has defined ‘persecution’”). The Board has examined the meaning of persecution in a variety of contexts. As analyzed infra Parts III–V, however, the Board’s discussions are neither consistent nor complete.
5 Haile v. Gonzales, 421 F.3d 493, 496 (7th Cir. 2005) (citing Sahi, 416 F.3d at 588–89).
7 Pathmakanthan v. Holder, 612 F.3d 618, 622 (7th Cir. 2010); Aguilar-Solis v. INS, 168 F.3d 565, 570 (1st Cir. 1999); Balazoski v. INS, 932 F.2d 638, 641 (7th Cir. 1991).
8 Bocova v. Gonzales, 412 F.3d 257, 263 (1st Cir. 2005); see also Stserba v. Holder, 646 F.3d 964, 972 (6th Cir. 2011) (labeling persecution “ambiguous”); Etugh v. INS, 921 F.2d 36, 39 (3d Cir. 1990) (stating that persecution “can only be given concrete meaning through a process of case-by-case adjudication”).
9 Compare Chaib v. Ashcroft, 397 F.3d 1273, 1277 (10th Cir. 2005) (stating conduct must be offensive), and Kovac v. INS, 407 F.2d 102, 107 (9th Cir. 1969) (same), with Mei Fun Wong v. Holder, 633 F.3d 64, 71–72 (2d Cir. 2011) (stating extreme conduct is required), and Fatin v. INS, 12 F.3d 1233, 1240 & n.10 (3d Cir. 1993) (same).
11 See JAMES C. HATHAWAY, THE LAW OF REFUGEE STATUS 101–05 (1991); see also Deborah E. Anker, Refugee Status and Violence Against Women in the “Domestic”
for this dearth of attention. First, the literature has tended to focus on the other prongs of the refugee definition. In addition to showing a well-founded fear of persecution, applicants cannot demonstrate that they qualify as refugees without establishing a nexus between the persecutory actions and one of the five protected grounds—commonly referred to as the “nexus requirement”—and that the harm was (or will be) perpetrated by the government or private actors the government is unable or unwilling to control. The nexus requirement serves as a major limiting principle to a successful claim. The perceived unfairness of the requisite causal connection between the persecutor’s motive and the resulting harm has been a main focus of scholarship. The role of the State, as well, has spawned much literature, particularly in the area of States’ inability to avert or punish the actions of private actors.

The second likely reason for the absence of scholarship on persecution dovetails with the reasons why Congress, immigration agencies, and the judiciary have avoided a unifying definition—namely, the seemingly transient nature of the term. As one scholar has noted, the lack of a coherent definition may be

_Sphere: The Non-State Actor Question_, 15 GEO. IMMIGR. L.J. 391, 400 (2001) (explaining that the Canadian Supreme Court has adopted Hathaway’s formulation). Hathaway defines persecution as the “sustained or systemic violation of basic human rights demonstrative of a failure of state protection.” HATHAWAY, supra, at 104–05. Because this Article focuses on U.S. refugee law, and for additional reasons explored in Part III.B–C, the definition advanced in this Article diverges from Hathaway’s in several respects.

12 This Article uses the word “applicant” to apply to those seeking refugee relief, rather than the more controversial “alien” descriptor applied under the INA. See 8 U.S.C. § 1101(a)(3) (2006).


15 See, e.g., HATHAWAY, supra note 11, at 105–06, 125–33; Anker, supra note 11, at 392–93; Elsa M. Bullard, Note, _Insufficient Government Protection: The Inescapable Element in Domestic Violence Asylum Cases_, 95 MINN. L. REV. 1867 (2011).

appropriate, “given the endless variety of situations the term might cover.” Other scholars have argued, “There has never been a succinct, definitive definition of ‘persecution,’ because the nature of persecution and our understanding of it keep changing.”

The hesitation to define persecution is certainly understandable given the large variety of inflicted harms that may fall under the persecution rubric. Any attempt to define the term must be flexible enough to account for the unfortunate “inventiveness of humanity to think up new ways of persecuting fellow [citizens].” Persecution’s elusive nature might provide a caution to its formulation, but adjudicators have pleaded for such a formulation. Indeed, the U.S. Supreme Court recently noted that a “more comprehensive definition” of persecution would be beneficial. While a unified definition would benefit adjudicators, it is critical for the tens of thousands of applicants who annually seek asylum protection in the United States alone; the consequences of adjudicators’ decisions may have life-or-death ramifications for applicants.

This Article is a first step toward forging a unified definition of persecution, and it will do so through the prism of U.S. refugee law. Part II will begin by addressing harm, which is at the core of persecution. As one commentator has noted, “The central issue remains that of risk of harm amounting to persecution.” It is much easier to identify that harm should be at the core of persecution assessments than to catalogue all the different types of harm that could conceivably factor into the persecution inquiry. Certainly a punch to the face is easily recognizable as harm, as well as prolonged confinement or the seizure of assets. However, there are less tangible manifestations of harm that can result from assailants’ actions, such as psychological harm. For persecution determinations to

---

21 See Negusie v. Holder, 555 U.S. 511, 524 (2009); see also Marquez v. INS, 105 F.3d 374, 379 (7th Cir. 1997) (describing the “largely ad hoc” approach to assessing persecution claims as “unfortunate[ly]”).
25 These psychological harms may be the only result of an assailant’s actions or an inexorable byproduct of more tangible harms. As discussed further infra Part II.D, psychological harm would be the sole repercussion of, for example, a death threat, while
accurately reflect the level of harm experienced by the applicant (or feared prospectively), all harms must be considered. Adjudicators cannot consider the significance of a type of harm that they fail to recognize in the first place. Consequently, Part II provides a taxonomy of each type of harm, obvious or not, that forms the foundation for assessing whether conduct rises to the level of persecution.

A harm taxonomy cannot be complete without a discussion of whether and how human rights should factor into the harm inquiry. The use of human rights to gauge whether harm exists represents one of the most significant disconnects between international refugee law and the U.S. system. In European countries and other jurisdictions, human rights provide the prism for assessing the existence and extent of harm. In the United States, by contrast, adjudicators very rarely mention human rights in their persecution assessments; they factor the significance of a human rights violation into their decisionmaking process even less frequently. Commentators regularly fault the U.S. approach for failing to interpret its refugee protection obligations in a manner that comports with the framework of ratified international instruments.

Even assuming arguendo that these commentators’ criticisms are warranted, the United States is not going to uproot entirely the adjudication methodologies it applies to refugee and asylum claims. More importantly for purposes of developing a harm taxonomy, the question becomes whether the United States’ failure to assess persecution claims from a human rights vantage point could lead adjudicators to overlook the existence of any form of harm that the descriptive categories of harm do not otherwise identify. If so, then the omission of such harms would necessarily render a harm taxonomy incomplete. This Article suggests that such gaps do exist and, consequently, an accurate assessment of harm requires adjudicators to view human rights infringements as a separate category of harm in certain instances. Specifically, a rights-based approach to assessing harm can help to fill in the gaps when the descriptive categories fail to appreciate the significance of the wrongful act. To illustrate, assume a State informs an asylum applicant that he cannot practice his religion, but the penalty for doing so is a very small fine. While a descriptive assessment of the harm could filter the repercussions into relevant categories such as small economic consequences and severe physical abuse may cause mental suffering in addition to other distinct forms of harm.


28 For examples of some of the anomalous cases that expressly mention the rights at issue, see Chen v. Holder, 604 F.3d 324, 334 (7th Cir. 2010); Nuru v. Gonzales, 404 F.3d 1207, 1222–23 (9th Cir. 2005).

29 See Hathaway & Cusick, supra note 23, at 481–84.
some form of mental distress, the nature of the harm would not be fully appreciated without accounting for the prohibition of an action that is so fundamental to one’s identity and inner-most belief structure.

Demonstrating that harm has occurred is not, in and of itself, sufficient to establish that the conduct in question amounts to persecution. Rather, the harm experienced must meet the requisite level of severity. A survey of the literature and the thousands of court cases that have measured the severity of persecution yield three distinct frameworks used to decide if harm is severe enough to constitute persecution. The bones of these frameworks and the problems associated with them are the initial focus of Part III. Under one framework, coined the minimal harm model, refugee status is not about the severity of harm; rather, protection largely depends on the persecutor’s motive for inflicting the harm and whether the State is responsible for the harm or fails to stop it. Such deemphasis, of course, renders superfluous an entire prong of the refugee definition and merely highlights the importance of developing a severity framework that comports with the objectives of applicable refugee conventions and national implementation frameworks.

The second framework—the cumulative harm model—hypothesizes that adjudicators should assess the severity of harm by aggregating together each experienced harm. Aggregating harm in such a way, however, does not illuminate the entire contextual canvass. Instead, it is more analogous to a severity point system, where, by way of illustration, a slap is one point, a punch yields two points, a short detention adds three points, and so on. It is easy to see why a process implicitly grounded in a point aggregation system would appeal to an adjudicator. The underlying assumption here is that a higher point total correlates to greater overall harm, which, in turn, renders it more likely that an applicant experienced (or will experience) harm that rises to the level of persecution. While this may be true in many instances, mere aggregation overlooks the context of the events in a manner that can skew the true extent of harm.


31 See, e.g., Mei Fun Wong v. Holder, 633 F.3d 64, 71–72 (2d Cir. 2011); Gjolaj v. Bureau of Citizenship & Immigration Servs., 468 F.3d 1080, 1084 (9th Cir. 2005).

32 For examples of cases applying components of the cumulative harm model, see Manzur v. U.S. Department of Homeland Security, 494 F.3d 281, 290 (2d Cir. 2007); Krotova v. Gonzales, 416 F.3d 1080, 1084 (9th Cir. 2005).

33 See Restrepo-Velasquez v. U.S. Att’y Gen., 194 F. App’x 865, 867 (11th Cir. 2006); Siswanto v. Gonzales, 177 F. App’x 523, 524 (8th Cir. 2006).
The final framework is the systematic abuse model, which posits that harm is only severe enough if meted out in a systematic way. This theory highlights the particularly impactful consequences of repeated abuse by a specific person, group, or entity. Nevertheless, the model only provides protection for an entirely too narrow subset of harm. Adjudicators should not discount the harm an applicant experienced simply because multiple parties delved it out over a period of time for potentially different reasons. Indeed, if each of the reasons were on account of a protected ground (which is what the refugee definition requires), then the nefarious actions should all bear potential relevance to the persecution assessment. The systematic nature of abuse may impact the severity of harm, but not whether adjudicators can even assess the harm that occurred.

After assessing these three currently employed frameworks, Part III proposes a new model for measuring the severity of harm: the continuous suffering model. Under this model, adjudicators would assess the severity of harm by looking at the applicants' continuous experiences in their home countries, and determine an applicant was persecuted when the overall experience has made living in the country sufficiently intolerable. This proposed model goes beyond the cumulative harm model’s emphasis on the severity of isolated instances of harm. Limited instances of acute abuse may indeed be sufficient to establish persecution, but perhaps not. The circumstances surrounding specific instances of harm may limit the harm’s impact. Thus, even if assailants severely beat an applicant on two occasions, the beatings’ impact on the applicant’s overall experience would diminish if there were no chance a third beating would take place and the assailants did not threaten the applicant with additional harm.

Conversely, an applicant may have experienced more numerous incidents that were not as severe, such as a slap to the face. Rather than thinking of each slap as one “point” in a severity aggregation equation, the continuous suffering model seeks to measure all the different harms these instances cause the applicant. Thus, for example, depending on the surrounding circumstances, the slaps may cause the applicant to live in a perpetual state of fear, alter his daily routine, worry about the safety of his family, constantly wonder when the next incident will occur, and fear that the confrontations will grow more severe. Information supporting the likelihood of escalation, such as evidence that the alleged persecutor intensified the

---

34 See, e.g., Pakasi v. Holder, 577 F.3d 44, 47–48 (1st Cir. 2009); Bocova v. Gonzales, 412 F.3d 257, 263 (1st Cir. 2005).


36 The ultimate inquiry is prospective; it focuses on whether the applicant fears persecution if returned to his home country. An applicant’s past experiences can be relevant in two ways. First, they can serve as circumstantial evidence that future harm is likely to ensue. See, e.g., Boykov v. INS, 109 F.3d 413, 416 (7th Cir. 1997) (stating that past “unfulfilled threats” may be “indicative of the danger of future persecution”). Second, and more beneficial for the applicant, past experiences that amount to persecution and satisfy the other prongs of the refugee definition provide the applicant with a rebuttable presumption that he will be persecuted if deported. See 8 C.F.R. §§ 1208.13(b)(1), 1208.16(b)(1)(i) (2013).
violent acts in confrontations with other parties, could serve as proof that the actions directed at the applicant caused more severe harm. The goal of this Article is not to judge whether these particular effects are sufficiently severe. Rather, it is simply to proffer a framework for measuring harm, assessed in all its forms on the basis of the discussed harm taxonomy, that accurately accounts for the true level of harm endured or feared after removal.

Finally, Part III will discuss how adjudicators should factor human rights abuses into their severity assessments. As commentators have noted, “[T]he concept of being persecuted in the refugee definition is not the equivalent of being exposed to any human rights violation.” Beyond a general recognition of this distinction, there is no consensus on how adjudicators should assess the severity of a rights violation. Most attempts assert that adjudicators should measure severity based on the fundamental nature of the infringed right. Although rights assessments contain an inherent level of subjectivity, this Article seeks to make the inquiry more objective by advocating that adjudicators start by distinguishing rights violations that are per se sufficiently severe from those that are not. For example, there is a recognizable distinction between a government infringing on an applicant’s right to marry and a government violating its obligation to refrain from torturing the applicant, even though both circumstances represent rights infringements. By definition, a persecutor cannot harm a victim in a manner that is severe enough to constitute torture but not persecution.

37 An adjudicator would not necessarily credit any deleterious consequence the applicant claims to have experienced or idiosyncratic reactions to harm that skews its relative severity. Rather, the applicant’s claim must be subjectively genuine and objectively reasonable. See Lie v. Ashcroft, 396 F.3d 530, 535–36 (3d Cir. 2005); Acewicz v. INS, 984 F.2d 1056, 1061 (9th Cir. 1993).

38 This Article does not claim that all cases that assess persecution claims fail to account for the context surrounding the incidents experienced by the applicant. Indeed, in the myriad cases that have addressed the issue, a number have made purposeful efforts to better assess the true extent of harm. See, e.g., Sulaiman v. Holder, 371 F. App’x 224, 226–27 (2d Cir. 2010); Manzur v. U.S. Dep’t of Homeland Sec., 494 F.3d 281, 290 (2d Cir. 2007). No attempt has been made, however, to incorporate these contextual considerations into a comprehensive catalogue of harm that serves as a normative framework for evaluating harm—certainly not in a universal way that would provide fairness and consistency to a process often chastised for lacking both qualities. See Ramji-Nogales et al., supra note 18, at 296.


40 See GUY S. GOODWIN-GILL & JANE MCADEM, THE REFUGEE IN INTERNATIONAL LAW 133 (3d ed. 2007); HATHAWAY, supra note 11, at 109–12.

Conversely, infringements of the right to marry encompass a vast array of divergent conduct. A State may permit its citizens to marry, but prohibit them from marrying before the age of twenty. An applicant who claims persecution by the government solely because it prohibited the applicant’s marriage at the age of nineteen should not be able to establish that the government’s prohibition constitutes persecution per se. Once adjudicators identify rights violations that are not persecution per se, they can then assess the violations’ severity within the context of the applicant’s experience as a whole, taking into account that a right’s infringement represents at least part of the harm the applicant endured.

Even if an applicant underwent sufficiently severe harm, an adjudicator may still determine that the applicant was not persecuted. For instance, although lifetime confinement is incredibly severe harm, this harm may be the consequence of a horrible act committed by the applicant that the State legitimately wants to punish. Thus, in addition to showing harm of sufficient severity, establishing persecutory conduct also requires an assessment of the legitimacy of the alleged persecutor’s actions. Part IV addresses the legitimacy component, analyzing the distinct circumstances when the legitimate actions of governments and private actors derail an applicant’s persecution claim.

To this point, this Article has discussed the three main components of persecution: harm, severity, and legitimacy. But this is not the end of the analysis. It is also necessary to consider what should not factor into the persecution assessment, and Part V seeks to do so. Adjudicators factor a number of irrelevant considerations into their persecution assessments. Part V demonstrates why these additional inquiries do not bear relevance to the persecution assessment. Irrelevant factors skew the persecution definition and detract attention from persecution’s core meaning. Additionally, and equally consequential, jurisdictions apply these irrelevant considerations to different extents and in divergent ways.
which creates inconsistent results for applicants. Asylum and refugee protection should not depend on the definitional idiosyncrasies of a particular jurisdiction.

Part VI draws on the previous analysis to propose a definition of persecution. The definition itself is purposefully straightforward and minimal: the illegitimate infliction of sufficiently severe harm. It includes the three major components but omits references to each of the added components that this Article has shown do not belong. Equally important to the definition itself is the evaluative methodology this Article advocates for assessing each component. Persecution assessments will always involve some level of subjectivity. Nevertheless, as noted in Part VII’s concluding thoughts, adjudicators’ judgment calls must be grounded in an accurate assessment of what it means to be persecuted.

II. TAXONOMY OF HARM

The extent of harm or prospective suffering cannot be accurately valued without an understanding of the broad spectrum of harms that persecutors can inflict on their victims. If persecution is at the core of refugee protection, then harm is at the core of persecution. Along with an inability to seek redress from the government, harm provides the impetus for refugees to flee their home countries or fear returning, and it is harm that refugee law seeks to prevent in certain contexts. As noted above, refugee law only protects victims of harm when the harm is sufficiently tethered to a protected ground and the State perpetuates the harm or abdicates in its responsibility to control the actions of private actors. See 8 U.S.C. § 1101(a)(42)(A) (2006); Gao v. Ashcroft, 299 F.3d 266, 272 (3d Cir. 2002).

The types of harm a victim may suffer include several with minimal descriptive ambiguity. This Part will first identify three categories of harm that have been well recognized as actionable. These categories encompass physical harms, confinement, searches and seizures, and economic harms. Harm, however, is not limited to highly tangible consequences. Psychological harms represent a broad spectrum of suffering that has defied comprehensive catalogue, despite the fact that mental suffering represents such a vast array of repercussions present in nearly every refugee inquiry. After reviewing psychological harm, this Part will conclude with a discussion of the role that human rights law plays in an understanding and evaluation of harm.

46 Compare Faddoul v. INS, 37 F.3d 185, 188 (5th Cir. 1994) (requiring that the persecutor intend to punish the victim), with Pitcherskaia v. INS, 118 F.3d 641, 647 (9th Cir. 1997) (explaining why an intent to punish requirement is not applicable in the Ninth Circuit).

47 As noted above, refugee law only protects victims of harm when the harm is sufficiently tethered to a protected ground and the State perpetuates the harm or abdicates in its responsibility to control the actions of private actors. See 8 U.S.C. § 1101(a)(42)(A) (2006); Gao v. Ashcroft, 299 F.3d 266, 272 (3d Cir. 2002).

48 See infra Part II.A.

49 See infra Part II.B.

50 See infra Part II.C.

51 See infra Part II.D.

52 See infra Part II.E.
It is worth noting that a single incident can encompass harms that fall into more than one of the identified categories of harm. In fact, such overlap is often the case. Thus, for example, psychological harm often accompanies physical punishment, and prolonged detention may lead to economic losses. Moreover, the existence of certain types of harm may themselves alter the severity of other attendant harms. The psychological consequences of proffered threats, for example, can intensify when the victim is detained because the victim’s confinement magnifies the persecutor’s ability to carry out the threat. This categorical overlap does not minimize the justifications for breaking down harm into separate categories. To the contrary, it highlights the importance of doing so. For it is not until harm is understood in all its forms that decisionmakers can assess and quantify the consequential level of suffering caused by the actions in question. Without a comprehensive understanding of harm, a collection of different harms cannot be viewed as anything other than “adverse treatment” or other nondescriptive labels that fail to appreciate the circumstances in a given case. In short, a harm taxonomy represents the foundational building blocks of persecution’s structural meaning.

A. Physical Harm

Physical harms represent one of the central types of persecutory conduct. Qualifying conduct may include beatings, sexual abuse, and mutilation. The different types of physical harms inflicted on victims are, of course, too numerous to list in full. Any conduct causing bodily injury or physical pain, however slight, is sufficient to be included under this form of harm.

The inclusion of physical harms as a basis for establishing persecution is noncontentious and universally accepted. This is not surprising, as physical harm is the most tangible harm to evaluate. Despite its universal acceptance, the relevance and importance of particular physical harms to a persecution assessment have, at times, not been expressed and analyzed with sufficient clarity. Discussions of physical harm should be limited to the expression of specific conduct rather than abstract concepts or terms with particular legal significance. For example, adjudicatory bodies regularly label as “torture” harm that they do not describe, and then factor this determination into their assessments of applicants’ persecution.

54 Ivanishvili v. U.S. Dep’t of Justice, 433 F.3d 332, 340–41 (2d Cir. 2006).
55 See Liu v. Ashcroft, 380 F.3d 307, 313 (7th Cir. 2004).
56 See Brucaj v. Ashcroft, 381 F.3d 602, 610 (7th Cir. 2004); Lopez-Galarza v. INS, 99 F.3d 954, 962–63 (9th Cir. 1996); cf. United States v. Powers, 59 F.3d 1460, 1465–66 (4th Cir. 1995) (“[R]ape, as much as any beating suffered by [the victim], is an act of violence.”).
57 See Bah v. Mukasey, 529 F.3d 99, 112 (2d Cir. 2008).
claims. The term encompasses harms that extend beyond physical pain. Without a description of each type of harm that factors into the persecution assessment, a comprehensive understanding of each harm’s significance diminishes.

B. Restraints and Deprivations of Privacy

There are many harms deemed relevant to a persecution analysis that can collectively be described as restraints and deprivations of privacy. Restraints include arrests, interrogations, imprisonment, and other means of detention. Slavery would also fall under this category, as would indentured servitude and any physical or psychological barrier that prevents a person from leaving a physical location or otherwise impedes movement.

Deprivations of privacy, in this context, encompass a limited subset of the circumstances that can potentially fall under the privacy rubric. Indeed, privacy is a term that may connote many things including autonomy, the right to be left alone, and control over personal information. Many of the ideas touched upon in described conceptions of privacy fall under other categories of harm. Here, however, privacy predominantly refers to surveillance and searches.

See, e.g., Ayele v. Holder, 564 F.3d 862, 866 (7th Cir. 2009); Rodriguez-Matamoros v. INS, 86 F.3d 158, 159 (9th Cir. 1996). Detailed analyses of torture are most often reserved for cases where statutory bars preclude an applicant’s eligibility for asylum, 8 U.S.C. § 1158(b)(2) (2006), and withholding of removal under the INA, id. § 1231(b)(3)(B). See, e.g., Tunis v. Gonzales, 447 F.3d 547, 550 (7th Cir. 2006).

See Kaplun v. Att’y Gen., 602 F.3d 260, 271 (3d Cir. 2010) (describing torture as a “term of art”); 8 C.F.R. §§ 1208.16(c), 1208.18 (2013) (regulations implementing the Convention Against Torture (CAT)).

See 8 C.F.R. § 1208.18(a)(1) (including mental suffering in the definition of torture); see also infra Part II.D (discussing psychological harms).

See, e.g., Gomez-Zuluaga v. Att’y Gen., 527 F.3d 330, 341 (3d Cir. 2008); Gjolaj v. Bureau of Citizenship & Immigration Servs., 468 F.3d 140, 142 (2d Cir. 2006); Begzatowski v. INS, 278 F.3d 665, 669 (7th Cir. 2002). Some opinions have referred more generically to “deprivation of liberty,” a construction that could very well encompass the discussed restraints, but remains too abstract to provide any meaningful guidance. See Mikhail v. INS, 115 F.3d 299, 303 n.2 (5th Cir. 1997); In re T-Z-, 24 I. & N. Dec. 163, 176 (B.I.A. 2007). The “deprivation of liberty” language was originally used in a House report that preceded the passage of the Refugee Act of 1980. See H.R. REP. NO. 95-1452, at 5 (1978).


See, e.g., infra Part II.D.2 (discussing the particularly high psychological harm associated with harms of a sexual nature that are a violation of bodily privacy); infra Part II.E (discussing human rights abuses, which encompass numerous privacy concerns).

See, e.g., Liu v. Ashcroft, 380 F.3d 307, 313 (7th Cir. 2004); Begzatowski, 278 F.3d at 669.
includes the act of watching a person’s activities, monitoring movements, and eavesdropping on conversations taking place via telephone or otherwise.\(^\text{65}\) Searches include searches of a person, a home or other physical location, and those directed at the collection of personal information, such as data stored on a computer or in an e-mail account.\(^\text{66}\) These examples are necessarily nonexclusive, as this is an evolving category whose expanded growth parallels exponential advances in technology. Indeed, in the not-too-distant past, it would have seemed unlikely that pandemic GPS monitoring could be accomplished so facilely.\(^\text{67}\)

\textit{C. Resource and Opportunity Limitation}

The third category concerns economic harms. This category was accepted as actionable even prior to Congress enacting the Refugee Act in 1980.\(^\text{68}\) The Board and every federal circuit court have recognized that economic harms are relevant to a persecution assessment.\(^\text{69}\) Economic harms include the direct confiscation of money, such as theft by private actors or a government’s decision to impose a monetary fine under the color of State authority, but also the destruction or confiscation of real property and other tangible possessions.\(^\text{70}\) Economic harms also encompass the rescission of government services that lead to the loss of a benefit the applicant does not have the means to pay for otherwise; health benefits and food rations would be examples.\(^\text{71}\) The loss of employment is another frequently analyzed economic harm, along with related actions that negatively impact victims’ future earning capacity by way of lost job opportunities or akin means.\(^\text{72}\) Victims of economic harms are usually subjected to multiple economic

\(\text{---}\)

\(^{65}\) See Ayele v. Holder, 564 F.3d 862, 871 (7th Cir. 2009) (discussing the relevance of the government’s purported surveillance of the applicant’s father); see also Gomez-Zuluaga, 527 F.3d at 342–43 (considering the surveillance of the applicant by the Fuerzas Armadas Revolucionarias de Colombia).


\(^{69}\) See, e.g., Li v. Att’y Gen., 400 F.3d 157, 169 (3d Cir. 2005); Yong Hao Chen v. INS, 195 F.3d 198, 204 (4th Cir. 1999); In re T-Z-, 24 I. & N. Dec. 163, 171 (B.I.A. 2007).

\(^{70}\) See Manzur v. U.S. Dep’t of Homeland Sec., 494 F.3d 281, 288 (2d Cir. 2007); Liu v. Ashcroft, 380 F.3d 307, 313 (7th Cir. 2004); Marenco v. INS, 67 F.3d 750, 755 (9th Cir. 1995).

\(^{71}\) See Li, 400 F.3d at 169; see also Chen v. Holder, 604 F.3d 324, 334 (7th Cir. 2010).

\(^{72}\) See Stserba v. Holder, 646 F.3d 964, 969 (6th Cir. 2011) (holding that a failure to recognize the applicant’s medical degree constitutes persecution); Li, 400 F.3d at 169; \textit{In re...}
consequences that manifest concurrently or sequentially. If a government fires someone and prohibits the individual’s future employment elsewhere in the government, such conduct can impact both the individual’s current paycheck and future opportunities, particularly if the individual possesses a skill set uniquely suited for government work, or where job prospects in the private sector are bleak.

The economic disadvantage, however, must be deliberately imposed. This prerequisite is merely a way to ensure that States do not afford refugee protection to individuals subjected to generally depressed economic opportunities affecting a country’s citizenry as a whole, forces of nature, or other analogous circumstances. Such limitations are reasonable. Arguably, all categories of harm have some requisite level of deliberateness. Circling back to physical harms to illustrate, getting hit in the head by a rock would not be relevant to a persecution assessment if the injury resulted from a rock that happened to shoot out from the tire of a passing motorist’s vehicle.

---

73 See, e.g., Mirzoyan v. Gonzales, 457 F.3d 217, 219 (2d Cir. 2006) (describing various limitations placed on education and employment); Li, 400 F.3d at 167–69 (detailing a situation where an individual was heavily fined, fired from his job, and blacklisted from government employment).

74 Cf. Borca v. INS, 77 F.3d 210, 215 (7th Cir. 1996) (reviewing how an applicant was fired from a medical job with the government and limited to a position as a farm laborer).


76 See HATHAWAY, supra note 11, at 117–18 (distinguishing refugees from economic migrants).

77 As discussed further infra Part III.B, this discussion excludes any reference to refugee claims premised on a pattern or practice of persecution. See 8 C.F.R. § 1208.16(b)(2)(iii) (2013).

78 Despite the deliberateness requirement, however, case law has never established whether the consequences of the actions have to be deliberate, or merely the actions that set in motion the chain of events that led to the end result. See, e.g., Krotova v. Gonzales, 416 F.3d 1080, 1087 (9th Cir. 2005) (noting in passing that “economic pressure” among other factors supported a finding of persecution). This distinction raises a number of interesting issues that are beyond the scope of this Article regarding the requisite state of mind of the persecutor and whether the actions or consequences must be intended specifically, or merely that the persecutor engaged in deliberate conduct reasonably certain to cause the resulting economic disadvantage. Cf. infra Part II.D.3 (discussing generally the requisite mental state of a persecutor who causes an applicant to suffer by harming a third party).
D. Psychological Harm

The harms discussed in the previous three categories all involve actions and consequences that are comprehensively tangible and lucid. Psychological harms are much harder to assess in this respect. The consequences are largely internal, and thus less subject to verification. Perhaps the opaque nature of mental states accounts for the failure of the Board and federal circuit courts to engage in a systematic effort to develop a framework for analyzing psychological harm. But psychological harms are undoubtedly an important ingredient in the persecution lexicon.

Two points show why psychological harms should be recognized as a distinct component of a persecution analysis. First, recognizing such harm is consistent with a central purpose behind refugee protection, namely to protect individuals who are compelled to flee a constant and unrelenting feeling of hopelessness for their safety and are deprived of any ability to seek recourse from the State. It is this fear for safety that makes psychological harms so poignant because the attendant deprivation of emotional tranquility can be continuous. Second, psychological harms are a frequent byproduct of other types of harms. Discounting the accompanying mental suffering would skew the true breadth of the harm a victim actually experiences. The consequences of physical harms can heal over a period of time, but the psychological repercussions might not mend so quickly, if at all.

Notwithstanding these two points, psychological harm’s inclusion need not hang its hat on aspirational goals. Rather, two additional reasons establish that such harm is a component of the persecution analysis. The first reason is premised on the fact that mental pain and suffering may constitute torture. Because the requisite harm needed to establish torture is more severe than that needed to establish persecution, mental pain and suffering that qualifies as torture “is a fortiori conduct that reaches the level of persecution.” Thus, the regulations implementing the Convention Against Torture (CAT) provide an express basis for finding that psychological harm is relevant to a persecution discussion. Second, in

---

79 See Niang v. Gonzales, 492 F.3d 505, 512 (4th Cir. 2007) (“[P]ersecution’ cannot be based on a fear of psychological harm alone . . . .”).
80 See HATHAWAY, supra note 11, at 101–02.
81 See Uwais v. U.S. Att’y Gen., 478 F.3d 513, 518 (2d Cir. 2007) (discussing violence perpetrated without fear of repercussion); U.N. Refugee Handbook, supra note 68, para. 53 (noting the heightened suffering caused by a “general atmosphere of insecurity”).
82 See infra Part II.D.2.
83 See 8 C.F.R. § 1208.18(a)(1) (2013) (“Torture is defined as any act by which severe pain or suffering, whether physical or mental, is intentionally inflicted on a person . . . .”).
84 Nuru v. Gonzales, 404 F.3d 1207, 1225 (9th Cir. 2005). Even though the threshold harm required to establish torture is higher than persecution, an applicant entitled to protection under the regulations implementing the Convention Against Torture may nevertheless fail to establish eligibility for asylum relief because of the nexus requirement. See 8 U.S.C. § 1101(a)(42) (2006).
practice adjudicatory bodies do regularly credit psychological harm, even though
they very seldom link the analyzed harm to mental suffering.85

Nearly all the triggers of psychological harm can be broken down into three
categories: threats, harms to self, and harms to third parties.86 A government’s
participation in persecutory conduct, or failure to control the harmful actions
of private actors, represents a quasi-fourth category.87

1. Threats

Threats can help gauge the requisite likelihood of persecution,88 but they are
also relevant to the substance of the harm analysis because of the psychological
consequences they cause the recipient. Numerous consequences can stem from
threats. They may cause the recipient to live in a constant state of fear that the
threat will be acted on. The level of distress may in turn lead the recipient to alter
his lifestyle or routine to try and avoid the threatened harm.89 In extreme cases,
threats may compel a move to a different part of the home country or cause the
applicant to flee the country altogether, which could lead to job loss and other
consequences tied to resource and opportunity limitation.90

The Board and federal circuit courts appear to recognize that threats can be
relevant in this way by factoring them into the assessment of whether the applicant
has established past persecution.91 Yet despite the thousands of asylum cases
where threats are at issue, only on rare occasions have adjudicators expressly noted
that threats represent a form of psychological harm; adjudicators scrutinize the
nature of the harm involved even less.92 This lack of scrutiny often leads

85 See, e.g., infra Part II.D.1 (discussing myriad cases that recognize threats as
potentially relevant).
86 For an example of psychological harm that does not squarely fall under one of these
categories, see Begzatowski v. INS, 278 F.3d 665, 669–70 (7th Cir. 2002), which describes
actions of military commanders “meant to humiliate” the applicant.
87 See infra Part II.D.4.
88 See Boykov v. INS, 109 F.3d 413, 416 (7th Cir. 1997) ("U[if]nfulfilled threats will
fall within that category of past experience more properly viewed as indicative of the
danger of future persecution."). For asylum claims, the applicant would have to prove that
there is some reasonable possibility that persecution will occur, while withholding of
removal under the INA requires a showing that persecution is more likely than not to occur.
89 See Ruano v. Ashcroft, 301 F.3d 1155, 1158 (9th Cir. 2002).
90 See Gonzalez v. INS, 82 F.3d 903, 906 (9th Cir. 1996) (noting that threats were an
impetus to flee); see also Duran v. Detroit News, Inc., 504 N.W.2d 715, 718 (Mich. Ct.
App. 1993) (explaining how the applicant fled Columbia after she received death threats).
91 See, e.g., Fei Mei Cheng v. Att’y Gen., 623 F.3d 175, 193 (3d Cir. 2010); Vilela v.
Holder, 620 F.3d 25, 29 (1st Cir. 2010); Tamara-Gomez v. Gonzales, 447 F.3d 343, 346–
49 (5th Cir. 2006); Ouda v. INS, 324 F.3d 445, 454 (6th Cir. 2003).
92 For examples in which courts have noted the link between threats and
psychological harm, see Makhoul v. Ashcroft, 387 F.3d 75, 80 (1st Cir. 2004) (recognizing
through judicial precedent that threats may at times constitute actionable psychological
adjudicators to use the immediacy and perceived severity of the threat as a cut-off point for its relevance. The First Circuit, for example, has stated that only certain threats are “menacing” enough to cause “actual suffering or harm.”

Such rigid cut-off points are unwarranted. Psychological harm from threats should be assessed on a continuum rather than determining which side of a coin they fall under. A pinprick is still physical harm even though it is minor. Just as physical pain comes in degrees, so too does psychological harm. A threat to harm an applicant who fails to comply with a demand might not entail imminent danger, but it can still cause continuous psychological suffering, as the applicant contemplates the potential repercussions of noncompliance. Similarly, the impact of threats levied against a detained applicant can be particularly acute, regardless of whether the presumed consequences would immediately come to fruition. Whether harms (individually and collectively) are severe enough to qualify as persecution is entirely distinct from the question of whether harm occurred in the first place. Greater recognition of the nature of the harm involved is essential for an accurate evaluation of how threats should factor into a persecution assessment.

So why, then, have adjudicators been so reluctant to consider the psychological harm that threats may evoke? The reason would appear to be driven in part by evidentiary concerns. The Ninth Circuit in Lim v. INS perhaps summarized the concern best, noting that because “claims of threats are hard to disprove,” a finding that a threat constitutes past persecution would “[f]lip the burden of proof every time an asylum applicant claimed that he had been threatened” and “unduly handcuff” the immigration agency. There are indeed a

93 See, e.g., Vilela, 620 F.3d at 29; Zakia Antor v. U.S. Att’y Gen., 280 F. App’x 803, 806–07 (11th Cir. 2008) (discounting threats as “harassing” without further discussion); Li v. Att’y Gen., 400 F.3d 157, 164–65 (3d Cir. 2005); Corado v. Ashcroft, 384 F.3d 945, 947 (8th Cir. 2004); Lim v. INS, 224 F.3d 929, 936 (9th Cir. 2000); Boykov, 109 F.3d at 416 (“We have left open the possibility that threats of a most immediate and menacing nature might, in some circumstances, constitute past persecution.”); cf. Mamouzian v. Ashcroft, 390 F.3d 1129, 1134 (9th Cir. 2004) (reviewing a line of cases in the Ninth Circuit that mention the potential relevance of death threats alone or when combined with other forms of harm).

94 Vilela, 620 F.3d at 29.


97 224 F.3d 929.

98 Id. at 936.
variety of evidentiary considerations present in an assessment of psychological harms. Then again, evidentiary problems rear their pernicious heads in most aspects of asylum claims. The alleged circumstances that purportedly forced the applicant to flee happened in another country, where proof might be hard to come by or simply nonexistent due to the country’s record-keeping practices. So it will often come down to the credibility of the applicant. The applicant can lie about the existence of a threat, the reasons why the applicant believes the threat would likely be carried out, and the attendant psychological trauma, just as easily as the applicant could lie about almost any of the events and circumstances alleged in support of a claim. Gauging the credibility of an applicant is a consideration with its own set of rules that are distinct from the harm that the applicant is trying to establish.

Moreover, even if an applicant is found credible, an immigration judge (and appellate bodies) need not accept as “gospel” every word uttered by the applicant during testimony and submitted into the record. There is a separate issue related to burden of proof. It makes no difference whether the applicant honestly believes the perpetrator will carry out the threat if the evidence lends itself to a contrary conclusion. In short, a recognition of the challenges associated with proving psychological harm is a legitimate point. However, these challenges do not provide a basis to discount nearly categorically a class of harms that should bear potential relevance to the question of whether the applicant established persecution.

2. Harm to the Applicant

The above discussion on threats addressed a circumstance where psychological harm results even though the applicant is not physically harmed. But psychological harm can serve as a distinct byproduct of a physical harm, separate and apart from the direct physical repercussions of the pain inflicted. If the level of mental suffering consistently correlated with the intensity (and consequences) of the physical pain inflicted, then perhaps the need to evaluate the attendant psychological harm would be minimal. After all, psychological harm would then just be seen as an assumed byproduct consistently and proportionately present in a

99 See Mitondo v. Mukasey, 523 F.3d 784, 788 (7th Cir. 2008).
101 See Kumar v. Gonzales, 444 F.3d 1043, 1060 (9th Cir. 2006) (Kozinski, J., dissenting).
103 See Sveaass, supra note 53, at 313–14 (discussing the causal connection in the context of torture).
review otherwise focused on the physical symptoms experienced by the victim. But this is not the case. The extent of psychological harm is not necessarily consonant with the ensuing physical harm.

Sexual abuse is a very apt example of a circumstance that potentially yields a disconnect between the extent of physical and psychological harm suffered. Although the physical touching in such circumstances could be minimal, no one could legitimately compare the harm suffered from sexual abuse with that of benign physical contact of a similar intensity in other contexts. Indeed, studies assessing the psychological impact of rape and other forms of sexual abuse on its victims have identified a broad range of symptoms, including depression, suicidal thoughts, chronic anxiety, and feelings of alienation. These effects may last for considerable periods of time after the event and, in some instances, these effects have been documented as similar to those experienced by torture victims.

The psychological effects of sexual abuse have not been lost on the adjudicatory bodies that assess persecution claims. The Board, for instance, noted in one case that rape encompasses “mental violence.” The federal courts of appeals have universally accepted that rape can establish persecution, and on rare occasion, several courts have explored in greater detail the psychological underpinnings to their assumption that the harm is severe. Adjudicatory bodies’ explicit or implicit acceptance of these underpinnings in limited instances provides a tangible example of the important role psychological harms play in an assessment of inflicted physical injuries. Yet, no effort has been made to recognize or apply more universally the general premise underlying these bodies’ analyses of sexual violence cases. Failing to recognize and assess these accompanying psychological harms more broadly vitiates a distinct repercussion that leads individuals to flee their home countries and supports their fear of returning. As

104 The proportional presence of psychological harm would not, however, provide a basis for discounting it when assessing the extent of harm suffered by the applicant. See infra Part II.D.4.


106 Aswad, supra note 105, at 1931–38.


109 See Brucaj v. Ashcroft, 381 F.3d 602, 610 (7th Cir. 2004); Lopez-Galarza, 99 F.3d at 962–63.

110 Indeed, even in Lopez-Galarza, the Ninth Circuit’s in-depth discussion of psychological harm only came to fruition when the court had to determine whether the applicant was entitled to a discretionary grant of asylum in the absence of a well-founded fear of persecution. 99 F.3d at 960–61; see also 8 C.F.R. § 1208.13(b)(1)(iii) (2013) (codifying the humanitarian asylum standard).
with threats, there are attendant evidentiary concerns. They do not provide a justification for failing to appreciate that the psychological harms caused by physical harms may provide an additional factor relevant to a persecution assessment.

3. Harm to Third Parties Sufficiently Connected to the Applicant

An applicant may also suffer psychological harm when an alleged persecutor inflicts pain or other suffering on a third party. In such a scenario, the applicant need not be physically harmed or threatened with harm. It is for this reason that some courts have concluded that a claim based on harm to another is not viable under the theory that it is simply a derivative asylum claim, or that emotional distress based on harm to a third party is not worthy of protection. Assumptions this broad are misguided. Persecution is based on harm. Psychological suffering is a type of harm. Injuries to third parties can cause psychological harm. Thus, the question should not be whether violence directed at third parties constitutes harm, but rather under what circumstances it can be harm relevant to a persecution determination.

The circumstances should be assessed by looking at the harm on a continuum that is based on four factors: the severity of the harm to the third party, the relationship between the third party and the applicant, the applicant’s sensory witness to the harm inflicted, and the intent of the perpetrator. On one end of the spectrum would be extremely severe harm inflicted on someone particularly close

111 Causation is a potential concern that particularly sticks out. See State v. Cressey, 628 A.2d 696, 700 (N.H. 1993) (stating that, in the context of sexual abuse of a child, “[m]any of the symptoms considered to be indicators of sexual abuse . . . could just as easily be the result of some other problem, or simply may be appearing in the natural course of the children’s development”).


113 See Mashiri v. Ashcroft, 383 F.3d 1112, 1120 (9th Cir. 2004) (“Zakia lived through violent attacks against all three members of her immediate family . . . .”); Singh v. INS, 94 F.3d 1353, 1360 (9th Cir. 1996) (factoring harm to family into its determination that the record compelled a finding of past persecution).

114 See Niang v. Gonzales, 492 F.3d 505, 512–13 (4th Cir. 2007); see also Gumaneh v. Mukasey, 535 F.3d 785, 790 n.3 (8th Cir. 2008) (declining to address the applicant’s claim that she would suffer psychological trauma if her daughter undergoes female genital mutilation based on waiver of the argument).

115 See Karanja v. Keisler, 251 F. App’x 891, 892 (5th Cir. 2007) (“[T]here is no authority which supports Karanja’s argument that the emotional distress he suffered based on the alleged rapes of his wife and daughter is a type of harm that constitutes past persecution.”).

to the applicant, in front of the applicant, and for the purpose of making the applicant suffer. If a group of armed men physically torture an applicant’s husband in front of her, then the applicant’s consequential mental suffering should be relevant to her persecution claim.\(^{117}\) It has taken conduct of comparable magnitude for the courts to expressly acknowledge the relevance of psychological harm in this scenario. In *Abay v. Ashcroft*,\(^ {118}\) the Sixth Circuit held that the applicant could establish a well-founded fear of persecution based on “being forced to witness the pain and suffering of her daughter” if they were deported to Ethiopia and the daughter was forced to undergo female genital mutilation (FGM).\(^ {119}\)

The correlation between greater harm to a third party and increased mental suffering on the part of the applicant is intuitive. The effects of witnessing the third party harm are documented.\(^ {120}\) The other two factors, however, require further discussion. First, for the relationship component, an applicant will suffer greater psychological harm when the injury is directed at a third party with close ties to the applicant. Thus, harm to immediate family members is usually more consequential than harm to a second cousin.\(^ {121}\) But relationships vary, so ultimately the extent of psychological suffering will be based on the context of the case and the circumstances of the applicant, taking into consideration the familial and relational divergences among cultures and ethnicities.\(^ {122}\)

Second, there is an additional burden to bear by an applicant who knows that a third party sustained injuries for the purpose of causing the applicant to suffer. The inclusion of a (quasi) intent factor is also critical because it conforms the

\(^{117}\) *Cf.* Rodriguez-Matamoros v. INS, 86 F.3d 158, 160 (9th Cir. 1996) (factoring into its persecution finding the fact that the applicant’s sister was killed in front of her).

\(^{118}\) 368 F.3d 634 (6th Cir. 2004).

\(^{119}\) *Id.* at 642; accord Kone v. Holder, 620 F.3d 760, 765–66 (7th Cir. 2010). *But see Niang*, 492 F.3d at 512 (labeling as “derivative” and refusing to recognize persecution claim of applicant based on potential forced FGM of applicant’s daughter).

\(^{120}\) See *AM. PSYCHIATRIC ASS’N*, supra note 116.

\(^{121}\) *See Kahssai v. INS*, 16 F.3d 323, 329 (9th Cir. 1994) (discussing the link between losing immediate family members and the “severe emotional . . . injury” that follows). Applicants used to be able to establish persecution per se on the basis of a forced sterilization or abortion performed on their spouses. *In re C-Y-Z*, 211 I. & N. Dec. 915, 918 (B.I.A. 1997) (en banc); *see also 8 U.S.C. § 1101(a)(42)(B) (2006)* (“[A] person who has been forced to abort a pregnancy or to undergo involuntary sterilization . . . or who has been persecuted for failure or refusal to undergo such a procedure . . . or a person who has a well founded fear that he or she will be forced to undergo such a procedure . . . shall be deemed to have a well founded fear of persecution on account of political opinion.”). However, this psychological harm is no longer determinative in many instances. See Shi Liang Lin v. U.S. Dep’t of Justice, 494 F.3d 296, 300 (2d Cir. 2007) (en banc). The change in the law was purportedly grounded in a reexamination of the statutory text of the INA. *Id.* In truth, the change was partially driven by an onslaught of coercive population control claims of questionable veracity.

scope of actionable harms to other persecution prerequisites. In the absence of a pattern or practice of persecution against a class of individuals the applicant belongs to, the applicant must establish that he was singled out for persecution.\textsuperscript{123} Thus, under the singled-out requirement, it could be argued that the perpetrator’s intent to cause the applicant emotional pain is a necessary prerequisite.\textsuperscript{124}

Nevertheless, such purposeful direction cannot serve as an absolute cut-off point, even if it appears intuitively reasonable, because the courts will rightfully find psychological harms to third parties relevant where the intent to cause the applicant pain is either attenuated, ambiguous, or nonexistent.\textsuperscript{125} Indeed, in the aforementioned \textit{Abay} case, the court did not assume that the individual who would force the applicant’s daughter to undergo FGM would do so for the purpose of causing the applicant pain.\textsuperscript{126} What criteria, then, should be used to draw the probative line in the sand? There are two considerations. The first is the nexus requirement, which mandates that harms be inflicted on account of a protected ground; harms do not fall outside the nexus requirement’s grasp simply because a third party serves as a middleman.\textsuperscript{127} The second is a case-by-case assessment of the unique factual circumstances that can shape what seems reasonable when assessing the singled out requirement of the claim. The relationship should be deemed relevant as long as there is a reasonably sufficient connection between the harm to the third party and the subsequent consequences to the applicant.\textsuperscript{128} The case-by-case assessment adds flexibility to the decisionmaking process to account for unique factual circumstances, yet it still lends credence to the primary basis for the singled out requirement, namely to ensure that applicants do not receive refugee protection for general civil strife and random acts of violence.\textsuperscript{129} Such

\textsuperscript{123} See Lie v. Ashcroft, 396 F.3d 530, 536 (3d Cir. 2005); Ramsameachire v. Ashcroft, 357 F.3d 169, 183 (2d Cir. 2004); Prasad v. INS, 47 F.3d 336, 340 (9th Cir. 1995) (“Particularized individual persecution, not merely conditions of discrimination in the country of origin, must be shown before asylum will be granted.”); see also 8 C.F.R. § 1208.13(b)(2)(iii) (2013) (codifying pattern or practice exception).

\textsuperscript{124} This is different from a specific intent to punish, which is a flawed criteria discussed infra Part V.D.

\textsuperscript{125} See Abay v. Ashcroft, 368 F.3d 634, 642 (6th Cir. 2004); see also Zeqiri v. Mukasey, 529 F.3d 364, 370 (7th Cir. 2008) (inferring that conduct beyond the express “singled out” requirement may be sufficient when the applicant is “otherwise ‘personally’ persecuted”); Makhoul v. Ashcroft, 387 F.3d 75, 80–83 (1st Cir. 2004) (noting that “under the right set of circumstances, a finding of past persecution might rest on a showing of psychological harm,” but refusing, on procedural grounds, to hear the applicant’s argument that he suffered “psychological torment” based on fears arising from a friend’s arrest).

\textsuperscript{126} See Abay, 368 F.3d at 642.


\textsuperscript{128} For examples of third party harms that illustrate a wide spectrum of perceived sufficient connectivity, see Mashiri v. Ashcroft, 383 F.3d 1112, 1120 (9th Cir. 2004).

\textsuperscript{129} See Santos-Lemus v. Mukasey, 542 F.3d 738, 746–47 (9th Cir. 2008) (describing the singled out requirement as “well-established”). For critiques of the singled out requirement, see HATHAWAY, supra note 11, at 90–97; Michael G. Heyman, Redefining Refugee: A Proposal for Relief for the Victims of Civil Strife, 24 SAN DIEGO L. REV. 449,
random acts of violence, however, can contribute to a general atmosphere of insecurity that, in turn, may bear relevance to a persecution assessment for reasons explored in the next section.

4. Government Action (or Inaction) and Mental Suffering

One of the prerequisites to establishing refugee status is the need to show that the State perpetrates the harm in question or abdicates to the actions of private actors.\textsuperscript{130} As one asylum expert has noted, “The purpose of refugee law is to provide alternative protection when an individual’s government has failed in its duty to safeguard basic rights.”\textsuperscript{131} The government involvement requirement also serves an important limiting principle by ensuring that refugee status is reserved for people who, at minimum when feasible, attempt to obtain redress from their own governments before seeking protection in another State.\textsuperscript{132} Thus, even if an applicant suffered unspeakable acts of violence, the applicant would not be entitled to asylum protection if the applicant could have sought refuge with the police but decided instead to flee the country.

Even though government involvement represents a distinct prong of the refugee definition, many of the facts probative of this part of the definition are also relevant to the persecution prong.\textsuperscript{133} Indeed, interelement factual overlap is common, both within\textsuperscript{134} and outside\textsuperscript{135} the immigration sphere. In the persecution prong, the failure of State protection is a relevant factor because it both causes harm and intensifies the severity of other harms.\textsuperscript{136} For whatever the harm


\textsuperscript{130} See Rahimzadeh v. Holder, 613 F.3d 916, 920 (9th Cir. 2010).

\textsuperscript{131} ANKER, supra note 26, § 4:7, at 182; see also Rahimzadeh, 613 F.3d at 923 (“[T]he question in an asylum case is whether the police could and would provide protection.”).

\textsuperscript{132} See Mazariegos v. Office of the U.S. Att’y Gen., 241 F.3d 1320, 1327 (11th Cir. 2001).

\textsuperscript{133} But cf. Sabrina v. Gonzales, 217 F. App’x 671, 672 (9th Cir. 2007) (discussing concurrently the harm suffered and role of government without linking the role of government to an additional potential form of harm); Mashiri, 383 F.3d at 1120–22 (limiting its discussion of “emotional trauma” to events other than the subsequently discussed inability of the government to control the persecutory acts).

\textsuperscript{134} For instance, government involvement and internal relocation are distinct segments in the refugee analysis, but the facts probative of each often overlap. See 8 C.F.R. § 1208.13(b)(1)(i)(B) (2013).


\textsuperscript{136} The distinct role of government involvement as a separate prong of the refugee definition on the one hand, and as a part of the persecution analysis on the other, is one of the reasons why the persecution definition proposed in this Article differs from
inflicted, mental suffering is exacerbated by a feeling of helplessness caused by an underlying belief that public officials or private actors can act with impunity.\textsuperscript{137} Mental suffering also stems from an awareness that a lack of any deterrence makes additional offenses more likely if the persecutor has the intent to carry them out. For example, a perpetrator is often more likely to carry out a threat if the perpetrator does not fear consequences, and a potential victim is more likely to be traumatized by an outstanding threat that the potential victim believes the perpetrator is more capable of carrying out. Thus, insecurity is a particularly acute cause of mental suffering because it is prevalent and ongoing.

A skeptic may reasonably question why there is any need to address, in the persecution prong, a form of harm that is presumptively universally present. Two reasons in particular stand out for why it should be recognized. First, the existence of suffering caused by the role of the State is still harm, and its broad application in the refugee context does not change this fact. Whatever threshold of severity adjudicators believe should apply to persecution,\textsuperscript{138} that threshold should be assessed on the basis of the level of harm applicants actually experienced (or fear will come to fruition in the future).

The second reason concerns the impact that harms may have on an applicant when the harms in question cannot directly factor into a persecution assessment. These indirect harms include harms to the applicant deemed irrelevant because they fail to satisfy one of the other elements of the refugee definition or acts of violence directed at the applicant’s fellow citizens.\textsuperscript{139} To illustrate, assume that private actors harmed an applicant on two different occasions. It may very well be the case that the government regularly ignores the first type of injury but has a slightly better track record of investigating and prosecuting the second inflicted harm—although the applicant is aware of several instances where others have suffered the second harm without subsequent action by the government.\textsuperscript{140} As it turns out, the applicant did not even attempt to seek redress after either harm. Since the record may not support the conclusion that the government necessarily would abdicate in its duty to protect the victim from the second harm, it could be argued

---

\textsuperscript{137} See Ayala v. U.S. Att’y Gen., 605 F.3d 941, 946 (11th Cir. 2010) (noting unfettered corruption and brutality in the Venezuelan police force); Uwais v. U.S. Att’y Gen., 478 F.3d 513, 518 (2d Cir. 2007) (discussing violence perpetrated without fear of repercussion); Ian Black, Rights Watchdog Accuses Syria of Crimes Against Humanity, GUARDIAN (London), June 2, 2011, at 22 (discussing serious harms perpetrated without consequence).

\textsuperscript{138} See infra Part III (discussing the severity issue).

\textsuperscript{139} See Mansour v. Ashcroft, 390 F.3d 667, 673 (9th Cir. 2004) (discounting the relevance of mistreatment because the applicant failed to establish that the government is unable or unwilling to prevent the mistreatment in question).

\textsuperscript{140} See generally Afriyie v. Holder, 613 F.3d 924, 931–34 (9th Cir. 2010) (illustrating some of the nuances associated with discerning whether a government is unable or unwilling to control harms).
that the second harm should not be relevant to the applicant’s persecution claim. Moreover, factoring in the circumstances of third parties who suffered comparable harms would seem to circumvent applicants’ need to establish that they were “singled out” for persecution.141

The answer to this quantitative conundrum accounts for the inflicted harms and the singled out requirement: although certain harms to the applicant, widespread harm in the country of origin, and other general conditions of instability cannot be directly used to find that an applicant has established persecution, they can be relevant to understanding the climate that couches the level of suffering experienced by the applicant on the basis of harm that is directed at the applicant. In the influential Handbook on Procedures and Criteria for Determining Refugee Status, the U.N. High Commissioner for Refugees alluded to this possibility when he said that a “general atmosphere of insecurity in the country of origin” may contribute to a persecution claim in certain instances.142 The analytical underpinnings for this pronouncement were never made clear, however, and thus they were not incorporated into U.S. refugee jurisprudence.143

E. Deprivations of Human Rights

Persecution encompasses harms other than human rights abuses,144 but a primary grounding of refugee law is to protect individuals and groups from human rights abuses that represent “a breakdown in the relationship between a person and his or her state.”145 Thus, it is important to assess how these rights fit into an understanding of the persecution definition. In particular, should the human rights grounding of the harms in question be considered expressly in persecution assessments, or is it sufficient to discuss the specific forms of harm falling into one of the four above-mentioned categories without reference to any human rights that

---

141 See 8 U.S.C. § 1208.13(b)(2)(iii) (2006); Prasad v. INS, 47 F.3d 336, 340 (9th Cir. 1995); supra Part II.D.3.
142 U.N. Refugee Handbook, supra note 68, para. 53. The Handbook discusses insecurity in the context of widespread discriminatory conduct, which is an artless means of gauging harm. See infra Part III.A.2.b.
143 The Ninth Circuit has incorporated a related “disfavored group” component into its persecution assessment. See Wakkary v. Holder, 558 F.3d 1049, 1062 (9th Cir. 2009). The underlying rationale and application of the “disfavored group” analysis, however, is grounded in a probability assessment rather than being harm based. Consequently, because of its focus on probability, the deficiencies in many of the premises underlying its probability analysis, and the disconnect between the standard and general principles of agency deference, the “disfavored group” analysis is ultimately unsuccessful.
144 U.N. REFUGEE STATUS, supra note 30, at 32; see also U.N. Refugee Handbook, supra note 68, paras. 51–52.
145 ANKER, supra note 26, § 4:3, at 165; see also JOAN FITZPATRICK, HUMAN RIGHTS PROTECTION FOR REFUGEES, ASYLUM-SEEKERS, AND INTERNALLY DISPLACED PERSONS 6 (Joan Fitzpatrick ed., 2002) (“Forced displacement is often the result of human rights violations; flight itself can involve additional harms and, after a person has stopped fleeing, he or she may face additional deprivations.”).
the harm might violate? Ultimately, the human rights underpinnings of many forms of harm should play a role in the persecution analysis when it helps to identify harm and subsequently gauge the severity of harm.

It is beyond debate that the United States is obligated to protect the human rights of individuals seeking refugee status. The preamble of the 1951 United Nations Convention Relating to the Status of Refugees couches the protection of refugees in the protection of the “fundamental rights and freedoms” referred to in the 1948 Universal Declaration of Human Rights. The Universal Declaration catalogues a vast array of rights, including the right not to be subjected to torture; to be free from arbitrary arrest, detention, or exile; to marry and have a family; and freedom of thought, conscience, and religion. These and other such human rights were incorporated into subsequent binding treaties ratified by the United States. Although the United States was not a party to the 1951 Refugee Convention, it did accede to its obligations by ratifying the 1967 United Nations Protocol Relating to the Status of Refugees in 1968. The Protocol is not self-executing and thus only instructive in U.S. immigration law. Nevertheless, the Refugee Act of 1980 was premised on conforming the refugee definition to that of the Convention and Protocol, giving the United States “statutory commitment” to “human rights and humanitarian concerns;” the drafters agreed that a purpose of refugee protection is to safeguard the human rights enumerated in the Universal Declaration.

---


147 Universal Declaration, supra note 41, arts. 5, 9, 16.

148 See ICCPR, supra note 41, arts. 9, 18, 19, 23; see also International Covenant on Economic, Social, and Cultural Rights, Dec. 16, 1966, S. Exec. Doc. D, 95-2 (1978), 993 U.N.T.S. 3 [hereinafter ICESCR] (signed but not ratified by the United States); Restatement (Third) of Foreign Relations Law of the United States § 701 reporters’ note 6 (1987) (“The binding character of the Universal Declaration of Human Rights continues to be debated, . . . but the Declaration has become the accepted general articulation of recognized rights. With some variations, the same rights are recognized by the two principal covenants, the [ICCPR and ICESCR].”).


150 See Stovic, 467 U.S. at 428 n.22.

151 S. REP. NO. 96-256, at 1, 6 (1979), reprinted in 1980 U.S.C.C.A.N. 141, 141, 146–47; see also Refugee Act of 1980, Pub. L. No. 96-212, 94 Stat. 102 (1980) (codified in various sections of 8 U.S.C.); INS v. Cardoza-Fonseca, 480 U.S. 421, 436 (1987) (“If one thing is clear from the legislative history of the new definition of ‘refugee,’ and indeed the entire 1980 Act, it is that one of Congress’ primary purposes was to bring United States
In persecution assessments subsequent to the passage of the Refugee Act, U.S. adjudicatory bodies only in limited instances expressly mention human rights or factor into their assessment the question of whether a harm deprives an applicant of a fundamental right. More often, adjudicatory bodies assess alleged persecutors’ conduct by discussing the harm descriptively. If decisionmakers factor into their judgments whether such conduct violates a human right, they only do so implicitly. Thus, for example, while courts regularly note that the deliberate imposition of substantial economic disadvantage toward an applicant may constitute persecution, they do not typically justify their conclusions by referring to human rights instruments that protect economic freedoms. Similarly, if courts mention that torture can qualify as persecution, they are not compelled to explain that freedom from torture is a nonderogable right under international conventions ratified by the United States.

There are legitimate reasons why reviewing courts would not need to expressly discuss the human rights grounding of each harm relevant to a particular case. For one, some are so established in current U.S. normative understandings of impermissible conduct that they need no explanation, such as the right not to be a slave. Additionally, a right might encompass harm of such varying treatment and severity that an abstract discussion of the right is not nearly as helpful as an analysis focused on the specific consequences of the harm in question. In some cases, legislative enactments trump potentially contrary interpretations of human rights instruments.

refugee law into conformance with the 1967 United Nations Protocol Relating to the Status of Refugees . . . .”); Sale, 509 U.S. at 169 n.20 (stating that “[o]ne of the considerations stated in the Preamble to the [Refugee] Convention is that the United Nations has ‘endeavored to assure refugees the widest possible exercise of . . . fundamental rights and freedoms,’” and implying that a narrow interpretation of how those fundamental rights and freedoms should apply in U.S. refugee law would render the Convention provision in question “a cruel hoax” (citation omitted)).

Deborah Anker attributes this deficiency in part to a “result-oriented analysis by some courts and adjudicators [that] continues to confuse the development of a meaningful framework for analyzing persecution.” See ANKER, supra note 26, § 4:3, at 169.

See, e.g., Li v. Att’y Gen., 400 F.3d 157, 168–69 (3d Cir. 2005); Guan Shan Liao v. U.S. Dep’t of Justice, 293 F.3d 61, 67 (2d Cir. 2002); Yong Hao Chen v. INS, 195 F.3d 198, 204 (4th Cir. 1999); Abdel-Masieh v. INS, 73 F.3d 579 (5th Cir. 1996). For an example of an anomalous case where economic deprivation is expressly linked to a protected right, see Chen v. Holder, 604 F.3d 324, 334 (7th Cir. 2010).

See, e.g., Begzatowski v. INS, 278 F.3d 665, 669 (7th Cir. 2002); Lal v. INS, 255 F.3d 998, 1003, 1006, 1009–10 (9th Cir. 2001).

This phenomenon is probably more prevalent than many human rights scholars would care to admit. See generally Duncan Kennedy, The Critique of Rights in Critical Legal Studies, in LEFT LEGALISM/LEFT CRITIQUE 178 (Wendy Brown & Janet E. Halley eds., 2002) (discussing the role of rights in American legal consciousness and questioning the ongoing coherence of rights discourse).

See Huang v. Holder, 591 F.3d 124, 129–30 (2d Cir. 2010) (giving Chevron deference to the BIA’s interpretation of whether certain coercive population control
These legitimate points justify adjudicators’ failure to expressly review human rights considerations in every case. Nevertheless, the dearth of attention paid to human rights considerations hinders an unabridged understanding of persecution’s parameters. The reason for this hindrance boils down to the fact that there are certain refugee claims whose success depends on how rights are perceived. An example would be a situation where an applicant is prevented from practicing religion or precluded from freedom of thought and expression.157 Certainly the above four categories can explain some of the resulting harm, such as the psychological repercussions stemming from these rights preclusions or any punishment that might result from noncompliance. But these rights encompass a quality that transcends regular categories of harm—conceptions of identity and personhood that are best assessed with an eye toward human rights protection.158 Part III will discuss when harms are sufficiently severe to constitute persecution, and the role that human rights should play in this assessment.159 But needless to say, human rights cannot serve as a gauge of severity unless they are accepted as one of the bases used to assess whether harm exists in the first place. Thus, while concurrently enmeshed in, and distinct from, the first four categories, human rights are a necessary gauge of harm that should inform a persecution assessment that, at bottom, is grounded in measuring harm.

III. SEVERITY OF HARM

The previous Part described the types of harm that can factor into a persecution analysis without delving into the question of when the harms identified would establish persecution. Certainly not all harm would warrant a finding of persecution. If persecution were interpreted so expansively, “a significant percentage of the world’s population would qualify for asylum in this country—and it seems most unlikely that Congress intended such a result.”160 Accordingly, harm can only establish persecution when it reaches a certain level of seriousness.

policies in China constitute persecution, despite the petitioner’s assertion that the harm she suffered was a “flagrant violation of fundamental rights, including the right to reproductive choice and bodily integrity” (citation omitted) (internal quotation marks omitted)).

157 See Bejko v. Gonzales, 468 F.3d 482, 486 (7th Cir. 2006); Kantonii v. Gonzales, 461 F.3d 894, 898 (7th Cir. 2006).


159 See infra Part III.C.

160 Fatin v. INS, 12 F.3d 1233, 1240 (3d Cir. 1993); see also Selimi v. Ashcroft, 360 F.3d 736, 740–41 (7th Cir. 2004) (recognizing that a lenient interpretation of “refugee” would cover a large number of otherwise ineligible asylum applicants); Laura J. Dietrich, United States Asylum Policy, in THE NEW ASYLUM SEEKERS: REFUGEE LAW IN THE 1980S, at 67, 67–68 (David A. Martin ed., 1988) (“[T]he United States cannot grant asylum to
Measuring that requisite level of severity is the focus of Part III. Simply identifying severity as the standard is not the end of the inquiry. The more important question is how to assess whether harm reaches the severity threshold. Section A discusses the use of benchmarks to measure severity. Adjudicators employ these benchmarks under the well-reasoned premise that persecution’s parameters are more accurately delineated by recognizing categories of harm that do not necessarily breach the severity threshold. Because the applied gauges are not equally instructive, Section A will distinguish the helpful gauges from those that end up hindering the severity assessment. Section B draws on the harm taxonomy developed in Part II to delve deeper into the true extent of suffering caused by injurious conduct. After reviewing the three theoretical frameworks that drive current conceptions of severity, Section B will propose a new unified model that seeks to measure more precisely the extent of harm suffered. Having laid out this framework, Section C will demonstrate how an appreciation of human rights law supplements the proposed model. However, the particular ways that rights-based claims are evaluated greatly affects their usefulness. Section C will propose an analytical methodology for assessing rights-based claims that diverts the discussion away from undermining abstraction.

A. The Use of Benchmarks to Gauge the Requisite Severity

Confronting the challenge of determining when harm is severe enough to constitute persecution, decisionmakers have carved out categories of harm that necessarily (or possibly) fall short of the persecution threshold. These carved-out categories provide a gauge for assessing the relative severity of the harm in the case at hand. This tactic is a commonly employed mode of analysis that can be quite helpful. The usefulness of these benchmarks, however, is premised on an accurate gauge. This section will review the benchmarks employed and distinguish those that work from gauges that fall short.

1. Helpful Benchmarks: Conduct Not Condoned

Helpful benchmarks have made clear that harm does not reach the requisite level of severity simply because it includes conduct the United States does not condone. In *Fatin v. INS*, the Third Circuit stated that “persecution does not encompass all treatment that our society regards as unfair, unjust, or even unlawful or unconstitutional.” Similar gauges have noted that persecution “does not include every sort of treatment our society regards as offensive.”

people who are not individually targets of persecution but suffer instead from general conditions in their own countries.”

12 F.3d 1233.

162 *Id.* at 1240; accord *Lukwago v. Ashcroft*, 329 F.3d 157, 167–68 (3d Cir. 2003); see also *Anacassus v. Holder*, 602 F.3d 14, 19 (1st Cir. 2010) (“To establish past persecution, ‘the totality of a petitioner’s experiences [must] add up to more than mere
These descriptions help to reinforce the general notion that States cannot provide protection to everyone who suffers harm. The specific language employed is particularly useful because it lays out quite clearly and succinctly distinct consequences of injurious conduct that cannot alone provide a basis for establishing persecution. Constitutional violations, for example, cannot form the definitive basis for finding that an applicant’s experiences establish persecution. The gauge does not discount the possibility that a constitutional infirmity may bear relevance to the severity inquiry, merely that it will not be dispositive. Thus, it appropriately steers the analysis to other criteria that are outcome determinative. Similarly, by making explicit that conduct does not qualify as persecution solely because it is unfair or unjust, decisionmakers can render sound judgments amidst internal and external dissonance, given the undesirable position of having to enter orders of deportation in circumstances where the evidence unequivocally shows that an applicant has been wronged.

2. Detrimental Benchmarks: Harassment and Discrimination

The most common benchmark for assessing severity distinguishes sufficiently severe harm from conduct that would simply qualify as “harassment” or “discrimination.” The comparison might appear equally instructive to the previously described gauges. However, this is not the case. Harassment and discrimination are often used to describe a perceived continuum, where harm crosses over the severity threshold as soon as it ceases to be seen as mere harassment or discriminatory conduct. This, in essence, can lead adjudicators to discomfiture, unpleasantness, harassment, or unfair treatment.” (alteration in original) (quoting Nikijuluw v. Gonzalez, 427 F.3d 115, 120 (1st Cir. 2005)).

Mansour v. Ashcroft, 390 F.3d 667, 672 (9th Cir. 2004); accord Li v. Gonzales, 405 F.3d 171, 177 (4th Cir. 2005). Adjudicators have incorrectly used the term “offensive” in other contexts for reasons explained infra Part V.C.

See Dietrich, supra note 160, at 67–68.

See Nelson v. INS, 232 F.3d 258, 263–64 (1st Cir. 2000) (reviewing unconstitutional acts that did not constitute persecution); Kacipa v. INS, 944 F.2d 702, 704–05, 707–08 (10th Cir. 1991) (same).

A significant number of concurring opinions express sympathy for the plights of applicants whose claims are ultimately denied. See, e.g., Truong v. Holder, 613 F.3d 938, 942–43 (9th Cir. 2010) (Reinhardt, J., concurring); Sebastian-Sebastian v. INS, 195 F.3d 504, 505 (9th Cir. 1999) (Wiggins, J., concurring).

See, e.g., Vilela v. Holder, 620 F.3d 25, 29 (1st Cir. 2010); Baba v. Holder, 569 F.3d 79, 85 (2d Cir. 2009); Nakibuka v. Gonzales, 421 F.3d 473, 476 (7th Cir. 2005); Li, 405 F.3d at 177; Sepulveda v. U.S. Att’y Gen., 401 F.3d 1226, 1231 (11th Cir. 2005); Ouda v. INS, 324 F.3d 445, 450 (6th Cir. 2003); In re M-F-W-, 24 I. & N. Dec. 633, 639 (B.I.A. 2008).

See, e.g., Baba, 569 F.3d at 85 (“[T]he dividing line between mere harassment and persecution is difficult to draw with precision . . . .”); Ciorba v. Ashcroft, 323 F.3d 539, 545 (7th Cir. 2003) (discussing when conduct “might cross the line from harassment to persecution”); Asani v. INS, 154 F.3d 719, 725 (7th Cir. 1998) (finding likely that harm
incorrectly incorporate into their decisionmaking process the assumption that conduct is persecution simply because it cannot be labeled as harassment or discrimination. The flaws in the discrimination benchmark are further compounded by the fact that discrimination is more related to the perpetrator’s state of mind than the harm the perpetrator inflicts.

(a) The Harassment-Persecution Distinction

The problem with looking at harassment and persecution as falling on a seamless continuum of harm becomes apparent when examining the variations in how adjudicators envision the parameters of harassment. As the Second Circuit has explained, “to ‘harass’ is ‘to vex, trouble, or annoy continually or chronically,’” and “‘harassment’ is ‘[w]ords, conduct, or action (usu[ally] repeated or persistent) that, being directed at a specific person, annoys, alarms, or causes substantial emotional distress in that person and serves no legitimate purpose.’”\(^\text{169}\) This definition conforms to typical conceptions of harassment, which produce the image of conduct causing mental anguish and annoyances of varying degree.\(^\text{170}\)

However, right outside the scope of mere harassing conduct lies a noticeable range of behavior whose severity should not qualify as persecution simply because it falls outside the parameters of harassment.\(^\text{171}\) Indeed, it is counterintuitive to label as “extreme conduct”—as adjudicators often describe persecution—harm that is a hair more severe than harassment.\(^\text{172}\) The severity gap between harassment and persecution demonstrates how the employed harassment gauge can lead to an overly broad conception of persecution. Equally detrimental, it can cause

\(^{169}\) *Ivanishvili*, 433 F.3d at 341 (alterations in original) (quoting BLACK’S LAW DICTIONARY 721 (7th ed. 1999); WEBSTER’S 3D NEW INT’L DICTIONARY 1031 (1981)).


\(^{171}\) See Naeem v. Holder, 321 F. App’x 516, 521 (7th Cir. 2009) (stretching the conduct reasonably equated to harassment).

\(^{172}\) See Fatin v. INS, 12 F.3d 1233, 1240, 1243 (3d Cir. 1993). Compare Mei Fun Wong v. Holder, 633 F.3d 64, 72 (2d Cir. 2011) (referring to persecution as an “extreme concept”), with *Ivanishvili*, 433 F.3d at 341–42 (analyzing whether the applicant was harassed or persecuted).
decisionmakers to eschew a more thorough examination of other indicia of harm severity.\(^{173}\)

Adjudicators have skirted the line between harassment and persecution by expanding the definition of harassment to include conduct that would not ordinarily fall within its grasp.\(^{174}\) In Nelson v. INS,\(^{175}\) the First Circuit labeled as little more than mere “harassment and annoyance” conduct that included “three episodes of solitary confinement of less than 72 hours, each accompanied by physical abuse[,] . . . periodic surveillance, threatening phone calls, occasional stops and searches, and visits to [the applicant’s] place of work.”\(^{176}\) Assuming arguendo that such conduct does not constitute persecution, placing it in the realm of harassment would stretch the term so far as to render it meaningless as a barometer for assessing the severity of harm.\(^{177}\) In short, harassment can only be useful as a gauge for persecution if it is given its ordinary meaning and applied with an understanding of the type and severity of harm it normally entails. Even when given its ordinary meaning, harassment should only serve, at most, as an example of conduct that falls short of persecution.

(b) The Discrimination-Persecution Distinction

Discrimination serves no better as a barometer for gauging persecutory conduct. Unlike harassment, discrimination is not a concept primarily concerned with an overt action in and of itself. Rather, it is more descriptive of the mindset the responsible party employs while engaging in an action, or whether the effects of the conduct in question demonstrate a distinction between characteristics attributable to, or positions held by, certain individuals or groups.\(^{178}\) The relevance


\(^{174}\) In some instances, adjudicators have further blurred the distinction by using harassment and persecution interchangeably. See, e.g., Avetova-Elisseva v. INS, 213 F.3d 1192, 1202 (9th Cir. 2000).

\(^{175}\) 232 F.3d 258 (1st Cir. 2000).

\(^{176}\) Id. at 264; see also Rupey v. Mukasey, 304 F. App’x 453, 456 (7th Cir. 2008) (finding that being “kicked, punched, and knocked down” on several occasions does not go beyond mere harassment” (quoting Ahmed v. Gonzalez, 467 F.3d 669, 673 (7th Cir. 2006))).

\(^{177}\) The First Circuit does appear to occasionally frame the question in a way that gives the impression that persecution is not necessarily established by conduct that is more severe than harassment. See Aguilar-Solis v. INS, 168 F.3d 565, 569–70 (1st Cir. 1999) (“[P]ersecution encompasses more than threats to life or freedom, but less than mere harassment or annoyance.” (citations omitted)). It does not appear, however, that the court has done so with enough clarity or consistency to provide a model gauge.

\(^{178}\) See BLACK’S LAW DICTIONARY (9th ed. 2009). Federal statutes prohibiting discriminatory conduct in particular circumstances often discuss concurrently the act of
of an act’s discriminatory nature falls largely under the nexus requirement of the refugee definition. The nexus prong requires, separate and apart from the harm itself, that the harm was perpetrated because of delineated characteristics the applicant may possess.\footnote{See 8 U.S.C. § 1101(a)(42)(A).}

The persecution prong, by contrast, focuses primarily on harm. If employees at a sporting event let females enter the stadium before males, then the males have been discriminated against but not persecuted. Conversely, if a hospital attendant purposefully withholds a drug from a particular ethnic group, and a member of the disfavored ethnicity sustains paralysis as a result, then the resulting harm would be closer to persecutory conduct. But, the resulting harm would rise closer to the level of persecution because of the severity of the harm. The harm is the outcome determinative factor, even though discrimination took place in both scenarios.\footnote{See supra Part II.D.}

Adjudicators purportedly using discrimination as a barometer often employ the term to indirectly measure harm in a way that has nothing to do with discrimination.\footnote{See In re O-Z-, 22 I. & N. Dec. 23, 26 (B.I.A. 1998).} In these assessments, discrimination is merely a shorthand way to characterize harm that fails to reach the requisite severity. Thus, in Korablina v. INS,\footnote{158 F.3d 1049 (9th Cir. 1998).} the Ninth Circuit found that the conduct in question was persecution rather than discrimination by concluding that the experiences of the applicant amounted to “the infliction of suffering or harm” rather than a “minor disadvantage or trivial inconvenience.”\footnote{Id. at 1044–45 (citations omitted) (internal quotation marks omitted).} Benign disadvantages and inconveniences, however, are synonymous with harm of insufficient severity. Discrimination’s role in this comparison is immaterial.

Although discrimination and harm are relevant to an assessment of refugee status for predominantly different reasons, the two concepts do overlap in certain contexts. As discussed previously, the reasons behind the infliction of harm can magnify the detrimental psychological impact of the event.\footnote{See 8 U.S.C. § 1101(a)(42)(A).} Accordingly, the repercussions of inflicted harm can increase when the perpetrator directs the harm toward the applicant because of a characteristic the applicant possesses; harm inflicted randomly or by mistake does not necessarily produce the same attendant consequences. Moreover, a general atmosphere of discrimination against a

particular class that the applicant belongs to may itself intensify the harm suffered by the applicant.\textsuperscript{185} These circumstances where discrimination can increase harm merely provide an additional reason to label as more severe the ensuing suffering. When discrimination as a distinct concept is divorced from the attendant methods and consequences of the act in question, it does little to help decisionmakers decide whether the conduct constitutes persecution.\textsuperscript{186} Consequently, discrimination is a largely flawed gauge that hinders the needed focus on the severity of harm. Discrimination should be understood to apply largely to the nexus requirement, and it should therefore be brought into an evaluation of the persecution prong when the act of discrimination itself has some independent impact on the level of suffering experienced by the applicant.

\textbf{B. Developing a Framework to Gauge the Severity of Harm}

While particular severity gauges are instructive in certain instances, they do not fundamentally address the appropriate framework needed to assess the severity of harm. Theoretically and in practice, opinions on the appropriate methodology diverge considerably. The varying approaches can roughly be broken down into three frameworks: the minimal harm model, the cumulative harm model, and the systematic abuse model.\textsuperscript{187} Adjudicators do not always adhere consistently to one particular framework.\textsuperscript{188} While this is partially the result of contemplated overlap in the theoretical assumptions guiding each model, it also stems from the scant attention paid to assessing the underlying drivers of harm. Following a discussion of these three models, this section will propose a new model for appraising the severity component of persecution, which will be referred to as the \textit{continuous suffering model}. The continuous suffering model seeks to identify the root causes of suffering. In so doing, this new model accounts for the warranted justifications driving each of the three established frameworks and discounts aspects of these frameworks that hinder a comprehensive assessment of harm’s severity.

\textsuperscript{185} See U.N. Refugee Handbook, \textit{supra} note 68, para. 55; see \textit{supra} Part II.D.4. For an example of an outlying case that does assess the impact of a generally discriminatory climate, see \textit{Duarte de Guinac v. INS}, 179 F.3d 1156, 1161–63 (9th Cir. 1999).

\textsuperscript{186} Additionally, while perceived in negative terms in persecution assessments, discrimination can also be beneficial to an applicant. See \textit{generally} Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701 (2007).

\textsuperscript{187} See \textit{infra} Part III.B.1 (reviewing these three frameworks).

\textsuperscript{188} \textit{Compare} Beskovic v. Gonzales, 467 F.3d 223, 226 (2d Cir. 2006) (holding a minor beating while detained suffices), \textit{with} Nezha v. Gonzales, 179 F. App’x 89, 90 (2d Cir. 2006) (holding an arrest and beating is not sufficient).
1. Main Analytical Frameworks Employed

(a) Minimal Harm Model

Under the minimal harm model of persecution, the frequency and severity of harm are only secondary considerations in determining whether the applicant has satisfied the persecution prong of the refugee definition. This model is premised on the notion that States should ordinarily provide refugee protection to applicants who demonstrate a nexus between any minimal harm and a protected ground and establish the requisite government involvement or inability to control private actors. By discounting the importance of showing serious harm, the minimal harm model implicitly renders a judgment call on the justifications for providing refugee protection in the first place. The model essentially proffers the position that refugee protection should be based on the legitimacy and purpose of the actions directed at the applicant rather than the actual harm suffered by the applicant. In Beskovic v. Gonzales, for example, the Second Circuit stated that “any physical degradation designed to cause pain, humiliation, or other suffering, may rise to the level of persecution if it occurred in the context of an arrest or detention on the basis of a protected ground.” Thus, under the minimal harm model, the persecution prong is satisfied as long as the actions against the applicant entail treatment civilized society would condone.

(b) Cumulative Harm Model

Unlike the minimal harm model, the cumulative harm model recognizes as germane to a persecution assessment both the number of incidents an applicant experiences and the severity of each harm. The model’s persecution inquiry is grounded in the foundational premise that instances of harm should not be viewed in isolation. Thus, an adjudicator should not discount two incidents of harm simply because neither individually rises to the level of persecution. Axiomatically, the sum total of multiple incidents can rise to the level of

---

189 See Mei Fun Wong v. Holder, 633 F.3d 64, 71–72 (2d Cir. 2011); Gjolaj v. Bureau of Citizenship & Immigration Servs., 468 F.3d 140, 142–43 (2d Cir. 2006); Gilaj v. Gonzales, 408 F.3d 275, 284–85 (6th Cir. 2005); see also Irasoc v. Mukasey, 522 F.3d 727, 730 (7th Cir. 2008) (stating that severity of harm is important but frequency is not determinative).
190 467 F.3d 223.
191 Id. at 226.
192 See Ritonga v. Holder, 633 F.3d 971, 975 (10th Cir. 2011); Fei Mei Cheng v. Att’y Gen., 623 F.3d 175, 192 (3d Cir. 2010); Chen v. Holder, 604 F.3d 324, 333–34 (7th Cir. 2010); Gjolaj, 468 F.3d at 143; Krotova v. Gonzales, 416 F.3d 1080, 1084 (9th Cir. 2005); Korabina v. INS, 158 F.3d 1038, 1044 (9th Cir. 1998); Surita v. INS, 95 F.3d 814, 819 (9th Cir. 1996).
193 See Bejko v. Gonzales, 468 F.3d 482, 486 (7th Cir. 2006).
persecution even where some, or even none, of the harms would separately cross
the severity threshold.194

Unsurprisingly, the severity of the overall harm an applicant experiences
expands as the frequency of harmful acts perpetrated against the applicant
increases. Some of the harms inflicted on the applicant might, however, lack a
causal connection to one of the five protected grounds.195 Adjudicators applying
the cumulative harm model discount from their severity assessment the harms that
do not satisfy the nexus requirement.196 These harms are discounted because the
model aggregates harm much like variables in an equation.

(c) Systematic Abuse Model

The systematic abuse model envisions persecution as a very narrow term of
art that can only be established under particularly dire circumstances.197 Harm, in
and of itself, is not enough to establish persecution.198 In fact, harm suffered on a
regular basis may still be insufficient because the model evaluates the probative
nature of harm in a very precise way. The evaluative criteria have nothing to do
with elements of the refugee definition other than persecution.199 Specifically, the
systematic abuse model distinguishes harm “reflective of a series of isolated
incidents” from harm perpetrated in a systematic way by a government, private
actor, or sufficiently connected class of actors.200 Only the latter systematic harm is
actionable under this theory.

By requiring harm that is both severe and systemic, the systematic abuse
model incorporates the analytical underpinnings of both previously discussed

---

194 See Manzur v. U.S. Dep’t of Homeland Sec., 494 F.3d 281, 290 (2d Cir. 2007).
195 The five protected grounds listed in 8 U.S.C. § 1101(a)(42)(A) (2006) are “race,
religion, nationality, membership in a particular social group, or political opinion.”
196 See, e.g., Restrepo-Velasquez v. U.S. Att’y Gen., 194 F. App’x 865, 867 (11th Cir.
2006); Siswanto v. Gonzales, 177 F. App’x 523, 524 (8th Cir. 2006).
197 In this context, the systematic abuse model is referring to “singled out” persecution
claims rather than those premised on a pattern or practice of persecution, which necessarily
require systematic and pervasive harm. See Ahmed v. Gonzales, 467 F.3d 669, 675 (7th
Cir. 2006).
198 See Pakasi v. Holder, 577 F.3d 44, 47–48 (1st Cir. 2009).
199 See Decky v. Holder, 587 F.3d 104, 111–12 (1st Cir. 2009); see also Hussain v.
Holder, 576 F.3d 54, 57 (1st Cir. 2009) (“[T]o qualify as persecution, a person’s experience
must rise above unpleasantness, harassment, and even basic suffering.” (citation omitted)
(internal quotation marks omitted)); Tasya v. Holder, 574 F.3d 1, 5 (1st Cir. 2009) (“The
abuse alleged . . . although unfortunate and unpleasant, was not frequent enough to compel
a finding of past persecution.”).
200 Bocova v. Gonzales, 412 F.3d 257, 263 (1st Cir. 2005) (citing In re O-Z-, 22 I. &
N. Dec. 23, 26 (B.I.A. 1998)); see also Ratnasingam v. Holder, 556 F.3d 10, 13–14 (1st
Cir. 2009) (discussing why the court determined Ratnasingam was subjected to isolated
incidents, not systematic maltreatment); Wiratama v. Mukasey, 538 F.3d 1, 7 (1st Cir.
2008) (holding Wiratama failed to show past persecution because he did not establish
systematic mistreatment).
The requirement that harm be severe corresponds to the cumulative harm model in that it assesses the level of harm endured by the applicant. The systematic prerequisite incorporates the minimal harm model’s emphasis on the nature and context underlying the actions of the persecutor. The systematic abuse model, however, focuses on the persecutor in a different way. Rather than concerning itself with the persecutor’s motive for inflicting any given harm, the systematic abuse model instead surveys whether the persecutor has demonstrated a pattern of bad behavior directed at the applicant. 201

2. Toward a Unified Approach: Continuous Suffering Model

The three models discussed above represent a wide spectrum of evaluative processes, and each envisions the requisite seriousness of harm very differently. The models provide a number of reasonable methodologies and presumptions about harm that warrant consideration in a unified approach to gauging severity. Nevertheless, some models succeed more than others. The minimal harm model is the least successful. It marginalizes the importance of persecution by diverting attention to other elements of the refugee definition. Persecution is too fundamental to the core justifications of refugee protection to relegate it to an afterthought. Most adjudicatory bodies recognize the central role that harm plays in the refugee definition—at least in principle if not application.

Even though the minimal harm model expands too broadly the level of harm that may qualify as persecution, its belief that one instance of harm can establish persecution cannot be wholly discounted without further inquiry. The sufficiency of one instance of harm is antithetical to the systematic abuse model, but it is congruent with the cumulative harm model. Although the cumulative harm model posits that all incidents of harm must be considered holistically, it does not rule out the possibility that the context surrounding one incident of harm may establish persecution. And indeed, many courts have found as much. 202

So the question becomes why, exactly, one harmful episode may establish persecution in some circumstances, but not others. The severity of the harm is the most obvious answer. In this bare form, however, it is an analytical dead end. Stating that the issue is severity does not explain how to gauge severity beyond general understandings of the types of conduct that would hurt more than others—a slap on the face versus a bone-breaking knock to the chest. Assessing severity by measuring harm at the time of the incident masks a more precise realization about the harm applicants actually experience: namely, the need to account for the continuous effects that initiated harms might have on the applicant. As noted above, 203 for example, sexual abuse cannot be analyzed as a temporally isolated

202 See, e.g., Irasoc v. Mukasey, 522 F.3d 727, 730 (7th Cir. 2008); Guo v. Ashcroft, 361 F.3d 1194, 1197–98, 1203 (9th Cir. 2004) (finding that each of two incidents independently established persecution).
203 See supra Part II.D.2.
event. There is the incident itself and then there are the subsequent psychological consequences and possible physical manifestations of harm.\textsuperscript{204} Similarly, while a death threat is one instance of harm, it can cause ongoing harm for a considerable period of time by virtue of the continuous fear it invokes in the applicant. Moreover, a general atmosphere of insecurity may exacerbate and prolong the harm.\textsuperscript{205} In the case of threats, the backdrop of insecurity provides reason for the applicant to believe the threat can be carried out. Overarching insecurity has many causes, including general conditions of unrest and random unpunished acts of violence. The fact that certain factors cannot independently provide a basis for establishing persecution does not diminish their influence on the extent of harm manifested in the applicant’s continuous experience.\textsuperscript{206}

Assessing the continuity of harm to pinpoint its severity, however, cuts both ways. Constant threats not acted on in any way may diminish their harmful effects despite their recurrence by showing that the actor responsible for the threats has no intention to carry them out.\textsuperscript{207} Similarly, the significance of a single horrible incident may diminish if the applicant spends years in the country unabated and the applicant’s conduct during those years demonstrates the minimal impact of the initial harm on the applicant’s continuous experience.\textsuperscript{208}

Thus, it is the nature of the harm inflicted and the surrounding context that lead to an accurate understanding of inflicted harm’s effects, and it is this set of effects that should form the basis for assessing severity.\textsuperscript{209} This is why the more preferable approach to gauging severity is what this Article has coined the continuous suffering model. This model is consistent with the cumulative effects model in that it looks at the harms collectively. The continuous suffering model is, however, a more refined approach because it is not based on simply aggregating

\textsuperscript{204} See Council on Scientific Affairs, \textit{supra} note 105, at 3185; Santiago et al., \textit{supra} note 105, at 1338–40; Aswad, \textit{supra} note 105, at 1931–39.

\textsuperscript{205} See generally U.N. Refugee Handbook, \textit{supra} note 68, para. 53.

\textsuperscript{206} See \textit{supra} Part II.D.4.

\textsuperscript{207} See Yuk v. Ashcroft, 355 F.3d 1222, 1235 (10th Cir. 2004) (noting that the length of time elapsed since a threat lessens the threat’s importance); Gonzalez v. INS, 82 F.3d 903, 909 (9th Cir. 1996); cf. Gomez-Zuluaga v. Att’y Gen., 527 F.3d 330, 343 (3d Cir. 2008) (“Weighing this final incident in the context of the prior incidents shows that Petitioner’s allegations regarding the imminence and menacing nature of the [earlier] threats are justified.”).

\textsuperscript{208} See Alyas v. Gonzales, 419 F.3d 756, 761 & n.2 (8th Cir. 2005) (noting that the applicant remained in the country for four unabated years after the inflicted harm); see also Melecio-Saquil v. Ashcroft, 337 F.3d 983, 985–87 (8th Cir. 2003) (reviewing how a dramatic change in country conditions diminished the feared harm).

\textsuperscript{209} The Second Circuit has demonstrated a particularly acute awareness of the importance of contextual considerations in assessing harm. See, \textit{e.g.}, Sulaiman v. Holder, 371 F. App’x 224, 226–27 (2d Cir. 2010); Manzur v. U.S. Dep’t of Homeland Sec., 494 F.3d 281, 290 (2d Cir. 2007).
harms deemed relevant like variables in a severity equation, as frequently seen in the cumulative effects model.210

A focus on the harm caused to the applicant is not entirely compatible with the systematic abuse model, but assessing persecution through the lens of harm continuity does account for several of the legitimate concerns that lead decisionmakers to consider the systemic nature of persecutors’ actions. The systematic abuse model seeks to guard refugee status from applicants who experience a set of harms perceived as “isolated.” Thus, to the extent it is concerned with harm severity, the premise for finding that such harms do not give rise to persecution is implicitly grounded in the belief that an incident’s “isolated” nature compartmentalizes its effects.211 The systematic nature of harm, therefore, is presumed to be what gives the harms continuity and provides an ongoing fear that justifies refugee protection.212

The rationale for disfavoring so-called isolated incidents in the systematic abuse model is well grounded, but it is underinclusive because the rationale for requiring some level of harm continuity should not wholly depend on delineated characteristics of the persecutor and a minimum floor of requisite harmful incidents.213 Rather, the continuity should be assessed in relation to the harm experienced. The nexus requirement sufficiently tethers the harms together by ensuring that each directly probative incident was directed at the applicant for reasons codified to be worthy of protection.214 The continuous suffering model accounts for the fact that multiple harms inflicted over time may increase overall severity, but it also provides a rationale for how to assess limited instances or incidents that are not causally related to each other.

There are several potential counterarguments to the continuous suffering model. The first concerns the situation where an applicant suffers harm and proceeds to immediately flee the homeland. The argument would be that any focus on continuity would penalize the applicant for taking the very reasonable step of avoiding additional harm.215 But recognizing that continuity is the most accurate

---

210 See Restrepo-Velasquez v. U.S. Att’y Gen., 194 F. App’x 865, 867 (11th Cir. 2006); Siswanto v. Gonzales, 177 F. App’x 523, 524–25 (8th Cir. 2006).
211 See In re O-Z-, 22 I. & N. Dec. 23, 25–26 (B.I.A. 1998) (recounting the INS position that persecution is not established by isolated “random violence”). The Board’s decision in In re O-Z- has been the impetus for some adjudicators to adhere to the systematic abuse model. See, e.g., Bocova v. Gonzales, 412 F.3d 257, 263 (1st Cir. 2005).
212 See U.N. Refugee Handbook, supra note 68, para. 55 (discussing ongoing fear and apprehension caused by a general atmosphere of discontent); see also Hassan v. Holder, 571 F.3d 631, 642 (7th Cir. 2009) (discounting an isolated shooting because recurring harms typically needed to establish past persecution).
213 See generally Tchemkou v. Gonzales, 495 F.3d 785, 793 (7th Cir. 2007) (questioning the narrow focus underlying the systematic abuse model); Sahi v. Gonzales, 416 F.3d 587, 588 (7th Cir. 2005) (same).
215 See Miljkovic v. Ashcroft, 376 F.3d 754, 756–57 (7th Cir. 2004) (“[A] person can still be a victim of [past] persecution even if he manages to elude his persecutors.”).
way to evaluate the severity of harm does not preclude any possibility that an adjudicator may find past persecution in the above circumstance. Rather, it merely provides the contextual framework for assessing the claim. Context and individual circumstances will always play a role that can transcend general principles and frameworks. The continuous suffering model provides a reasonable caution by focusing the analysis on the methods decisionmakers should use to assess harm.

Additionally, conceiving of a focus on continuity as penalizing the applicant for immediately fleeing undermines the soundness of a persecution assessment by meshing the likelihood of future harm with an evaluation of harm the applicant already endured. A determination that an applicant has not presumptively established a well-founded fear of persecution by virtue of showing past persecution might not be favorable to the applicant, but correctly applying the law should not be viewed as punishment. Refugee law provides a mechanism for establishing eligibility in such situations. The applicant can simply use what happened in the past as evidence that the applicant will face persecution if deported—that is, the applicant can establish a well-founded fear of persecution without any presumption. If the applicant cannot establish a well-founded fear of persecution under a proper application of law, then the applicant is not entitled to refugee protection, even if the attendant circumstances evoke heartfelt sympathy—a recurrent and unfortunate reality that flows from States’ inability to protect or shelter the number of people harmed throughout the world. Thus, viewing as punishment a finding of no past persecution for an applicant who flees immediately, while certainly understandable, is ultimately a red herring to a proper evaluation of persecution law.

A second potential argument against the continuous harm model is that it marginalizes the fact that systematic abuse against an applicant is more worthy of protection than repeated abuse inflicted at different times by different actors. The continuous suffering model, however, merely shifts the systematic nature of harm from a necessary prerequisite to a highly probative factor. Assessing the systematic nature of harm as a necessary prerequisite to establishing persecution is more aptly applicable to pattern or practice of persecution claims. For in such claims, a decisionmaker needs to consider the extent of harm inflicted on members of a group that the applicant belongs to, rather than the individual applicant. Consequently, the systematic nature of harm in pattern or practice claims is simply a circumstantial mechanism used to infer the likelihood of harm to the applicant.

217 See Irasoc v. Mukasey, 522 F.3d 727, 729–30 (7th Cir. 2008) (stating that a viable persecution claim need not depend on establishing past persecution).
219 See Wakkary v. Holder, 558 F.3d 1049, 1061 (9th Cir. 2009) (discussing the link between systematic conduct and a pattern or practice claim).
220 See Rasiah v. Holder, 589 F.3d 1, 5 (1st Cir. 2009); Santoso v. Holder, 580 F.3d 110, 112 (2d Cir. 2009); 8 C.F.R. § 1208.16(b)(2).
Such inferences are not the primary focus of ordinary “singled out” persecution cases because adjudicators can evaluate actual harm suffered or feared by the applicant.

Moreover, it may very well be the case that systematic abuse by an actor or entity increases the applicant’s suffering. As noted previously, the very fact that government actors inflict harms on a recurrent basis may itself increase the applicant’s suffering by eliminating in the applicant’s mind any possibility that redress can be sought out or that protection will be provided.\footnote{See supra Part II.D (reviewing psychological harms).} Evaluating such circumstances under the continuous suffering model simply changes the evaluative lens. In the end, it is a lens that seeks to illuminate the full extent of endured harm.

C. Factoring Human Rights into the Severity Assessment

As a central grounding of refugee protection, human rights law plays a necessary role in measuring the severity of harm. Adjudicators have demonstrated some acceptance that human rights bear relevance to a persecution assessment. Less clear is how exactly a rights-based approach can aid the inquiry.

The United Nations and a number of human rights instruments provide guidance on the relative importance of prescribed rights by distinguishing derogable rights from nonderogable rights. “States may never legitimately restrict” nonderogable rights, such as the right to life and freedom of thought.\footnote{U.N. REFUGEE STATUS, supra note 30, at 32 (emphasis added); see also ICCPR, supra note 41, art. 4(2).} By contrast, derogable rights like the right to freedom of movement may be subject to limitation in certain instances.\footnote{See U.N. REFUGEE STATUS, supra note 30, at 32; ICCPR, supra note 41, art. 4(2).} Beyond derogable and nonderogable rights, there are certain rights deemed aspirational, where States are required to “work progressively toward their objectives.”\footnote{U.N. REFUGEE STATUS, supra note 30, at 32.} These include certain “economic, social and cultural rights.”\footnote{Id.}

The distinctions articulated by the United Nations and incorporated into several human rights instruments have been the subject of much debate.\footnote{See Theodor Meron, On a Hierarchy of International Human Rights, 80 AM. J. INT’L L. 1, 4 (1986).} As one commentator has noted, choosing “which rights are more important than other rights is exceedingly difficult,” absent notable exceptions.\footnote{Id.} Though some human rights advocates would prefer to identify all rights as equally important,\footnote{See Hansje Plagman, The Status of the Right to Life and the Prohibition of Torture Under International Law: Its Implications for the United States, 2003 J. INST. JUST. INT’L STUD. 172, 172–73.} the
existence of some level of distinction is well recognized despite the imprecision in the evaluative task.  

It is unsurprising that the rights distinctions made within the human rights spectrum yield several problematic consequences for measuring persecutory conduct. There can be a disconnect between the perceived importance of a violated right and the actual level of harm the violation causes the applicant. For example, while economic rights are merely aspirational, economic harms can cause immeasurable hardship. Another perceived deficiency stems from the fact that rights distinctions reflect normative judgments. The normative quality inherent in valuing particular rights has led some commentators to question the value of looking to human rights instruments at all to assess persecution claims, with one commentator describing the link between asylum law and human rights as “arbitrary” and “frail” because a persecution finding “hinges on the subjective choice by doctrinal writers, judges and decision makers.”

Recognizing that there is a subjective aspect to describing and quantifying rights does not purge the inevitable judgment call adjudicators must make when assessing harm in persecution claims. These judgments will be made regardless of whether human rights conceptions factor into the decision. The use of a rights-based approach when appropriate can help decisionmakers frame the harms in some systematic way—that is, help decisionmakers express more precisely the overarching nature of the harm. Despite helping adjudicators frame harm more systematically, the ultimate success of a rights-based approach to gauging harm must recognize the precarious aspects of the current regime mentioned above, particularly the perceived disconnect between the stated importance of a right and the severity of harm the applicant actually experiences. Thus, a cogent evaluative mechanism would need to make the inquiry as objective as possible to take the rights-based approach out of the realm of abstraction that has led many decisionmakers and commentators to question its usefulness to an assessment of whether harm constitutes persecution.

---

229 See RESTATEMENT (THIRD) ON FOREIGN RELATIONS LAW OF THE UNITED STATES § 702 cmt. m (1987) (recognizing the relative importance of rights); Meron, supra note 226, at 4 (noting that “personal, cultural, and political bias” make consensus nearly impossible).

230 See Xiu Zhen Lin v. Mukasey, 532 F.3d 596, 597–98 (7th Cir. 2008); see also U.N. REFUGEE STATUS, supra note 30, at 32.

231 Noll, supra note 39, at 494; see also FITZPATRICK, supra note 145, at 8–15 (providing a different conception of the human rights that persecutors can violate).

232 Even if the deprivation of a right constitutes persecution, refugee eligibility would still be premised on a sufficient nexus between the rights deprivation and a protected ground. See Chen v. INS, 95 F.3d 801, 806–07 (9th Cir. 1996) (negating the relevance of a violation of a right deemed fundamental under the Universal Declaration of Human Rights because of the nexus requirement); Sun Wen Chen v. Att’y Gen., 491 F.3d 100, 118 (3d Cir. 2007) (McKee, J., concurring in part & dissenting in part) (“Congress did not authorize refugee status for any violation of ‘fundamental human rights.’”). Human rights advocates have criticized the nexus requirement in this context. See Karen Musalo, Irreconcilable
In light of these considerations, this Article proposes a two-tiered evaluative framework that differs from international instruments and U.N. guidance. Under the first tier, a fact finder would assess whether any abrogation of the right in question could necessarily be sufficient to constitute persecution. If not, the inquiry would move to the second tier. Under the second echelon, the adjudicator would measure the harm by assessing the consequences that attach to the abrogation of the right, taking into consideration as warranted how fundamental of a right the harm concerns. The inquiry in its entirety would take place against the backdrop of the continuous suffering model.

This two-tiered framework is not a contrary or alternative approach to the continuous suffering model. Rather, it supplements the regular persecution analysis by emphasizing two key considerations. The first is the need to further illuminate the nature of the rights involved to aid in an assessment of harm severity. Second, as illustrated below, it provides a needed mechanism for assessing rights violations that skirt the fence between the two tiers.

Many of the rights violations that fall into the first tier are so intuitive that adjudicators could sort them into this category without serious contemplation. For example, if an applicant is harmed to a level that constitutes torture, then the conduct violates the right to be free from torture. We do not subsequently ask whether the harm is severe enough to constitute persecution because the very existence of the rights violation is so recognizably sufficient. The courts of appeals have placed several rights violations into this tier. In Kazemzadeh v. U.S. Attorney General, for example, the Eleventh Circuit adopted the Seventh Circuit’s position that “having to practice religion underground to avoid punishment is itself a form of persecution.”

A second tier is warranted because not all rights violations automatically establish persecution. For example, a judge would not necessarily find that an applicant has established persecution simply because the State denied the applicant the right to marry, even though the right to marry is contemplated in international law. Differences? Divorcing Refugee Protections from Human Rights Norms, 15 Mich. J. Int’l L. 1179, 1182 (1994).

---

233 See supra Part III.B.2 (introducing the continuous suffering model).
234 See ICCPR, supra note 41, arts. 4(2), 7.
235 See Nuru v. Gonzales, 404 F.3d 1207, 1222–23 (9th Cir. 2005) (“[T]orture is illegal under the law of virtually every country in the world and under the international law of human rights. We cannot therefore ever view torture as a lawful method of punishment.”).
236 577 F.3d 1341 (11th Cir. 2009).
237 Id. at 1354 (citing Muhur v. Ashcroft, 355 F.3d 958, 960–61 (7th Cir. 2004)); see also Zhang v. Ashcroft, 388 F.3d 713, 719–20 (9th Cir. 2004) (per curiam) (“[T]o require Zhang to practice his beliefs in secret is contrary to our basic principles of religious freedom and the protection of religious refugees.”).
Rather, the context of the case and other factual circumstances would guide the analysis. *Chen v. Ashcroft* illustrates this point. Local government regulations applicable to the applicant prevented him from marrying before the age of twenty-five. The applicant alleged that he was eligible for asylum relief because the Chinese government denied his request to marry when he was nineteen. The Third Circuit held that the Board was not “bound to conclude that” the minimum age requirement “amounted to persecution.” In support of its holding, the court noted that a “law or practice . . . does not necessarily rise to the level of ‘persecution’ simply because it does not satisfy American constitutional standards or diverges from the pattern followed by other countries.”

Rights evaluations differ based on whether they concern harm already endured or prospective harm. For prospective harms, applicants in certain instances will essentially assert that they will suffer harm based on actions they proactively choose to take. This is where the importance of recognizing and evaluating rights comes into play. No one would sympathize with an applicant who claims he will suffer injury from a fistfight he plans to initiate upon return to his home country. Conversely, applicants have the prerogative to engage in conduct encompassed within a protected right. The extent of this prerogative, however, is based on the relative importance of the right. If the right is considered fundamental, then the ability to avoid punishment by failing to engage in the conduct is more likely to preclude an adjudicator from relying solely on the more tangible consequences the applicant will face for engaging in such conduct.

With the scope of the discussed tiers in mind, two examples applying the proposed evaluative framework will now be discussed. The first concerns slavery. The right not to be a slave is encapsulated in numerous human rights instruments as a nonderogable right. But will a violation of this right necessarily lead to a finding of persecution? Common conceptions of prolonged slavery unequivocally would, but confinement falling under the slavery rubric can last for considerably shorter periods of time. Would one day of forced servitude be sufficient? The objective here is not to pass judgment on the correct answer, but rather to illustrate the inquiry. Whatever the answer, it is not a discussion that one would need to

---

238 See Universal Declaration, *supra* note 41, art. 16(1) (providing the right to any person who has reached “full age”); Loving v. Virginia, 388 U.S. 1 (1967); see also Al-Ghorbani v. Holder, 585 F.3d 980, 996 (6th Cir. 2009).
239 381 F.3d 221 (3d Cir. 2004).
240 *Id.* at 223.
241 *Id.*
242 *Id.* at 231.
243 *Id.* at 230–31.
244 See Zhang v. Ashcroft, 388 F.3d 713, 719–20 (9th Cir. 2004) (per curiam); see also Hamby v. Neel, 368 F.3d 549, 566–67 (6th Cir. 2004) (reviewing when rights are considered fundamental).
245 See Iao v. Gonzales, 400 F.3d 530, 532 (7th Cir. 2005).
246 See, e.g., ICCPR, *supra* note 41, arts. 4(2), 8(1).
partake in if the issue concerned torture. Each protected right presents a unique inquiry that raises divergent questions and considerations.

A second example will revisit the appellate courts’ conclusion that deprivations of the right to practice religion freely are per se persecution. Consider a circumstance where a foreign country prohibits the open practice of Christianity, but the rarely enforced penalty is a small fine.\(^\text{247}\) If an applicant credibly establishes that he practices Christianity and will continue to do so if deported, would the applicant have a well-founded fear of persecution?\(^\text{248}\) The critical question is whether it is possible that the answer would be no. And therein lies the heart of the tiered rights distinction.\(^\text{249}\) If certain conditions may render a rights violation insufficiently severe, then the violation of the right is not per se persecution, and a case-specific analysis must occur.\(^\text{250}\) If no effort is made to incorporate human rights as a distinct concept in evaluations of persecution, however, then decisionmakers cannot develop any systematic criteria to evaluate certain harms where the outcome can determine whether the government deports an applicant.\(^\text{251}\)

**IV. LEGITIMACY OF HARM**

Even if harm reaches the requisite severity, an applicant cannot establish persecution—either in the past or feared prospectively—if the inflicted harm is normatively justified or otherwise permissible. “Legitimacy” is the term used here to encapsulate the various circumstances where harm that may be sufficiently severe still fails to establish persecution.\(^\text{252}\) Oftentimes, the legitimacy of a State’s

---

\(^\text{247}\) Cf. Singh v Minister for Immigration & Multicultural Affairs (2000) 178 ALR 742 (Austl.) (denying a claim based on the applicant’s homosexuality because laws in his home country prohibiting homosexuality were not enforced in urban areas).

\(^\text{248}\) See Sahi v. Gonzales, 416 F.3d 587, 588 (7th Cir. 2005) (proposing a religious persecution hypothetical that does not include the consequences of noncompliance with the proposed restriction to religious freedom). The credibility analysis is especially critical in prospective harm cases based on human rights deprivations. See generally Rempell, supra note 100, at 210–26 (discussing the impact of applicants’ inferred untruthfulness on adverse credibility determinations).

\(^\text{249}\) For additional examples of rights warranting contemplation, see Chen v. Holder, 604 F.3d 324, 333–34 (7th Cir. 2010) (right to have a child); United States v. Shi Yan Liu, 239 F.3d 138, 140 (2d Cir. 2000) (right to be free from certain general searches); Brian Murphy, Bahrain Woman Gets Year in Jail for Critical Poems, DESERET NEWS (June 12, 2011 9:43 AM), http://www.deseretnews.com/article/700143597/Bahrain-woman-gets-year-in-jail-for-critical-poems.html (freedom of expression).

\(^\text{250}\) Cf. U.N. Refugee Handbook, supra note 68, para. 57 (“[P]enal prosecution for a reason mentioned in the definition (for example, in respect of ‘illegal’ religious instruction given to a child) may in itself amount to persecution.” (emphasis added)).

\(^\text{251}\) See Marouf & Anker, supra note 39, at 784–85 (advocating a human rights-based approach to evaluating refugee claims).

\(^\text{252}\) See Price, supra note 129, at 425 (arguing that asylum should be reserved “for those who are exposed to harm illegitimately inflicted by authoritative agents”).
actions is assessed by looking at whether the harm was perpetrated on account of one of the five protected grounds. Nevertheless, legitimacy should properly be regarded as a component of the persecution definition that is sufficiently distinct from the nexus prong because a causal connection between the harm and a protected ground is not a necessary condition for establishing that the State’s actions were illegitimate.

The most common scenario requiring an explicit legitimacy inquiry is where the State seeks to punish the applicant for engaging in certain conduct or failing to engage in conduct the State seeks to compel. The sanctions imposed by the State can be physical confinement, economic penalties, or the loss of benefits and privileges. Whatever the extent of the penalty, the inquiry remains the same. Initially, it is important to assess whether the stated reason for the penalty is merely a pretext for an unjustifiable desire to inflict harm. If the penalty is merely a pretext, then the harm is necessarily illegitimate. On the other hand, if the penalty is not overtly pretextual, then the inquiry shifts to an evaluation of whether the penalty is disproportionately severe given the nature of the violation involved. While a life sentence for first-degree murder may seem proportional, a comparable sentence for stealing a loaf of bread is not. The disproportionality of a penalty might itself provide circumstantial evidence of the punishment’s pretextual nature.

The distinction between murder and larceny is easy to assess, but the circumstances raising legitimacy concerns often present closer questions. For example, China’s justification for imposing sanctions on couples that violate the State’s one-child policy is grounded in a need to forestall population growth that

---

253 Violating a country’s travel restrictions would be an example. See Tesfamichael v. Gonzales, 411 F.3d 169, 177 (5th Cir. 2005); Rodriguez-Roman v. INS, 98 F.3d 416, 427–28 (9th Cir. 1996); Coriolan v. INS, 559 F.2d 993, 1000 (5th Cir. 1977).

254 Actions the State seeks to compel may include the payment of taxes and compulsory military service. See, e.g., Borovsky v. Holder, 612 F.3d 917, 923 (7th Cir. 2010); Sadeghi v. INS, 40 F.3d 1139, 1142 (10th Cir. 1994); M.A. v. INS, 899 F.2d 304, 312 (4th Cir. 1990). The State may also seek to take away citizenship rights for reasons more closely tied to an applicant’s personal characteristics than particular conduct. See Haile v. Holder, 591 F.3d 572, 573–74 (7th Cir. 2010); Haile v. Gonzales, 421 F.3d 493, 495–97 (7th Cir. 2005).

255 See Long Hao Li v. Att’y Gen., 633 F.3d 136, 151 (3d Cir. 2011); Chanco v. INS, 82 F.3d 298, 302 (9th Cir. 1996); see also U.N. Refugee Handbook, supra note 68, para. 85 (“Prosecution for an offence may . . . be a pretext for punishing the offender for his political opinions . . .”).

256 See, e.g., Long Hao Li, 633 F.3d at 151; Vumi v. Gonzales, 502 F.3d 150, 158 (2d Cir. 2007); Tesfamichael, 411 F.3d at 177; Chanco, 82 F.3d at 302; In re Laipenieks, 18 I. & N. Dec. 433, 459 (B.I.A. 1983). Even if applicants establish disproportionately severe treatment, they cannot satisfy the refugee definition unless they also establish a nexus to a protected ground. See Ahmed v. Ashcroft, 396 F.3d 1011, 1015 (8th Cir. 2005). Prosecutions initiated on account of a protected ground, however, will almost certainly be illegitimate. See Li Wu Lin v. INS, 238 F.3d 239, 244 (3d Cir. 2001).

257 Long Hao Li, 633 F.3d at 151.
would otherwise explode and encroach the functioning of the State. The extent of the sanctions appears to be driven by locality, but some regions only impose a modest fine. The legitimacy of a modest fine is based on an inverse relationship between the nature of the harm and the severity of the penalty, taking into account a sovereign nation’s prerogative to regulate conduct within its borders and international human rights norms that transcend national boundaries. In this respect, the legitimacy of State action bears analogy to the constitutional framework guiding the permissible scope of government infringements of rights expressly or implicitly delineated in the U.S. Constitution. The harm involved in this example is monetary, but also extends more broadly to deprivations of rights to make certain familial decisions without arbitrary government interference. Thus, the legitimacy component of the persecution assessment represents another reason why it is important for adjudicators to develop a framework for identifying and evaluating protected rights and freedoms.

The legitimacy inquiry is not limited to government sanctions used to prevent or compel behavior. For instance, State actions may focus on information gathering in the form of searches or surveillance. Legitimacy concerns may also emerge in an assessment of harm directed at applicants by nongovernmental organizations or individuals. The need to analyze legitimacy is particularly

---

258 See Penny Kane, The Second Billion 64–65 (1987) (stating that China’s population control law was originally premised on protecting citizen health and general prosperity); see also Paula Abrams, Population Control and Sustainability: It’s the Same Old Song but with a Different Meaning, 27 Envtl. L. 1111, 1113–15 (1997) (discussing population control and resource allocation).


261 See Robert C. Post, Foreword: Fashioning the Legal Constitution: Culture, Courts, and Law, 117 Harv. L. Rev. 4, 89 (2003) (reviewing how the permissible scope of infringement is based on whether the right in question is fundamental).

262 See ICCPR, supra note 41, art. 17 (protecting against arbitrary interference with family); Universal Declaration, supra note 41, art. 12 (prohibiting arbitrary interference with family); id. art. 16(1) (recognizing a right to found a family); Planned Parenthood of Se. Pa. v. Casey, 505 U.S. 833 (1992) (discussing the scope of a woman’s liberty to make child-bearing decisions); Prince v. Massachusetts, 321 U.S. 158, 165–67 (1944) (discussing parental freedom in the context of religious beliefs and conduct).

263 See supra Parts II.E, III.C.

264 See Bureau of Democracy, Human Rights, & Labor, supra note 259, at 13–14 (discussing China’s surveillance of human rights activists to control behavior).

265 See Ayala v. U.S. Att’y Gen., 605 F.3d 941, 948 (11th Cir. 2010). Private actors may have a right to inflict harm on an applicant in certain instances, such as self-defense.
critical when assessing harm committed in communities with cultures and practices that diverge from Western norms of acceptable behavior. Some cultures practice nonpunitive force-feeding while rites of passage in other communities include skin mutilation neither condoned nor penalized by the State. Whether a government or private actor perpetrates the harm in question, legitimacy will still be a necessary factor to consider. Thus, legitimacy represents one of the three pillars of a persecution definition that must also focus on the scope of harmful conduct and the severity of the harm inflicted.

V. DEFINITIONAL DESCRIPTORS WARRANTING ELIMINATION

Common conceptions of persecution include a number of descriptors that are inaccurate or irrelevant. The consequences of these flawed descriptors are not benign. Rather, conceptions of persecution contrary to its true tenets obscure its meaning and parameters. Part V will identify four descriptors that hinder an understanding of persecution. They are: references to threats to life or freedom, references to other elements of the refugee definition, labels equating persecution to offensive conduct, and the requirement that persecutors intend to punish their victims.

A. References to Threats to Life or Freedom

Life or freedom is a phrase employed in international instruments and domestic immigration statutes and regulations. In the Refugee Convention, Article 33 prohibits States from expelling a refugee “where his life or freedom would be threatened on account of” a protected ground. Under the INA, an applicant is entitled to withholding of removal when “life or freedom would be threatened in that country” on account of a protected ground. Similarly, U.S. immigration regulations couch the withholding of removal requirements under the rubric of life or freedom. The conduct that would threaten an applicant’s life or freedom has

But persecution assessments rarely confront the legitimacy of retaliatory harm because the motive for such retaliatory conduct rarely satisfies the nexus requirement.


Refugee Convention, supra note 146, art. 33(1). The one prescribed exception is where the refugee may constitute a danger to the country for certain reasons. See id. art. 33(2).


8 C.F.R. § 1208.16(a)–(b) (2013).
sometimes been distinguished from the type and level of harm required to establish persecution.\textsuperscript{270} In other instances, adjudicators have used the \textit{life or freedom} language interchangeably with persecution.\textsuperscript{271} The prevalence of the \textit{life or freedom} phrase within refugee law and the divergent interpretations the phrase has received raises a number of questions about its meaning and how it relates to the persecution requirement under the refugee definition. As this section will show, \textit{threats to life or freedom} is a phrase devoid of any meaning distinct from persecution. Moreover, it hinders a proper assessment of the true circumstances that warrant asylum relief or protection under the INA’s withholding of removal provisions—namely, protection against harm of a sufficient severity.

A discussion of the \textit{life or freedom} phrase requires an initial review of the distinctions between withholding of removal protection and asylum relief. Both require an applicant to establish that deportation will lead to harm in the home country. An entitlement to withholding of removal, however, is presumably more difficult to establish than asylum eligibility.\textsuperscript{272} Withholding of removal requires an applicant to prove that harm is more likely than not to occur, whereas asylum eligibility only requires a reasonable possibility of harm—that is, a fear that is “well-founded.”\textsuperscript{273} Additionally, while the INA only entitles applicants to withholding of removal “if the Attorney General decides that the alien’s life or freedom would be threatened in that country,”\textsuperscript{274} asylum applicants must establish a well-founded fear of \textit{persecution}.\textsuperscript{275} These distinctions are based on Congress’s effort to conform the INA to the Refugee Convention.\textsuperscript{276} The INA’s withholding of removal provisions are a codification of the nonrefoulement requirement in Article 33 of the Refugee Convention, which \textit{prohibits} States from deporting to the feared country applicants who satisfy the requisite standards.\textsuperscript{277} Asylum, by contrast, is concerned with Article 1 of the Refugee Convention, which provides a general definition of a “refugee” without any reference to \textit{life or freedom}.\textsuperscript{278} Because the benefits afforded to applicants granted asylum are greater than those applicable to

\textsuperscript{270} \textit{See}, e.g., Kobugabe v. Gonzales, 440 F.3d 900, 901 (7th Cir. 2006).
\textsuperscript{271} \textit{See}, e.g., Tamang v. Holder, 598 F.3d 1083, 1091 (9th Cir. 2010).
\textsuperscript{272} For a critique of U.S. Supreme Court decisions that affirmed the heightened standard for withholding of removal, see Hathaway & Cusick, supra note 23, at 485–88.
\textsuperscript{275} \textit{Id.} §§ 1101(a)(42), 1158(b)(1).
\textsuperscript{277} Refugee Convention, supra note 146, art. 33.
\textsuperscript{278} \textit{Id.} art. 1; \textit{see also} Refugee Convention Protocol, supra note 149, art. 1 (updating the refugee definition).
withholding of removal under the INA, the ultimate grant of asylum is discretionary.\textsuperscript{279}

Several scholars have put forward convincing arguments as to why the use of the phrase \textit{life or freedom} in the Refugee Convention was not intended to have a meaning distinct from the refugee definition in Article 1 of the Refugee Convention. Professor James Hathaway and Anne Cusick noted that Article 31 references refugees as individuals whose “life or freedom was threatened in the sense of Article 1.”\textsuperscript{280} Presumably then, \textit{life or freedom} would be synonymous with persecution because the refugee definition couches harm in terms of persecution.\textsuperscript{281} Paul Weis has noted numerous instances in drafts of the Refugee Convention where persecution and \textit{life or freedom} were used interchangeably.\textsuperscript{282} Thus, as Hathaway and Cusick pointed out, it appears that \textit{life or freedom} is merely “‘shorthand’ language . . . chosen simply in order to avoid the need to repeat the whole of the refugee definition,” a point driven home further by the fact that the verbose \textit{travaux préparatoires} do not mention any distinction between persecution and \textit{life or freedom}.\textsuperscript{283} The general discussions of the refugee and nonrefoulement provisions in the \textit{travaux préparatoires} led Professors Guy Goodwin-Gill and Jane McAdam to conclude that “[s]o far as the drafters of the [Refugee] Convention were aware of a divergence between the words defining refugee status and those requiring non-refoulement, they gave little thought to the consequences.”\textsuperscript{284}

If the intent of Congress was simply to conform the Refugee Act to the Refugee Convention, then there would appear no justification for viewing \textit{life or freedom} any differently than persecution. However, subsequent interpretations by the U.N. High Commissioner for Refugees (UNHCR) would appear to give \textit{life or freedom} a quasi-literal reading. In its highly influential refugee handbook, the UNCHR stated, “There is no universally accepted definition of ‘persecution,’” but subsequently explained that “it may be inferred that a threat to life or freedom on account of race, religion, nationality, political opinion or membership of a particular social group is always persecution.”\textsuperscript{285} At the very least, this explanation implies that there may be a class of harms outside of threats to life or freedom that may constitute persecution.

\textsuperscript{279} See 8 U.S.C. § 1158(b)(1)(A); see also Hathaway & Cusick, \textit{supra} note 23, at 485 (noting that the Refugee Convention drafters did not want to mandate that States afford applicants the generous benefits associated with refugee status).

\textsuperscript{280} Hathaway & Cusick, \textit{supra} note 23, at 523 (quoting Refugee Convention, \textit{supra} note 146, art. 31(1)).

\textsuperscript{281} See Refugee Convention, \textit{supra} note 146, art. 1.

\textsuperscript{282} \textit{THE REFUGEE CONVENTION, 1951}, at 303 (Paul Weis ed., 1995).

\textsuperscript{283} Hathaway & Cusick, \textit{supra} note 23, at 523–24; see also GOODWIN-GILL & MCADAM, \textit{supra} note 40, at 234 (noting that no distinction in standard of proof was made in the \textit{travaux préparatoires}); Joan Fitzpatrick, \textit{Revitalizing the 1951 Refugee Convention}, 9 HARV. HUM. RTS. J. 229, 235 n.30 (1996) (noting the very limited exceptions to Article 33’s obligatory nature).

\textsuperscript{284} GOODWIN-GILL & MCADAM, \textit{supra} note 40, at 233.

The U.S. Supreme Court has been more explicit. In *INS v. Stevic*, the Court stated in dicta that persecution is a “seemingly broader concept than threats to ‘life or freedom.’” In decisions subsequent to *Stevic*, the federal courts of appeals and the Board regularly (and unsurprisingly) adhere to the express language of the withholding of removal provisions of the INA, steadfastly describing the conduct that needs to be shown as a threat to life or freedom. Some courts have gone even further by noting explicitly that an entitlement to withholding of removal is harder to establish than eligibility for refugee relief for reasons that include the distinction between persecution and threats to life or freedom. This has led courts to state that persecution must entail “more than threats to life and freedom.” The *life or freedom* language has bled into adjudicators’ understanding of persecution in more creative ways. The Third Circuit, for example, characterizes persecution as “‘threats to life, confinement, torture, and economic restrictions so severe that they constitute a threat to life or freedom.’” Conversely, some courts have found that persecution and threats to life or freedom are the same.

Whether *life or freedom* is believed to be synonymous with persecution or merely a component of its framework, its deleterious effects on an understanding of persecution’s meaning remains the same. Regardless of whether the drafters intended persecution and *life or freedom* to encompass different treatment, no meaningful distinction can actually apply in practice without leading to absurd or anomalous results. When evaluating the *life or freedom* rubric descriptively, it becomes clear that it is not an accurate way to conceptualize the harm needed to establish persecution. This is because the *threats to life or freedom* description is both overinclusive and underinclusive of the types of harms that would qualify as persecution—and should qualify as sufficient for withholding of removal protection.

---

287 Id. at 428 n.22.
288 See Tamang v. Holder, 598 F.3d 1083, 1091 (9th Cir. 2010); In re Acosta, 19 I. & N. Dec. 211, 230 (B.I.A. 1985).
289 See Kobugabe v. Gonzales, 440 F.3d 900, 901 (7th Cir. 2006); see also Qiu v. Ashcroft, 329 F.3d 140, 148 (2d Cir. 2007), overruled on other grounds by Shi Liang Lin v. U.S. Dep’t of Justice, 494 F.3d 296 (2d Cir. 2007) (en banc) (perceiving a showing of a threat to life or freedom as an additional prerequisite to withholding apart from establishing eligibility for asylum).
290 Ivanishvili v. U.S. Dep’t of Justice, 433 F.3d 332, 341 (2d Cir. 2006) (citation omitted) (internal quotation marks omitted)); accord Li v. Gonzales, 405 F.3d 171, 177 (4th Cir. 2005); Dandan v. Ashcroft, 339 F.3d 567, 573 (7th Cir. 2003); Aguilar-Solis v. INS, 168 F.3d 565, 569–70 (1st Cir. 1999).
291 Li v. Att’y Gen., 400 F.3d 157, 166–67 (3d Cir. 2005) (quoting Fatin v. INS, 12 F.3d 1233, 1240 (3d Cir. 1993)). The circular “threat to life” language would not make sense unless it is based on the distinction between severity and probability.
292 See Fisher v. INS, 291 F.3d 491, 497 (8th Cir. 2002) (“Persecution involves a threat to one’s life or freedom . . . .”).
Threats to life or freedom is overinclusive of the types of conduct warranting relief and protection because it implies that any deprivation to life or freedom would be sufficient. This, of course, is not the case because persecution requires harm of sufficient severity. Some of the harms that fall under the life or freedom description would constitute per se persecution, such as the “inherent right to life.” Freedom may also encompass harm that would necessarily constitute persecution depending on how it is defined. Some interpretations of freedom have limited it to physical restraints, while others have equated freedom with a broader amalgamation of rights afforded under human rights norms. Nevertheless, either interpretation would lead to the inclusion of certain conduct that does not establish persecution per se, and thus renders life or freedom overinclusive.

The extent of life or freedom’s underinclusiveness is also influenced by the perceived scope of freedom, but both a narrow and broad reading of the term would demonstrate some level of disconnect. The case of Mirzoyan v. Gonzales helps to illustrate. The applicant was denied admission to a prestigious college, could not find a job in her field, and was eventually fired from the unskilled courier position she was forced to take. The Second Circuit concluded that these interferences with her economic opportunities “might constitute a ‘substantial economic disadvantage,’ but her treatment is unlikely to constitute a ‘threat to [her] life or freedom.’” The reviewing body in Mirzoyan was confronted with a circumstance where the applicant experienced harm that might qualify as economic persecution even though the type of harm experienced did not neatly fall within the threat to life or freedom rubric—at least according to the court. But it serves no purpose to parse the nature of the harm involved to determine if it falls under a life or freedom conception that distracts attention from persecution’s harm-based core. Since economic deprivations are recognized as an actionable type of harm, the only question should be whether they are sufficiently severe to qualify for relief or protection against deportation.

The disconnect between the life or freedom language and the proper focus on harm has caused reviewing bodies to implicitly recognize the uselessness of life or

293 See Butt v. Keisler, 506 F.3d 86, 90 (1st Cir. 2007); Gilaj v. Gonzales, 408 F.3d 275, 284–85 (6th Cir. 2005).
294 ICCPR, supra note 41, art. 6(1). But see id. art. 6(2) (limiting the application of the death penalty to the “most serious crimes”).
295 See Marquez v. INS, 105 F.3d 374, 379 (7th Cir. 1997) (equating lack of freedom with imprisonment).
296 See Popova v. INS, 273 F.3d 1251, 1260–61 (9th Cir. 2001) (equating freedom with liberty).
297 457 F.3d 217 (2d Cir. 2006).
298 Id. at 219.
299 Id. at 223; see also In re T-Z-, 24 I. & N. Dec. 163, 170–71 (B.I.A. 2007) (discussing the perceived difference in the two definitions).
300 See Guan Shan Liao v. U.S. Dep’t of Justice, 293 F.3d 61, 67 (2d Cir. 2002); In re T-Z-, 24 I. & N. Dec. at 171.
freedom as a distinct concept. This implicit recognition can be seen in a macroevaluation of immigration opinions over time, where the vernacular for discussing harm swivels back and forth from life or freedom to persecution. Thus, courts regularly note that an applicant is eligible for withholding of removal if the applicant can establish a “probability of persecution.” Numerous decisions have concluded that a claim for withholding of removal automatically fails when the applicant does not establish asylum eligibility. The denial of withholding of removal would not be automatic unless life or freedom and persecution were perceived to entail synonymous inquiries. Consequently, even though reviewing bodies may hollowly provide the distinctive language between life or freedom and persecution, the decisions largely view the distinction between withholding of removal and asylum as nothing more than a different probability standard.

In short, refugee claims should be evaluated by looking at the severity of harm. The life or freedom phrase was simply a shorthand way for drafters of the Refugee Convention to refer to conduct closely akin to persecution. And regardless of whether the drafters intended that life or freedom have a meaning distinct from persecution, in practice it does not. In fact, life or freedom is not an accurate phrase for assessing harm because it is both overinclusive and underinclusive of the conduct warranting relief and protection. Adjudicators should recognize its hollow nature and refrain from using it as a literal gauge for evaluating withholding of removal claims, a means to distinguish withholding of removal and asylum, or as a factor in the persecution assessment itself.

B. References to Other Refugee Prerequisites

Decisionmakers needlessly obscure the meaning of persecution by including additional aspects of the refugee definition within the scope of persecution descriptions. The Fourth and Eighth Circuits label persecution as the “infliction or threat of death, torture, or injury to one’s person or freedom, on account of one of the enumerated grounds in the refugee definition.” In the Second Circuit, the

---

301 See, e.g., Tamang v. Holder, 598 F.3d 1083, 1091 (9th Cir. 2010); In re Acosta, 19 I. & N. Dec. 211, 229 (B.I.A. 1985).
302 See, e.g., Pavlovich v. Gonzales, 476 F.3d 613, 619 (8th Cir. 2007); Sanchez v. U.S. Att’y Gen., 392 F.3d 434, 437 (11th Cir. 2004); Korablina v. INS, 158 F.3d 1038, 1045–46 (9th Cir. 1998).
303 See, e.g., Bhosale v. Mukasey, 549 F.3d 732, 736 (8th Cir. 2008); Camara v. Ashcroft, 378 F.3d 361, 367 (4th Cir. 2004).
304 See Wakkary v. Holder, 558 F.3d 1049, 1053 (9th Cir. 2009) (“[W]ithholding of removal under 8 U.S.C. § 1231(b)(3) provides relief to applicants who fear persecution according to the same substantive criteria as asylum, but with a higher standard of objective reasonableness . . . .”).
305 Li v. Gonzales, 405 F.3d 171, 177 (4th Cir. 2005) (quoting Kondakova v. Ashcroft, 383 F.3d 792, 797 (8th Cir. 2004)); Rife v. Ashcroft, 374 F.3d 606, 612 (8th Cir. 2004); see also Regalado-Garcia v. INS, 305 F.3d 784, 787 (8th Cir. 2002) (listing each of the five protected grounds in its formulation of the persecution definition).
court has defined persecution as “the infliction of suffering or harm upon those who differ on the basis of a protected statutory ground.”

The nexus requirement, however, is a prong of the refugee definition separate and apart from persecution. Including it in the persecution definition diverts attention from persecution’s harm-based core.

The Seventh Circuit has also folded the nexus requirement language into the persecution definition, but it has done so in a manner that further obscures the meaning of legitimacy. In Zhou Ji Ni v. Holder, for example, the court stated that persecution encompasses certain harmful conduct inflicted “for political, religious, or other reasons that this country does not recognize as legitimate.” Because political opinion and religion are two of the five protected grounds required to establish eligibility for refugee relief, the Seventh Circuit is presumably using the notion of legitimacy as a shorthand reference to the nexus requirement. Legitimacy, however, has a distinct meaning and purpose in the persecution inquiry. It serves as a gauge for assessing when severe harmful conduct nevertheless fails to establish persecution. Thus, the use of the legitimacy language to refer to the nexus requirement further compounds the problems caused by any reference to the nexus requirement in a persecution description.

A final incorrect overlap can be seen in the Fifth Circuit’s description of persecution. The court has stated that persecution requires that the inflicted harm occur “under government sanction.” As with the nexus requirement, the refugee definition includes a separate government involvement component, namely that the government must participate in the harm suffered by the applicant or be unable or unwilling to control it. Separating persecution from the government involvement prong is particularly critical because of the distinct roles government action (or inaction) plays as a separate element in the refugee definition on the one hand, and as a factor relevant to assessing harm on the other.

---

306 Ivanishvili v. U.S. Dep’t of Justice, 433 F.3d 332, 341 (2d Cir. 2006); see also Mohammed v. Keisler, 507 F.3d 369, 371 (6th Cir. 2007) (“Persecution is not statutorily defined, but this Court has shaped some of its contours. . . . [T]he context must indicate that the asylum applicant is targeted for abuse based on his membership in a protected category.”).


308 635 F.3d 1014 (7th Cir. 2011).

309 Id. at 1017 (quoting Tama-Mercea v. Reno, 222 F.3d 417, 424 (7th Cir. 2000)); accord Khan v. Filip, 554 F.3d 681, 690 (7th Cir. 2009); Bace v. Ashcroft, 352 F.3d 1133, 1137 (7th Cir. 2003).


311 See supra Part IV (discussing the meaning and application of legitimacy in the persecution inquiry).

312 Eduard v. Ashcroft, 379 F.3d 182, 187 (5th Cir. 2004); Mikhail v. INS, 115 F.3d 299, 303 n.2 (5th Cir. 1997); Abdel-Masieh v. INS, 73 F.3d 579, 583 (5th Cir. 1996).

313 See Gao v. Ashcroft, 299 F.3d 266, 272 (3d Cir. 2002); Navas v. INS, 217 F.3d 646, 655–56 (9th Cir. 2000).

314 See supra Part II.D.4 (discussing the role of government action or abdication in a persecution assessment). The Fifth Circuit’s belief that harm must occur “under
differentiate between the overarching elements of the refugee definition, then there is little chance they will recognize and evaluate properly the analytical subtleties within particular elements such as persecution.

C. Offensiveness

Some of the most prominent and widely used definitions of persecution incorporate some analogy between persecution and offensive conduct. The offensiveness language was first introduced to the persecution definition in 1969 when the Ninth Circuit defined persecution as “the infliction of suffering or harm upon those who differ (in race, religion, or political opinion) in a way regarded as offensive.”315 When legislators debated whether to include a definition of persecution in the Refugee Act, they ultimately declined to do so under the premise that the Ninth Circuit’s description of persecution was well established.316 The Board subsequently incorporated the offensiveness language into its understanding of persecution with approval in the seminal case of In re Acosta,317 after concluding that the Refugee Act did not intend to alter the scope of persecution.318 Following the Board’s decision, the Fifth and Tenth Circuits adopted the Ninth Circuit’s description of persecution.319

315 Kovac v. INS, 407 F.2d 102, 107 (9th Cir. 1969) (citing WEBSTER’S THIRD NEW INTERNATIONAL DICTIONARY 1685 (1965)).
319 See, e.g., Karouni v. Gonzales, 399 F.3d 1163, 1170 (9th Cir. 2005); Chaib v. Ashcroft, 397 F.3d 1273, 1277 (10th Cir. 2005); Wiransane v. Ashcroft, 366 F.3d 889, 893 (10th Cir. 2004); Abdel-Masih v. U.S. INS, 73 F.3d 579, 583–84 (5th Cir. 1996); Ghaly v. INS, 58 F.3d 1425, 1431 (9th Cir. 1995); cf. Li v. Gonzales, 405 F.3d 171, 182 (4th Cir. 2005) (Gregory, J., dissenting) (providing a similar definition of persecution as the Ninth Circuit). Although they do not regularly define persecution in this way, several other circuit courts have done so on occasion. See, e.g., United States v. Geiser, 527 F.3d 288, 295 (3d Cir. 2008); Anton v. INS, 50 F.3d 469, 472 (7th Cir. 1995); Zalega v. INS, 916 F.2d 1257, 1260 (7th Cir. 1990). The Second Circuit has gone so far as to eliminate the “offensive” language altogether and simply refers to a protected ground. In Ivanishvili v. U.S. Department of Justice, 433 F.3d 332 (2d Cir. 2006), for example, the court paraphrased the above offensive language to read that “persecution is the infliction of suffering or harm upon those who differ on the basis of a protected statutory ground.” Id. at 341 (basing its definition on the definition provided by the Board in In re S-A-, 22 I. & N. Dec. at 1336, which proffered persecution as “the infliction of suffering or harm upon those who differ (in race, religion or political opinion) in a way regarded as offensive”). The persecution definition originally stated in Kovac has been contorted in several other ways as well. See,
The meaning of “offensive” in these descriptions is open to two interpretations; neither interpretation is consistent with the actual scope of persecution. The first interpretation is that the definition is simply clarifying that conduct qualifies as offensive when it is inflicted on account of one of the five protected grounds—that is, race, religion, ethnicity, political opinion, or social group.\textsuperscript{320} In this respect, the offensiveness of conduct is simply an expression of the motives persecutors employ to inflict harm that the Refugee Convention and Refugee Act deem worthy of protection. The nexus requirement, however, should not appear in a persecution definition for the reasons already discussed.\textsuperscript{321}

In addition to analogizing between offensive conduct and the nexus requirement, adjudicators have also equated offensiveness with harm severe enough to qualify as persecution. In \textit{Singh v. INS},\textsuperscript{322} for example, the Ninth Circuit stated that “whether discrimination, harassment, or violence directed at a particular group on account of a protected ground is sufficiently offensive to constitute persecution under the Act must be decided on a case-by-case basis.”\textsuperscript{323} Thus, under this interpretation, there is a continuum of offensive conduct that at some point crosses a threshold whereby the conduct is severe enough to constitute persecution. Devoid of context, this is a somewhat reasonable point, for in the absence of a legitimate basis for the harm,\textsuperscript{324} the infliction of any harm would offend moral sensibilities. But this is not what the persecution description actually denotes. Rather, the description \textit{equates} persecution to offensive conduct, which disregards the need to establish harm of sufficient severity.\textsuperscript{325} Accordingly, adjudicators’ use of offensive conduct as a threshold to establish persecution is an incorrect characterization that should be omitted.

\textsuperscript{320} See Duarte de Guinac v. INS, 179 F.3d 1156, 1162 (9th Cir. 1999) (noting that the victim “possessed a particular ‘offensive’ characteristic”).

\textsuperscript{321} See supra Part V.B.

\textsuperscript{322} 94 F.3d 1353 (9th Cir. 1996).

\textsuperscript{323} \textit{Id.} at 1359; see also Witjaksono v. Holder, 573 F.3d 968, 977 (10th Cir. 2009) (“Verbal taunts, while offensive, fall within the bounds of harassment and discrimination, not persecution.”).

\textsuperscript{324} See supra Part IV (discussing legitimacy).

\textsuperscript{325} Consequently, the meaning of offensive conduct in this description differs from the appropriate use of offensiveness as a severity gauge discussed \textit{supra} Part III.A. Offensive conduct under the interpretation discussed in Part III does not equate offensive conduct to persecution, whereas this interpretation does. The incongruence between the two interpretations further compounds the confusion caused by the offensiveness language.
D. Punishment

1. Intent to Punish as a Prerequisite to Persecution

The Fifth Circuit has affirmatively incorporated an intent to punish requirement into its conception of persecution, stating that persecution requires "‘a showing that harm or suffering will be inflicted upon [the applicant] in order to punish [the applicant] for possessing a belief or characteristic a persecutor sought to overcome.’”\(^{326}\) Conversely, the Ninth Circuit has repeatedly rejected punishment as a prerequisite.\(^{327}\) Its reasons for doing so were explored in detail in *Pitcherskaia v. INS*.\(^{328}\) In *Pitcherskaia*, the court stated that "‘punishment’ is neither a mandatory nor a sufficient aspect of persecution” because "punishment implies that the perpetrator believes the victim has committed a crime or some wrong; whereas persecution simply requires that the perpetrator cause the victim suffering or harm."\(^{329}\) The court observed, “Human rights laws cannot be sidestepped by simply couching actions that torture mentally or physically in benevolent terms such as ‘curing’ or ‘treating’ the victims.”\(^{330}\)

The Ninth Circuit’s analysis stands on much stronger footing. Requiring intent to punish would fail to provide protection to victims of atrocious harm who otherwise satisfy the statutory requirements of the refugee definition. It would also cut against the pliable understanding of refugee status espoused by the U.N. High Commissioner for Refugees.\(^{331}\) Certainly the refugee definition cannot be circumvented simply to encompass sympathetic (but otherwise ineligible) cases. Nevertheless, the perceived scope of conduct that qualifies as persecution should be informed by the decades of circumstances that have evolved our understanding of the term since the birth of modern refugee law. As the Ninth Circuit noted in *Pitcherskaia*, “The fact that a persecutor believes the harm he is inflicting is ‘good for’ his victim does not make it any less painful to the victim, or, indeed, remove the conduct from the statutory definition of persecution.”\(^{332}\)

\(^{326}\) Faddoul v. INS, 37 F.3d 185, 188 (5th Cir. 1994) (quoting Guevara Flores v. INS, 786 F.2d 1242, 1249 (5th Cir. 1986)); accord Kane v. Holder, 581 F.3d 231, 238 (5th Cir. 2009); Zhao v. Gonzales, 404 F.3d 295, 307–08 (5th Cir. 2005). But see Eduard v. Ashcroft, 379 F.3d 182, 187 (5th Cir. 2004) (providing an alternative definition of persecution).

\(^{327}\) See Gormley v. Ashcroft, 364 F.3d 1172, 1177 (9th Cir. 2004); Pitcherskaia v. INS, 118 F.3d 641, 647 (9th Cir. 1997).

\(^{328}\) 118 F.3d 641.

\(^{329}\) Id. at 647–48.

\(^{330}\) Id. at 648.

\(^{331}\) See, e.g., U.N. High Comm’r for Refugees, Guidelines on International Protection: “Membership of a Particular Social Group” within the Context of Article 1A(2) of the 1951 Convention and/or its 1967 Protocol Relating to the Status of Refugees, ¶ 3, U.N. Doc. HCR/GIP/02/02 (May 7, 2002) (stressing the importance of assessing the meaning of a particular social group in an “evolutionary manner”).

\(^{332}\) Pitcherskaia, 118 F.3d at 648.
instances of excruciating suffering that, even assuming the existence of a nexus to a protected ground and the requisite government involvement, would not qualify the applicant for refugee status if an intent to punish was a prerequisite. For example, in some regions of Africa and Asia, village elders perform genital mutilation on young females in the community. Anesthesia is not typically administered prior to the painful procedure, and victims can suffer a range of horrible complications. Despite the severe pain these elders inflict, their motive for carrying out the procedure is not ordinarily grounded in an intent to punish the young women. Rather, their motivations focus on the preservation of virginity before marriage and fidelity during marriage.

While a specific intent to punish may increase the psychological suffering of the victim, there is no legitimate reason why it should serve as a definitive threshold. Indeed, most appellate bodies implicitly recognize the shortcoming of a punishment cutoff by virtue of their acceptance of FGM as a viable form of persecution. Moreover, none of the appellate bodies that require an intent to punish have ever proffered a reasonable justification. To the contrary, the Board decisions that originally led the Fifth Circuit to incorporate punishment into its interpretation of persecution have been mitigated by subsequent Board decisions. In the seminal case *In re Acosta*, the Board held that establishing persecution requires a showing that “harm or suffering . . . be inflicted upon an individual in order to punish him for possessing a belief or characteristic a persecutor sought to overcome.”

---


334 See *Bah*, 529 F.3d at 102; FOUND. FOR WOMEN’S HEALTH, RESEARCH & DEV., *FEMALE GENITAL MUTILATION: INFORMATION PACK* 4 (2002).

335 See supra Part II.D.3.

336 See *Bah*, 529 F.3d at 112; Niang v. Gonzales, 492 F.3d 505, 510 (4th Cir. 2007); Agbor v. Gonzales, 487 F.3d 499, 502 (7th Cir. 2007); Hassan v. Gonzales, 484 F.3d 513, 517 (8th Cir. 2007); Niang v. Gonzales, 422 F.3d 1187, 1197 (10th Cir. 2005); Mohammed v. Gonzales, 400 F.3d 785, 796 (9th Cir. 2005); Toure v. Ashcroft, 400 F.3d 44, 49 n.4 (1st Cir. 2005); Abay v. Ashcroft, 368 F.3d 634, 638 (6th Cir. 2004); see also *In re A-T*, 24 I. & N. Dec. 296, 296–97 (B.I.A. 2007); *In re Kasinga*, 21 I. & N. Dec. 357, 365–67 (B.I.A. 1996) (en banc).


338 786 F.2d 1242 (5th Cir. 1986).
ever since. In the interim, however, factual circumstances arose in cases appearing before the Board that led it to reject the need to establish punishment to sustain a finding of persecution. In *In re Kasinga*, the Board held that “‘punitive’ or ‘malignant’ intent is not required for harm to constitute persecution,” and that formulations of persecution requiring punishment have their “antecedents in concepts of persecution that predate the Refugee Act of 1980.” The Board has not subsequently issued an opinion that so forcefully disavows the punishment requirement, but the analytical shortcomings of the punishment prerequisite are evident in the Board’s subsequent body of jurisprudence, which rarely discusses punishment in such an outcome determinative manner. Consequently, the flaws in the punishment threshold, the failure of adjudicators to justify a punishment prerequisite, and reviewing bodies’ express and implicit rejection of a required intent to punish, all support eliminating punishment as a prerequisite to establishing persecution.

2. Punishment as Distinct from Harm

In contrast to some adjudicatory bodies’ belief that persecution requires punitive intent, several appellate courts view punishment as one of the vehicles that may establish persecution. The Seventh Circuit describes persecution as “punishment or the infliction of harm for political, religious, or other reasons that this country does not recognize as legitimate.” The Sixth Circuit has similarly held that persecution “entails ‘punishment or the infliction of suffering or harm,’” although it more often incorporates punishment into its persecution definition by phrasing it in negative terms. In *Mikhailevitch v. INS*, for example, the court stated that persecution “requires more than a few isolated incidents of

---

339 *Id.* at 1249. On occasion, the Fifth Circuit has tried to temper its punishment requirement. See *Arif v. Mukasey*, 509 F.3d 677, 680–81 (5th Cir. 2007) (stating that the punishment prerequisite is “generally require[d]”).


341 *Id.* at 365. Strangely, in support of its contention that punishment is not a prerequisite of persecution, the Board cited to *In re Acosta*, see *id.*, a case that explicitly stated that punishment is a prerequisite to a finding of persecution, see *In re Acosta*, 19 I. & N. Dec. at 222; see also *Pitcherskaia v. INS*, 118 F.3d 641, 647–48 (9th Cir. 1997) (recognizing and rejecting the explicit incorporation of punishment into the persecution definition by the Board in *In re Acosta*).


343 *Zhou Ji Ni v. Holder*, 635 F.3d 1014, 1017 (7th Cir. 2011); *Toure v. Holder*, 624 F.3d 422, 428 (7th Cir. 2010); *Bace v. Ashcroft*, 352 F.3d 1133, 1137 (7th Cir. 2003).


345 146 F.3d 384 (6th Cir. 1998).
verbal harassment or intimidation, unaccompanied by any physical punishment, infliction of harm, or significant deprivation of liberty.\textsuperscript{346}

Two reasons caution against the Sixth and Seventh Circuits’ incorporation of punishment into their persecution descriptions. Primarily, “punishment” and “infliction of harm” do not have semantic boundaries distinguishable in any meaningful way, much less a way that could contribute to an understanding of the conduct that could qualify as persecution. By defining persecution as “punishment or the infliction of harm,” the courts are stating that punishment should be understood to mean something other than the infliction of harm. It could refer to the intent of the persecutor to punish,\textsuperscript{347} which would obscure an accurate definition by listing in the disjunctive a frame of mind that provides no information about the conduct needed to establish persecution. An alternative interpretation would entail some separate class of conduct distinct from “harm.” But “harm” is a generic term that encompasses the full spectrum of conduct probative of persecution, as seen in the above-discussed taxonomy of harm.\textsuperscript{348} Consequently, neither possible interpretation of punishment succeeds.

The second reason to question the persecution definitions in the Sixth and Seventh Circuits is the courts’ failure to justify how punishment fits into a persecution assessment. Rather, the courts appear to have incorporated punishment accidentally, by virtue of word choices not grounded in any careful consideration of their semantic significance. In its 1990 opinion \textit{Zalega v. INS},\textsuperscript{349} the Seventh Circuit defined persecution as “the infliction of suffering or harm upon those who differ (in race, religion, or political opinion) in a way regarded as offensive.”\textsuperscript{350} One year later, in \textit{Osaghae v. INS},\textsuperscript{351} the court cited \textit{Zalega} for the proposition that persecution means “punishment for political, religious, or other reasons that our country does not recognize as legitimate.”\textsuperscript{352} The \textit{Osaghae} opinion did not provide any justification for its new conclusion that punishment is essential to a finding of persecution. To the contrary, the new definition appears to be nothing more than a perceived paraphrase of definitions offered in earlier cases.

The Seventh Circuit’s determination that persecution requires an intent to punish did not last beyond \textit{Osaghae}. Subsequent cases combined \textit{Osaghae}’s punishment language with previous alternate definitions to derive the definition currently employed, namely that persecution requires a showing of “punishment or

\textsuperscript{346} Id. at 390; accord Vincent v. Holder, 632 F.3d 351, 356 (6th Cir. 2011); Pilica v. Ashcroft, 388 F.3d 941, 950 (6th Cir. 2004).
\textsuperscript{347} See supra Part V.D.1 (reviewing the viability of a punitive intent prerequisite); see also Faddoul v. INS, 37 F.3d 185, 188 (5th Cir. 1994) (construing persecution to require punitive intent).
\textsuperscript{348} See supra Part II. For the same reason, the definition would not succeed if “punishment” is meant to refer to the legitimacy component of the persecution definition.
\textsuperscript{349} 916 F.2d 1257 (7th Cir. 1990).
\textsuperscript{350} Id. at 1260.
\textsuperscript{351} 942 F.2d 1160 (7th Cir. 1991).
\textsuperscript{352} Id. at 1163.
the infliction of harm.”

When incorporating punishment into its understanding of persecution, the Sixth Circuit relied on the definition created by the Seventh Circuit and its mistaken belief that the Ninth Circuit viewed punishment as relevant to a persecution assessment. Thus, the inclusion of punishment in the persecution definitions in both circuits stems from the same deficiency. Whether incorporated by accident or intentional design, punishment, in and of itself, is still not a viable means to establish persecution.

VI. PERSECUTION DEFINED

Incorporating the principles discussed in the Article and eliminating those aspects unhelpful, contradictory, or detrimental to a coherent definition, this Article proposes that persecution should be defined as the illegitimate infliction of sufficiently severe harm. Context distinguishes persecution from otherwise harmful conduct. In the same way that harmful conduct can become a rights deprivation when assessed from a human rights vantage point, so too does illegitimate severe harm become persecutory conduct when assessed within the context of refugee law. To otherwise incorporate into the persecution definition other aspects of the refugee definition that would expressly tie it to persecution would dilute and obscure the true nature of persecution’s harm-based core.

Harmful conduct includes physical harms, restraints and deprivations of privacy, resource and opportunity limitation, psychological harms, and infringements on human rights standing alone or attendant the previously mentioned categories. The severity of the harmful acts in question and the harmful effects such actions have on the applicant should be assessed within the context of the applicant’s continuous experience in the home country (or the anticipated experience if deported). Guiding the severity inquiry is the recognition that persecution does not entail all treatment that can be viewed as offensive, unfair, unjust, or even unlawful or unconstitutional. Harassment, by its ordinary meaning, is but one example of conduct that fails to cross the severity threshold. While the discriminatory nature of a harmful act may itself add to

353 See Sivaainkaran v. INS, 972 F.2d 161, 164 n.2 (7th Cir. 1992).
354 Mikhailievich v. INS, 146 F.3d 384, 389–90 (6th Cir. 1998); see also Pitcherskaia v. INS, 118 F.3d 641, 647 (9th Cir. 1997) (disavowing punishment as a precondition to establishing persecution).
355 Compare HATHAWAY, supra note 11, at 104–05 (defining persecution as “sustained or systemic violation of basic human rights demonstrative of a failure of state protection”), with Minister of Immigration & Multicultural Affairs v Khawar (2002) 210 CLR 1, 37 (Austl.) (“[The Refugee Convention] is one of several important international treaties designed to redress violation[s] of basic human rights, demonstrative of a failure of state protection.”).
356 See supra Part II.
357 See supra Part III.B–C.
358 See supra Part III.A.1.
359 See supra Part III.A.2.a.
suffering, ordinarily the severity of harm should be based on measuring the consequences of the discriminatory conduct.\textsuperscript{360} Even if harm reaches the requisite level of severity, an applicant cannot establish persecution if the harmful conduct was, or will be, legitimately inflicted.\textsuperscript{361} By contrast, the definition of persecution should not include unhelpful or incorrect qualifiers, such as an incorporation of the \textit{threats to life or freedom} language, references to other elements of the refugee definition, gauges linking offensive conduct to persecution, or a requirement that a persecutor inflict a harm for punitive reasons.\textsuperscript{362}

\textbf{VII. Conclusion}

For nearly a century, the international community has sought to afford refugee protection to the millions of individuals forced to flee their homelands.\textsuperscript{363} While the reasons for offering protection are multifaceted,\textsuperscript{364} the United States recognized the importance of ratifying the Refugee Convention Protocol and adopting the Refugee Act.\textsuperscript{365} Affording refugee protection is indeed important, for States do not (or cannot) always protect their citizens against the numerous measures groups and individuals take to harm their fellow humans.\textsuperscript{366} But an understanding of what persecution means drives the circumstances under which refugee applicants warrant protection. This Article sought to advance an understanding of persecution and how decisionmakers responsible for deciding the fates of prospective refugees should conceive of the term. The definition advanced hones persecution’s core while eliminating descriptors that are peripheral to its focus. By providing a deeper understanding of the meaning and parameters of this “elusive”\textsuperscript{367} and “protean”\textsuperscript{368} term, this Article hopes to aid decisionmakers charged with the fateful task of determining when to extend the welcoming arms of the United States and when to clamp them shut. Persecution decisions will always involve some level of subjective judgment. But whatever intuitions and beliefs are brought to the evaluative table, they should be grounded in a concordant understanding of what it means to be persecuted.

\textsuperscript{360} \textit{See supra} Part III.A.2.b.
\textsuperscript{361} \textit{See supra} Part IV.
\textsuperscript{362} \textit{See supra} Part V.
\textsuperscript{363} \textit{See} HATHAWAY, \textit{supra} note 11, at 2.
\textsuperscript{364} \textit{See id.}
\textsuperscript{366} \textit{See} Ramji-Nogales et al., \textit{supra} note 18, at 379 (“[T]he nature of persecution and our understanding of it keep changing.”).
\textsuperscript{367} Pathmakanthan v. Holder, 612 F.3d 618, 622 (7th Cir. 2010).
\textsuperscript{368} Bocova v. Gonzales, 412 F.3d 257, 263 (1st Cir. 2005).
THE SKEPTIC’S GUIDE TO INFORMATION SHARING AT SENTENCING

Ryan W. Scott*

ABSTRACT

The “information sharing model,” a leading method of structuring judicial discretion at the sentencing stage of criminal cases, has attracted broad support from scholars and judges. Under this approach, sentencing judges should have access to a robust body of information, including written opinions and statistics, about previous sentences in similar cases. According to proponents, judges armed with that information can conform their sentences to those of their colleagues or identify principled reasons for distinguishing them, reducing inter-judge disparity and promoting rationality in sentencing law.

This Article takes a skeptical view of the information sharing model, arguing that it suffers from three fundamental weaknesses as an alternative to other structured sentencing reforms. First, there are information collection challenges. To succeed, the model requires sentencing information that is written, comprehensive, and representative. Due to acute time constraints, however, courts cannot routinely generate that kind of information. Second, there are information dissemination challenges. Sharing sentencing information raises concerns about the privacy of offenders and victims. Also, the volume and complexity of sentencing decisions create practical difficulties in making relevant information accessible to sentencing judges. Third, the model’s voluntariness is an important drawback. The information sharing model rests on the heroic assumption that judges will respond to information about previous sentences by dutifully following the decisions of their colleagues. That is unrealistic. Judges just as easily can disregard the information, ignore it, or even move in the opposite direction.

Despite those grounds for skepticism, information sharing can play a valuable role as a supplement to other sentencing reforms. In particular, information sharing would benefit from a system of

* © 2013 Ryan W. Scott. Associate Professor, Indiana University Maurer School of Law, Bloomington. Thanks to Amy Baron-Evans, Craig Bradley, Brian Broughman, Anuj Desai, Richard Frase, Nancy Gertner, Cecilia Klingele, Ion Meyn, Marc Miller, Ellen Podgor, Kevin Reitz, Meghan Ryan, Michelle Spak, Kate Stith, Lua Yuille, and participants in faculty workshops at Yale Law School, SMU Dedman School of Law, the University of Wisconsin Law School, and the University of Cincinnati College of Law for valuable comments on earlier drafts. Thanks as well to Benjamin Hugon and Daniel Bradley for outstanding research assistance, and to Christopher Kozelichki for assistance with the empirical study.
sentencing guidelines, whether mandatory or advisory, and from open access to the information on the part of defense counsel and prosecutors.

INTRODUCTION

For decades, prominent scholars and judges in the United States have proposed an “information sharing model” for structuring criminal sentencing decisions. In indeterminate sentencing systems, which prevailed throughout the United States until the 1970s, judges enjoyed broad and essentially unchecked discretion to select the appropriate punishment for criminal offenses. With broad statutory ranges, no appellate review, and no obligation to give reasons for their decisions, judges largely were left to their own intuitions in choosing a sentence. One consequence was stark inter-judge disparity. Similarly situated offenders stood to receive widely disparate sentences depending on the values, preferences, and biases of the sentencing judge. Another was that no rational and principled body of sentencing law could develop. To address those weaknesses, reformers in the 1970s and 1980s proposed various methods of structuring sentencing decisions.

One leading approach, which this Article will describe as the information sharing model, has attracted strong support among scholars and is experiencing something of a renaissance. The idea is that judges should have access to a robust store of information about previous sentences in similar cases. Armed with statistical data, details about past offenses and offenders, and written opinions from similar cases, sentencing judges can achieve better results. Information sharing will promote inter-judge consistency and rationality, the argument goes, because judges who understand the reasons for previous sentences can conform to them or identify principled points of distinction. A distinguished and varied group of scholars and judges has endorsed some form of information sharing at sentencing, including Professor Marc Miller, Justice Michael Wolff, Professor Kate Stith and Judge José Cabranes, Judge Nancy Gertner, and Judge Robert Sweet.

The information sharing model is frequently advanced as an alternative to more intrusive forms of structured sentencing, such as sentencing guidelines. “Sentencing information systems” and other forms of electronic data sharing

---

already form a crucial component of the sentencing process in some U.S. states and in jurisdictions overseas. Missouri, for example, has attracted national attention for its information-based discretionary sentencing system, which strives to equip sentencing judges with better data about previous outcomes in similar cases. And just last year, Ireland’s criminal courts launched an ambitious sentencing information system as a national pilot project. Meanwhile, mounting interest in evidence-based sentencing—using tools like risk assessment instruments—has highlighted the need for a better information sharing infrastructure for sentencing judges.

Surprisingly, however, the literature rarely grapples with basic questions about the information sharing model. Is information sharing at sentencing feasible? Can courts and other stakeholders in the criminal justice system effectively collect, disseminate, and make use of a large volume of information about criminal sentences? Can the information sharing model achieve its objectives by reducing inter-judge disparity and promoting rationality?

Count me a skeptic. For structural and practical reasons, voluntary information sharing is a poor stand-alone model for promoting consistency and rationality in sentencing law. Despite considerable enthusiasm among scholars and commentators, there is little evidence that the information sharing model can serve as an effective alternative to other structured sentencing models. Nonetheless, information sharing can operate as a valuable supplement to other reforms, especially sentencing guidelines. And previous experiments with information sharing at sentencing offer important lessons about what works. Think of this Article as a skeptic’s guide to the information sharing model. It advances three related sets of claims.

The first set of claims is conceptual. Information sharing suffers from fundamental weaknesses as a mechanism for promoting inter-judge consistency and rationality. There are daunting information collection obstacles. To achieve its objectives, the information sharing model depends on case-level sentencing information that is written, comprehensive with respect to relevant facts, and representative of outcomes in similar cases. But because of the complexity of sentencing decisions, there is reason to doubt that sentencing courts can routinely generate that kind of information. There are also formidable information sharing obstacles. Sentencing judges rely on highly sensitive personal information about offenders, raising privacy concerns about any program of data dissemination. Also, as a practical matter, it is difficult to make the large volume of relevant information available to judges in a useful format. The voluntariness of the information sharing model is also an important drawback. The model assumes that judges will respond to information about earlier sentences by dutifully aligning

---


7 Carol Coulter, Website on Court Sentencing Launched, IRISH TIMES (Dublin), Aug. 3, 2010, at 4.

their decisions with others. That is unrealistic. Judges can just as easily disregard the information, ignore it, or even move in the opposite direction from their colleagues’ reasoning, at the expense of inter-judge consistency and rationality.

The second set of claims is empirical. To illustrate challenges with information collection, this Article contains an original empirical study of data reporting practices from a federal district court. The study analyzes more than four hundred “Statement of Reasons” documents, which federal judges must complete in connection with every criminal sentence. Because these documents are ordinarily nonpublic, the study is the first of its kind in the United States. The results reveal that, despite mandatory reporting requirements, judges rarely provide the kind of written opinions necessary to support an effective information sharing model. In 48.6% of cases in which a written description was required, the sentencing judge did not provide one. Just 6.0% of cases prompted a written explanation of approximately one page of discussion or more. Further, the class of cases in which the judge provided a lengthy explanation differed in important ways from the population of criminal cases as a whole. The study thus confirms some of the challenges that courts face in collecting written, comprehensive, and representative sentencing information.

To illustrate challenges with information sharing and voluntariness, this Article discusses two previous experiments with the information sharing model. Several jurisdictions in the United States and in foreign countries have developed sentencing information systems (SISs), which are interactive computer systems designed to provide judges with statistics and other information about previous sentences in similar cases. Yet no research has shown that SISs contribute to inter-judge consistency and rationality, and most systems have atrophied due to judicial neglect. Similarly, a handful of federal district courts in the 1970s experimented with sentencing councils, which are voluntary meetings of sentencing judges to discuss upcoming cases. Research revealed, however, that the councils failed to reduce inter-judge disparity because the sentencing judge retained discretion to disregard the council’s advice. Today, they are all but abandoned.

The third set of claims is prescriptive. If the Article’s conceptual and empirical claims are sound, then two features of a sentencing system might make information sharing more effective in promoting inter-judge consistency and rationality. First, information sharing is more likely to succeed as a supplement to a system of sentencing guidelines, rather than a stand-alone mechanism for structuring sentencing discretion. That is because guidelines provide a shared vocabulary about sentencing, operationalize complex sentencing concepts, and channel the attention of sentencing courts. Second, information sharing would have greater impact if defense counsel and prosecutors enjoy open access to the store of sentencing information. Although open access would accentuate privacy concerns, harnessing the adversary process would greatly improve the visibility of sentencing information and help guard against errors.

Thus, although styled as a skeptic’s guide, this Article expresses cautious optimism about the information sharing model. A carefully designed information sharing system can improve sentencing outcomes. Nonetheless, it is doubtful that
voluntary information sharing, standing alone, can meaningfully reduce inter-judge disparity or promote rationality in sentencing law. Information sharing therefore should be considered a supplement, not an alternative, to other structured sentencing reforms.

The Article proceeds in four parts. Part I describes the information sharing model and the broad support it has attracted among scholars and judges as a mechanism for reducing inter-judge disparity and improving rationality in sentencing outcomes. It also distinguishes the information sharing model from two popular alternatives: the common law model and the sentencing guidelines model.

Part II develops the Article’s conceptual claims, describing the formidable obstacles an information sharing model will face. Information collection will be difficult because the model depends on case-level information that is written, comprehensive, and representative. Information sharing will raise legal concerns about offender privacy and practical concerns about the usefulness of statistics and case information. Voluntariness also can undermine the information sharing model by leaving judges free to ignore or repudiate the reasoning of their colleagues.

Part III develops the Article’s empirical claims. It first reports the results of the empirical study, illustrating the challenges in collecting information using unique data from sentencing documents in a federal district court. It then discusses two analogous reform efforts—sentencing information systems and sentencing councils—and the mostly discouraging research concerning their effectiveness.

Part IV develops the Article’s prescriptive claims. It contends that the information sharing model would be more effective as a supplement, not an alternative, to other reform efforts. In particular, it contends a system of sentencing guidelines and open access to sentencing information would improve the chances of success.

I. THE INFORMATION SHARING MODEL

Among scholars and judges, one frequently discussed method of structuring sentencing decisions is the robust exchange of information among sentencing judges. That approach—call it the information sharing model—differs from alternative structured sentencing reforms, such as sentencing guidelines or a judge-made common law of sentencing. Yet proponents believe that it can accomplish many of the same goals, reducing inter-judge sentencing disparity and promoting rationality in sentencing law.

A. Two Key Objectives of Sentencing Reform

In indeterminate sentencing systems, which prevailed in almost all U.S. jurisdictions until the 1970s, judges enjoyed essentially unfettered discretion in choosing the type and severity of sentences.9 Grounded in the once-dominant theory that rehabilitation was the principal goal of criminal punishment,
indeterminate sentencing sought to maximize judges’ ability to individualize sentences and thereby help offenders to become productive members of society.\(^\text{10}\) Legislatures, in defining criminal offenses, often authorized a wide range of punishments. For a single violation, for example, the court might have the option of imposing a fine, a period of probation, or a term of imprisonment ranging anywhere from a few days to many years.\(^\text{11}\) No U.S. jurisdictions provided meaningful rules or guidance about how to select an appropriate punishment.\(^\text{12}\) The decision of the sentencing court was essentially unchallengeable, with no right to appeal.\(^\text{13}\) In fact, judges had no obligation even to give reasons for the sentence selected.\(^\text{14}\) By design, this “black box” gave judges enormous discretion to tailor sentences to the needs of criminals.

In the 1970s and 1980s, however, indeterminate sentencing came under sustained criticism by scholars and policymakers, and in the last thirty years criminal sentencing has undergone radical transformation. Calls for structured sentencing addressed a wide range of concerns, including dissatisfaction with the rehabilitative ideal and a desire for “truth in sentencing” undermined by parole.\(^\text{15}\) But two central claims of sentencing reformers are particularly relevant here.

First, indeterminate sentencing produced unacceptable levels of inter-judge sentencing disparity.\(^\text{16}\) Vested with enormous discretion and subject to little oversight, judges were largely left to their own intuitions in selecting an appropriate sentence. As a result, the preferences, philosophy, and biases of the

\(^{10}\) Douglas A. Berman, Conceptualizing Booker, 38 ARIZ. ST. L.J. 387, 389 (2006); see United States v. Mueffelman, 327 F. Supp. 2d 79, 83 (D. Mass. 2004) (Gertner, J.) (explaining that judges were expected to choose sentences “almost like a doctor or social worker exercising clinical judgment”).


\(^{12}\) MICHAEL TONRY, SENTENCING MATTERS 6 (1996).

\(^{13}\) STITH & CABRANES, supra note 3, at 9, 197 n.3.


\(^{16}\) The term “sentencing disparity” requires clarification because it is essentially meaningless standing alone. Many disparities, or differences, between sentences are entirely justified based on legitimate differences between offenses and offenders. Whether particular differences between sentences are justified is contestable and depends on some underlying theory of punishment. See Kevin Cole, The Empty Idea of Sentencing Disparity, 91 NW. U. L. REV. 1336, 1336–37 (1997). This Article uses the term inter-judge disparity to describe differences in sentences driven not by differences in offense or offender characteristics, but by the preferences, personality, and biases of the sentencing judge. See Ryan W. Scott, Inter-Judge Sentencing Disparity After Booker: A First Look, 63 STAN. L. REV. 1, 6 n.23 (2010).
judge played an important role in determining the sentence. Similarly situated offenders, convicted of similar crimes, could receive starkly different sentences depending on which judge was assigned to the case. Reformers argued that inter-judge disparity offends fundamental rule-of-law values such as equality, objectivity, and consistency. For the same reasons, inter-judge disparity also harms the reputation of the courts. In addition, inter-judge disparity potentially undermines the deterrent effect of criminal law by making punishment less certain and predictable. Hoping to achieve greater consistency between judges, Congress identified the reduction of inter-judge disparity as a primary goal in the Sentencing Reform Act of 1984. Many state legislatures have followed suit.

Second, indeterminate sentencing resulted in irrational sentencing law. With sentences overwhelmingly unexplained and unreviewable, neither courts nor legislatures had developed well-reasoned, intelligible, and principled sentencing law. Reformers hoped to develop a more rational system in which sentencing courts would thoughtfully develop a body of coherent sentencing rules and principles. Congress cited improved “rationality” in sentencing decisions as an

---

17 TONRY, supra note 12, at 7 (summarizing research that demonstrated “[u]nwarrented disparities, explicable more in terms of the judge’s personality, beliefs, and background than the offender’s crime or criminal history”).


22 See, e.g., S. REP. NO. 98–225, at 45 (1983) (“Sentencing disparities that are not justified by differences among offenses or offenders are unfair both to offenders and to the public.”); U.S. SENTENCING COMM’N, SENTENCING GUIDELINES AND POLICY STATEMENTS 1.2 (1987); Breyer, supra note 15, at 4.

23 See, e.g., KAN. STAT. ANN. § 74-9101 (Supp. 2012); MINN. STAT. ANN. § 244 app. § I (West Supp. 2010).

24 E.g., Edward M. Kennedy, Foreword to Federal Sentencing Guidelines Symposium, 29 AM. CRIM. L. REV. ix, ix (1992) (describing the Sentencing Reform Act as “a long overdue framework for a coherent federal sentencing policy” designed to “rationalize this critical stage of the federal criminal justice system”).
anticipated and desired benefit of the Sentencing Reform Act of 1984. 25 Many states, too, have identified rationality as a key goal of sentencing reform. 26

To be sure, there is lively debate among scholars about whether inter-judge consistency and rationality in sentencing ought to be high-priority goals. 27 Excessive concern about sentencing disparity, for example, may distract attention from other important goals in designing a just sentencing system. 28 Rather than attempt to resolve that debate, however, this Article accepts for the sake of argument that inter-judge consistency and rationality are desirable. Not only is that premise enshrined in law in many jurisdictions, 29 but structured sentencing reforms are also often advertised as a way of achieving those objectives. 30 It is fair to ask whether they can deliver on their promises.

B. Structured Sentencing Models: Guidelines, Common Law, and Information Sharing

To reduce inter-judge disparity and promote rationality, reformers have proposed various methods of structuring sentencing decisions. Broadly, three general approaches or models for structured sentencing have emerged: sentencing guidelines, a common law of sentencing, and the information sharing model.

The first model, and today the most prominent in the United States, is a system of sentencing guidelines. As of 2008, twenty states and the federal government had adopted some form of sentencing guidelines. 31 Although there is considerable variety in guidelines systems, sentencing guidelines generally consist of detailed rules promulgated by an independent sentencing commission. 32 At sentencing, the judge is required to make factual findings about the offense conduct, the effect on victims, and the offender’s criminal history and personal characteristics. Based on those factors, the guidelines specify a sentence or

25 S. REP. NO. 98-225, at 150 (stating that sentencing reforms are intended to provide “enough guidance and control of the exercise of [sentencing] discretion to promote fairness and rationality . . . in sentencing”).
26 E.g., KAN. STAT. ANN. § 74-9101(b)(1) (directing the sentencing commission to “establish rational and consistent sentencing standards”); 42 PA. CONS. STAT. ANN. § 2153(a)(12) (West 2013) (describing similar standards).
27 Professor Stith and Judge Cabranes, in particular, disavow those objectives as driving purposes of their proposed reforms. STITH & CABRANES, supra note 3, at 172–73.
28 See, e.g., id. at 106, 121–24 (calling the reduction of inter-judge disparity “a worthwhile goal for sentencing reform,” but also a “complex goal” that should not be the “myopic focus”); Albert Alschuler, The Failure of the Sentencing Guidelines: A Plea for Less Aggregation, 58 U. CHI. L. REV. 901, 902 (1991); Cole, supra note 16, at 1337.
29 See supra notes 22–26 and accompanying text.
30 See infra notes 68–74 and accompanying text.
32 Id. at 4 (describing a “continuum” of approaches adopted by different states in implementing sentencing guidelines); Richard F. Sparks, Sentencing: Guidelines, in 4 ENCYCLOPEDIA OF CRIME AND JUSTICE, supra note 19, at 1458.
sentencing range (such as 63–72 months of imprisonment). In some systems the range is mandatory or presumptive, binding judges to impose a sentence within the guideline range except in unusual circumstances.33 In others, including the federal system, the guideline range is advisory and thus allows a greater degree of flexibility and discretion.34 But the judge nonetheless must make the required findings, accurately calculate the guideline range, and give due consideration to that range when selecting a sentence.35

A second model with broad support among scholars is a common law of sentencing. Most strongly associated with Great Britain, a common-law approach has been adopted by a few U.S. states.36 In the common-law model, sentencing decisions are subject to review by appellate courts and may be reversed or altered. Although there is considerable variety in how the common law may operate—the appellate courts’ power to vacate or revise, the standard of review, the prevalence of “guideline judgments”—the essential feature of a common law of sentencing is the regulation of sentencing judges by other courts in the judicial hierarchy.37 As described by Professor Kevin Reitz, the powers of appellate courts in the common-law model include “determinations of the eligible goals of punishment decisions, and the creation of legal doctrine that translates those objectives into rules of decision.”38 Like other bodies of common law, a common law of sentencing evolves incrementally as appellate courts announce rules, carve out exceptions, draw distinctions, and occasionally overrule their prior decisions. But it culminates in a body of binding precedent, and sentencing judges must impose a sentence in conformity with that case law or risk reversal on appeal.

The third model for structuring sentencing discretion, and the focus of this Article, is the information sharing model. Under this approach, judges imposing sentences are not subject to rules promulgated by a sentencing commission or announced by appellate courts. Instead, sentencing judges have access to a robust store of information—statistics, written opinions, and other case information—about previous outcomes in similar cases.39

---

33 Sentencing guidelines in Minnesota, Washington, and Kansas fit that description. See, e.g., KAN. STAT. ANN. § 74-9101 (Supp. 2012); KAUNDER & OSTROM, supra note 31, at 11 (explaining that in Kansas, “[t]he sentencing judge must impose the presumptive sentence stated in the guideline, unless there are substantial and compelling reasons for departure”).


37 Id. at 672–73.


39 Eric Luna, Gridland: An Allegorical Critique of Federal Sentencing, 96 J. CRIM. L. & CRIMINOLOGY 25, 102 (2005) (describing a model in which “today’s courts draw[] upon the analysis and conclusions of prior judicial opinions”). Professor Luna describes this
A host of prominent judges and scholars have endorsed some form of information sharing as a way to structure judicial discretion in criminal sentencing.40 Professor Marc Miller has long urged more thorough judicial opinion writing and data dissemination to facilitate the development of coherent and principled sentencing law.41 He has also written extensively on “sentencing information systems”—searchable databases of sentencing data—as a possible reform for broken structured sentencing regimes.42 Justice Michael Wolff has extolled the advantages of an “information-based discretionary sentencing system” whose centerpiece is voluntary information sharing to assist sentencing judges.43 Professor Kate Stith and Judge José Cabranes, in their influential work criticizing the U.S. Sentencing Guidelines, have argued that the basic model for regulating sentencing discretion should be guidance from other judges through written opinions and data about previous cases.44 Judge Nancy Gertner has proposed extensive information sharing among sentencing judges at the federal level, urging federal judges to consult sentencing statistics and written opinions as a way to “make better sentencing decisions in each individual case.”45 Similarly, Judge Robert Sweet has proposed that judges should have access to a store of written opinions and statistics—a combination of “common law principles and modern technology”—that would improve sentencing decisions.46 Professor J.C. Oleson, among others, has endorsed sentencing information systems as a method of facilitating evidence-based sentencing.47

Recently, the information sharing model has attracted high-profile attention among policymakers. In February 2012, testifying before the U.S. Sentencing Commission about the future of the federal sentencing guidelines, representatives of the Judicial Conference of the United States stressed the crucial role of information sharing for sentencing judges.48 As Judge Paul Barbadoro explained,
judges “appreciate knowing whether their sentences are in step with other sentences by other judges for similar cases,” and they need a sentencing system that provides that kind of information. \(^49\) In September 2010, the State of Missouri, which has explicitly embraced the information sharing model, made national headlines for its efforts to shape sentencing outcomes by giving judges access to more information about punishment costs. \(^50\) As of January 2013, similar legislation is under consideration in Vermont. \(^51\) Overseas, in August 2010, the criminal courts of Ireland launched a website designed to share information about sentencing outcomes, the product of an ambitious four-year effort to coordinate the actions of sentencing judges. \(^52\)

As a real-world example of the information sharing model in action, consider how sentencing decisions are structured in the State of Missouri. In some ways, Missouri law preserves a traditional indeterminate system: within broad statutory ranges specified by the legislature, judges are free to impose any sentence. \(^53\) No appellate review of sentences is available, except to the extent that a sentence falls outside the statutory range. \(^54\) Neither the legislature nor a sentencing commission provides guidance to judges about eligible purposes of punishment, or about factors to consider at sentencing. \(^55\) And nothing in state law requires that judges issue written opinions—or even statements in open court—giving reasons for the

\(^49\) Id. at 9–10; Judge Barbadoro quotes Judge Richard Arcara, who stated in regional hearings that it was “crucial” for judges to have “information about how the sentence that we are considering compares overall with sentences recommended for this type of conduct.” Id. at 9 n.27. Along with Judge Arcara, Judge Jon McCalla testified that “historical data” on sentencing is “greatly valued” and necessary to allow judges “to make the difficult decisions required in sentencing on a consistent basis.” Id. at 10 n.27.


\(^52\) Coulter, supra note 7, at 4.

\(^53\) Scott, supra note 6, at 175.

\(^54\) See State v. Cook, 440 S.W.2d 461, 463 (Mo. 1969).

\(^55\) See Wolff, supra note 2, at 97 (explaining that because “[t]here is no overt guidance in Missouri law as to the purposes for punishment,” judges must “approach sentencing pragmatically and, to a degree, subjectively”).
sentence imposed.\footnote{56} By design, “[j]udicial discretion is the cornerstone of sentencing in Missouri courts.”\footnote{57}

Yet Missouri breaks from a fully indeterminate system by adopting the information sharing model. As Justice Wolff of the Missouri Supreme Court has explained, Missouri strives to lend structure to sentencing decisions by insisting upon “fully informed discretion.”\footnote{58} The Missouri Sentencing Advisory Commission (MSAC), charged by the legislature to make recommendations about sentencing,\footnote{59} has created an interactive website to share sentencing information.\footnote{60} Using the website, judges and lawyers can fill out a form that captures key offense and offender characteristics.\footnote{61} Based on that information, the system reports recommended, aggravated, and mitigated sentencing options.\footnote{62} The recommendations do not reflect the Commission’s own judgments, but “reflect sentencing practices of Missouri’s judges” based on years of historical data.\footnote{63} The idea is that judges should have access to accurate and up-to-date information about what their colleagues have done in cases that share those characteristics. Importantly, the information is provided on a “purely voluntary” basis.\footnote{64} Judges have no obligation to take the information into account, or even look it up. But the MSAC believes that, by providing useful information, it can win the “hearts and minds” of judges and persuade them to follow its recommendations.\footnote{65} It is the perfect strategy for the “show me” state: show judges the information, then leave them alone.

As the Missouri example makes clear, the information sharing model differs sharply from sentencing guidelines or a common law of sentencing. It is intended to assist sentencing judges, rather than constrain them.\footnote{66} The judiciary generates information for its own benefit, not at the direction of an external regulatory body like a sentencing commission. The information sharing model does not depend on legal rules that block judges from selecting sentences outside a specified range (as in a mandatory guidelines systems) or in conflict with binding precedent (as in a common-law system). To the contrary, judges retain essentially the same wide

\footnote{56}{See Mo. Ann. Stat. § 558.019 (West Supp. 2013); Mo. R. Crim. P. 29.07(b) (requiring only that the court “render the proper judgment and pronounce sentence”).}
\footnote{57}{Mo. Sentencing Advisory Comm’n, Recommended Sentencing: Report and Implementation Update 11–13 (2005).}
\footnote{58}{Wolff, supra note 2, at 97; see also Mo. Sentencing Advisory Comm’n, supra note 57, at 11–13.}
\footnote{59}{Mo. Ann. Stat. § 558.019.6(3).}
\footnote{60}{Michael Wolff, Missouri Provides Cost of Sentences and Recidivism Data: What Does Cost Have to Do with Justice?, 24 Fed. Sent’g Rep. 161, 162 (2012).}
\footnote{61}{Id.}
\footnote{62}{Id.}
\footnote{63}{Wolff, supra note 2, at 98.}
\footnote{64}{Id. at 97.}
\footnote{65}{Id. at 97–98, 100–101.}
\footnote{66}{Id. at 98 (discussing how the Missouri system is designed to help actors in the system to “focus on shared information” about offenses and offenders).}
discretion present in an indeterminate sentencing regime. Nor does the information sharing model compel judges to make any specific determinations or to focus on particular factors (as in an advisory guidelines system). Instead, the information sharing model insists, in Judge Gertner’s words, only that “judges need to see what other judges are doing.” The information sharing model arms judges with better information, but leaves them free to consult and act upon that information strictly on a voluntary basis.

Proponents of the information sharing model contend that it can accomplish the same goals as other structured sentencing reforms. Information sharing will reduce inter-judge disparity, the argument goes, because disparity between judges principally results from a lack of information about what other judges have done in similar cases. Closing the information gap will thus close the disparity gap. Judge Robert Sweet has predicted that an information sharing model based on written opinions and sentencing statistics “would provide, almost automatically, a firmament of reference points and a body of reasoning developed by the courts,” thereby “alleviating unwarranted disparity” between judges. Similarly, Judge Gertner argues that making sentencing opinions available to other judges is “critical” to “cabin[ing] discretion” because “[i]n order to avoid inter-judge disparities, judges must be able to see the decisions made in the courtroom next door.” Noting that judges frequently agree about the ordinal ranking of offense severity, even when they disagree about the cardinal severity of sentences, Professor Stith and Judge Cabranes predict that simply requiring written sentencing opinions and disseminating sentencing data—to serve as “quantitative guideposts to judges”—will reduce inter-judge disparity. In recommending Missouri’s approach, Justice Wolff remains hopeful that voluntary information sharing will “eliminate some of these overall disparities and gross differences.”

---

67 Gertner, supra note 4, at 278–79.
68 Cf. Zenoff, supra note 19, at 1451 (“[Some] observers believe that sentencing disparity would virtually disappear if judges had access to data about their colleagues’ decisions.”).
69 Sweet et al., supra note 5, at 943, 946.
70 Gertner, supra note 4, at 279.
72 Wolff, supra note 2, at 118.
In addition, proponents argue the information sharing model will produce more rational sentencing law. For one thing, the process of exchanging information will encourage more thoughtful decisions. Robust exchange of information, Judge Gertner explains, would encourage judges “to give coherent explanations, to articulate rules of general application” as part of a continuing dialogue among sentencing courts.73 Prompting judges to think through every sentence and to explain their reasons for the benefit of their colleagues would improve the quality of the resulting decisions.74 At the same time, information sharing at sentencing would encourage more principled decisions. For Judge Sweet, an important advantage of a robust store of information about previous cases is that it will produce a “coherent and ever-adapting body of law” that helps to intelligently translate “broad sentencing policies to individual cases.”75

Importantly, information sharing is not mutually exclusive with other reform efforts like sentencing guidelines or a common-law model.76 Indeed, many scholars endorse some combination of approaches. Professor Stith and Judge Cabranes, Judge Sweet, and others endorse information sharing as a supplement to a judge-made “common law of sentencing.”77 Judge Gertner and the Judicial Conference of the United States seek to promote information sharing in the federal system to support a system of advisory sentencing guidelines.78 They reason that, even in jurisdictions with more formal or intrusive structured sentencing programs, information sharing can serve a valuable function.79 Sentencing law should assist judges, even as it constrains them.

Nonetheless, for several reasons, it is useful to disentangle the information sharing model from other approaches. First, some jurisdictions (including Missouri and Ireland) have adopted what might be called a pure information sharing model, in which the sole mechanism for structuring sentencing decisions is a program of formal information sharing.80 An assessment of information sharing at sentencing

73 Judge Nancy Gertner, Thoughts on Reasonableness, 19 FED. SENT’G REP. 165, 167 (2007).
75 Sweet et al., supra note 5, at 945.
76 In New South Wales, Australia, for example, a searchable database of sentencing information has coexisted with other structured sentencing reforms for several decades. See discussion infra Part III.B.
77 STITH & CABRANES, supra note 3, at 170–72, 176; Reitz, supra note 38, at 1450–51, 1489, Sweet et al., supra note 5, at 928, 937.
78 See Hearing, supra note 48, at 9–10 & n.27 (statement of Judge Paul J. Barbadoro); Gertner, supra note 4, at 279–80.
79 See Gertner, supra note 4, at 279–80; Sweet et al., supra note 5, at 928.
80 See Coulter, supra note 7 (describing Ireland’s new information sharing initiative); supra notes 53–65 and accompanying text (describing Missouri’s information sharing model).
is of crucial relevance to legislators, judges, and other stakeholders in those systems.

Second, proponents of the information sharing model frequently recommend it as an alternative to more intrusive reforms, especially sentencing guidelines. In Missouri, the MSAC describes its information-based discretionary system as the product of a conscious choice between two models: information sharing, designed to assist judges and to win their approval, and a “regulatory,” “rule-driven system” of sentencing guidelines that can hope to command merely “obedience.”\(^{81}\) In Ireland, the information sharing model was advertised and embraced as a way “to avoid the proliferation of mandatory sentences with all their flaws.”\(^{82}\) The same scenario has played out in other jurisdictions, with the information sharing model positioned as a direct competitor to alternative reforms like sentencing guidelines.\(^{83}\)

Third, the information sharing model deserves separate attention because it fundamentally differs from command-and-control reforms like sentencing guidelines and judge-made common law. It rests on different assumptions about how legislators and sentencing commissions can influence judges, and it faces different challenges in shaping sentencing outcomes. Whether as a stand-alone system or as a supplement to more elaborate regulations, the information sharing model is designed to perform a distinct function, worthy of separate consideration.

Thus, the information sharing model is a distinct approach to structured sentencing, embraced by many scholars, judges, and policymakers. Its proponents argue that information sharing can reduce inter-judge disparity and promote rationality in sentencing. The remainder of this Article evaluates those claims, first exposing the information sharing model’s weaknesses and later exploring what constructive role it can play in pursuing its objectives.

II. FUNDAMENTAL WEAKNESSES OF THE INFORMATION SHARING MODEL

There is reason for skepticism that information sharing can meaningfully reduce inter-judge disparity and promote rationality in sentencing law. Conceptually, the information sharing model suffers from three fundamental weaknesses. First, there are challenges in collecting sentencing information. To succeed, the information sharing model depends on written, comprehensive, and representative information about sentencing practices. But as a practical matter that information is difficult to assemble. Second, there are challenges in disseminating sentencing information. Concerns about the privacy of offenders and victims, coupled with challenges in making the information accessible and useful for

---

\(^{81}\) See Wolff, supra note 2, at 100–01.

\(^{82}\) Tom O’Malley, *Creativity and Principled Discretion over Sentencing a Necessity*, IRISH TIMES (Dublin), Dec. 19, 2011, at 20.

judges, arise even if extensive information is available. Third, *voluntariness* presents serious challenges. By ignoring or disregarding information about how their colleagues have handled similar cases, judges can exacerbate inter-judge disparity and undermine rationality.

Before reviewing the empirical evidence and possible strategies for overcoming them, a description of each set of challenges is in order.

A. Information Collection Challenges

The most daunting challenge to the information sharing model is collecting sufficient information about the reasoning and results of past sentencing decisions. To succeed, the information sharing model requires a store of information about previous cases—for example, written opinions, offense and offender data, or aggregate statistics—made available to judges at sentencing. But two characteristics of sentencing decisions make information sharing particularly difficult in this context.

One characteristic is volume. Courts in the United States impose a staggering number of criminal sentences each year. Because plea bargaining has become the dominant method of adjudicating guilt or innocence, only a tiny fraction of cases end in trial. As of 2004, however, roughly two-thirds of criminal cases ended in a guilty plea, conviction, and sentence. In state systems, each year more than 1.07 million adults receive a sentence for a felony conviction. In the federal system, more than eighty-six thousand criminal sentences are imposed annually. Those overall figures also mask considerable variability in volume between jurisdictions and courtrooms. In some state courts, judges may impose fifty sentences per week or more than two thousand per year.

The other characteristic is complexity. Sentencing decisions typically require that judges consider a startling number of factors. Under a typical sentencing

---

84 See Luna, *supra* note 39, at 102.
85 EXAMINING THE WORK OF THE STATE COURTS, 2002: A NATIONAL PERSPECTIVE FROM THE COURT STATISTICS PROJECT 61 (Brian J. Ostrom, Neal B. Kauder & Robert C. LaFountain eds., 2003) (noting that approximately 3% of criminal cases in state courts nationwide were resolved at trial in 2001).
86 *Id.* at 61, 89 (explaining that nationally 65% of cases result in a guilty plea, and most trials result in a conviction).
statute, the judge must take into account the totality of the circumstances surrounding both the offense and the personal characteristics of the offender. The offense may be limited to a single incident, or it may involve a sprawling series of actions over many years. The judge must consider the offender’s acts, omissions, and state of mind, as well as the harm caused by the crime. Considering the offender’s personal characteristics requires a review of the offender’s entire life, before and after the offense. The number of moving parts is staggering: criminal history, assistance to the government, educational background, employment history, mental health, good deeds, public service, drug and alcohol addiction, childhood opportunities, family life, prospects for treatment, actions in pretrial detention, and on and on. In the words of one federal probation officer, at sentencing a judge must consider “a narrative of the individual from the day of his birth to the moment of his conviction.” And the inquiry does not end with the defendant since the court also may consider how the sentence would affect the offender’s family, the victims, and the victims’ families.

Compounding the complexity, sentencing judges must consider those facts in light of a wide range of purposes of punishment, which may be in tension with one another. Most jurisdictions in the United States have a “laundry list” statute that directs sentencing judges to take into account retribution, general and specific deterrence, incapacitation, rehabilitation, and various and sundry other goals, without assigning priority to any of them. In many corners of legal doctrine, judges and lawyers complain about the hopeless imprecision of “multi-factor balancing tests.” Think of sentencing as the ultimate example: a test with an infinite number of factors and no instructions about how to balance them.

These features of sentencing decisions make it difficult to generate the kind of robust store of information required to support the information sharing model.

---

90 E.g., 18 U.S.C. § 3553(a)(1) (2006) (“The court, in determining the particular sentence to be imposed, shall consider . . . the nature and circumstances of the offense and the history and characteristics of the defendant . . .”).


92 See, e.g., FED. R. CRIM. P. 32(d)(2)(B) (requiring that presentence reports include information about the financial, social, psychological, and medical impact of the offense on victims); see also Crime Victims’ Rights Act, 18 U.S.C. § 3771(a)(4) (granting crime victims the right “to be reasonably heard” at sentencing).

93 See e.g., 18 U.S.C. § 3553(a); MODEL PENAL CODE § 1.02(2) (1962); KEVIN REITZ, MODEL PENAL CODE: SENTENCING 71 (2003) (describing the “multiple choice” approach adopted by statute in most states).

Specifically, to reduce inter-judge disparity and promote rationality, the information made available to judges must be (1) written information, (2) comprehensive with respect to material facts, and (3) representative of other sentences in similar cases. On each score there is reason for skepticism.

1. *Written Information About Sentences*

Initially, the information sharing model depends on sentencing information that is *written*. It is not enough that judges identify relevant facts, formulate reasons, choose a sentence, and then enter judgment. Those facts and reasons must be reduced to writing and collected in some central store of information if future courts hope to rely upon them for guidance.95

At a minimum, that means a lot of data entry. For each case, someone would need to record information about the case, perhaps working from a long checklist of relevant offense and offender characteristics. Given the complexity of sentencing decisions, the amount of information collected in each case could be enormous, and the process correspondingly costly. In the federal system, for example, the U.S. Sentencing Commission employs around thirty full-time staff to review sentencing documents and perform data entry,96 and many states handle a much higher volume of criminal cases.

The deeper problem, however, is that the information sharing model requires more than raw data. A checklist of facts, standing alone, is a poor substitute for a written sentencing opinion because it does not disclose the judge’s reasoning. Although it may suggest possibilities, documenting potential aggravating or mitigating factors, it captures neither the factors the judge deems most relevant nor the judge’s process of prioritizing and balancing them. Leaving future judges to guess about the rationale for the sentence would not improve the rationality of sentencing outcomes. The information sharing model strives to foster a thoughtful and continuing dialogue in which judges discern the basis for previous sentences and then either accept that reasoning or draw principled distinctions. Such a dialogue is impossible without some written account of the judge’s reasoning. Thus, as proponents readily acknowledge, written opinions are the lifeblood of the information sharing approach.97

There is reason to doubt, however, that sentencing courts can generate enough written opinions to support effective information sharing. For years, scholars have been urging sentencing judges to issue full-fledged published sentencing decisions more frequently.98 Yet the overwhelming majority of sentencing decisions in the United States remain unpublished—indeed, never written down.99

95 Gertner, *supra* note 4, at 278.
96 Interview with Paul Hofer, former Special Projects Dir., U.S. Sentencing Comm’n, June 2009.
97 See STITH & CABRANES, *supra* note 3, at 170; see also Gertner, *supra* note 4, at 279 (“Wider use and availability of formal sentencing opinions is therefore critical . . . .”).
Typically, rather than prepare a written explanation, the judge announces the reasons for the sentence in open court. Although many sentencing hearings are audio recorded, most are not transcribed because they are never needed. And even if a written transcript is prepared, only a fraction of those transcripts ever become public or otherwise available to other judges in future cases.

Hearing transcripts, moreover, do a poor job of capturing the judge’s reasoning. Statements in open court are primarily directed at the offender, lawyers, witnesses, and observers in the courtroom—not future judges. Often cluttered with irrelevant material, jarred by interruptions, and disorganized, a sentencing transcript is a poor substitute for a written opinion explaining the reasons for a sentence.

The vanishingly small rate of written sentencing opinions is not the product of laziness or obstinacy. District court judges operate under acute time constraints. As Professor Frank Bowman has observed in another context, “The coin of the realm, the scarcest resource, in federal district court is time,” and many judges understandably “begrudge the time it takes to deal with sentencing issues.” The sheer volume of sentencing decisions forces judges to forego a written opinion in most cases.

2. **Comprehensive Information About Sentences**

Second, to accomplish its goals, the information sharing model depends on sentencing information that is comprehensive. A written opinion or other record of a case must capture the full range of potentially relevant facts and factors—
including, crucially, those that the judge does not find especially salient. Otherwise future judges cannot reliably compare new cases to previously decided cases.

To illustrate the need for comprehensive information, suppose that two judges impose sentences in identical burglary cases. Both offenders violently broke into a private residence at night, armed with a loaded handgun, and stole personal property worth $5,000 before being confronted by the terrified homeowner. Both offenders also pleaded guilty, expressed remorse, have identical criminal histories consisting of a single petty juvenile conviction five years ago, and enjoy strong support from loving families.

The first judge imposes a sentence of two years of probation. In a written opinion, the judge explains that there is no need for a prison sentence in light of the offender’s spotless criminal record, prospects for rehabilitation, and low risk of recidivism. But the opinion is not comprehensive. It notes the dollar amount of the theft, but does not specifically mention the handgun or the fact that the homeowner confronted the burglar.

The second judge reads the written opinion and mistakenly believes that the cases are very different. The second judge imposes a sentence of three years of imprisonment. Unlike the first case, the judge reasons, this case involved an in-person confrontation with a homeowner startled by a burglar in the middle of the night. That encounter, along with the loaded handgun, made this offense much more dangerous and the offender more culpable. Those factors, in the second judge’s view, outweigh the others.

The result is stark inter-judge disparity, although the judges did not realize it. Even though the offenses and offenders were in fact identical, the second judge received incomplete—and therefore misleading—information about the earlier case. Without comprehensive written opinions, capturing even factors the judge deems relatively unimportant, the information sharing model can malfunction.105

Unfortunately, there is reason to doubt that comprehensive sentencing information can be routinely captured. The basic problem is limited time, which prevents even the most committed judges from routinely writing detailed opinions that predict, disclose, and discuss all factors that future judges might find relevant. In addition, however, two distinctive characteristics of sentencing opinions make it difficult to generate comprehensive sentencing information.

First, written sentencing opinions are designed not merely to announce the sentence, but also to persuade the reader that the sentence is reasonable. The central challenge of the sentencing judge is to weigh a complex mixture of aggravating and mitigating factors and to strike an appropriate balance.106 Written opinions, therefore, do not catalogue in exhaustive detail all potentially relevant

---

105 Austin Lovegrove has made a similar observation in evaluating the Scottish Sentencing Information System. See Lovegrove, supra note 105, at 39–40.
106 Schild, supra note 99, at 125 (“Sentencing consists in trying to reconcile a number of totally irreconcilable facts.”).
factors. Judges naturally focus on the facts and factors they find most persuasive while downplaying others.107

Second, the review structure of sentencing decisions also makes it tempting to write an abbreviated opinion. In any given case, there is little risk that the judgment will be questioned because no other district court judge reviews the case or writes a dissent.108 Although appellate review of sentences is available in most jurisdictions, criminal defendants frequently waive the right to appeal,109 and in any event the chances of a successful appeal are very low.110

Moreover, when judges anticipate that the parties may appeal from a sentence, they have strategic incentives to provide fewer details. Judges do not like to be reversed on appeal,111 and detailed sentencing opinions sometimes increase the risk of reversal. An appellate court might be forced to vacate and remand based on a stray reference to a prohibited factor, a misstatement of a relatively minor fact, or an artless phrase that sounds too much like a legal error. Outside of especially complex or controversial cases, writing a long opinion is asking for trouble. A simple announcement from the bench or a terse written order gives the parties little to go on, and therefore little to attack.

In theory, exhaustive data entry could compensate for gaps in the written explanation. A member of the court staff, for example, could complete a checklist in every burglary case that indicates the presence of a weapon or an encounter with a victim. That way, future judges could see a complete picture of the case even if the written opinion contains omissions. As a practical matter, however, such comprehensive data entry is prohibitively costly. Every sentencing decision involves a theoretically infinite number of facts and factors, especially when including those not relevant in the particular case. No court system can realistically code and transmit that kind of hyperdetailed information for every sentence.112 Accordingly, the need for comprehensive information forces a choice between

107 Cf. ANTONIN SCALIA & BRYAN A. GARNER, MAKING YOUR CASE: THE ART OF PERSUADING JUDGES 94 (2008) (advising lawyers, in drafting a brief, to persuade the reader by “putting some facts in high relief and some in low relief—and . . . omitting others altogether”).

108 Cf. John O. McGinnis & Charles W. Mulaney, Judging Facts Like Law, 25 CONST. COMMENT. 69, 104–05 (2008) (arguing that judges are less susceptible to confirmation bias than legislators because “[i]f a judge ignores facts in the majority opinion, he will suffer the embarrassment of a strong dissenter (a factual ‘whistleblower’) who points out an opinion’s factual flaws”).


110 In the federal system, for example, 87.4% of sentencing appeals are affirmed or dismissed. U.S. SENTENCING COMM’N, supra note 88, tbl.56.


112 See Schild, supra note 99, at 133–34 (explaining that use of parameter values in defining criminal records would discourage judges because of the sheer number of values, making it difficult to define criminal records).
keeping data-collection costs manageable and preventing errors that undermine inter-judge consistency and rationality.

3. Representative Information About Sentences

Third, to accomplish its goals, the information sharing model depends on sentencing information that is representative. According to proponents, information sharing reduces inter-judge disparity and promotes rationality by giving judges a context or “picture” of how each new case compares to previous cases.¹¹³ Judges can then align their sentences to fit that picture. The trouble is that, in several ways, the picture can be skewed.

One possibility is that the body of sentencing information may disproportionately reflect some kinds of cases. Judges, after all, do not choose at random whether to write a detailed sentencing opinion. Sometimes a judge chooses to issue a published opinion because the sentence presents a novel legal issue, and the bulk of the opinion is dedicated to that issue.¹¹⁴ Other times a judge chooses to issue a published opinion because the case involves extreme or unusual facts or the sentence imposed may generate controversy.¹¹⁵ Time constraints not only make it impossible to prepare a written explanation in every case, but also force judges to be selective about which sentences warrant extended discussion. Given the choice, judges often focus on groundbreaking, extreme, or otherwise special cases.

That kind of imbalance, although understandable, presents serious problems for the information sharing model. A store of sentencing information that consists disproportionately of extreme or unusual cases is incomplete, and therefore potentially deceptive, as a guide for judges in ordinary cases. Indeed, it can cause the information sharing model to backfire. Relying on a skewed body of information can undermine rationality by creating inconsistency with “invisible” sentences that did not warrant a published opinion.¹¹⁶

Another risk is that the body of sentencing information may disproportionately reflect the work of especially prolific judges. It is no insult to the judiciary to recognize that different judges have different levels of enthusiasm for criminal sentencing. Some judges excel at writing sentencing opinions and

¹¹³ Gertner, supra note 4, at 279; Gertner, supra note 73, at 166–67; Sweet et al., supra note 5, at 943, 946.
¹¹⁴ E.g., United States v. Pimental, 367 F. Supp. 2d 143, 149–54 (D. Mass. 2005) (Gertner, J.) (holding that the Sixth Amendment forbids the consideration of acquitted conduct at sentencing, or alternatively that acquitted conduct may be considered only if it is found beyond a reasonable doubt).
¹¹⁶ See Austin Lovegrove, Judicial Decision Making, Sentencing Policy, and Numerical Guidance 42 (1989) (“[I]t is acknowledged that there is disparity in sentencing, and it is important that a few disparate cases should not appear to be the norm.”).
would eagerly participate in an information-driven inter-judge dialogue, perhaps earning induction into Professor Doug Berman’s “Sentencing Judges Hall of Fame.” But many others have little patience for sentencing and little inclination to invest more time in preparing written opinions. Given the option, some judges may contribute more and better information than their colleagues.

That poses a problem for the information sharing model because a nonrepresentative store of sentencing information, dominated by some voices while others remain silent, can exacerbate inter-judge disparity. Future judges, guided by a skewed sense of previous sentencing patterns, might inadvertently misalign their decisions with those of less prolific judges. The risk is particularly acute if outlier judges contribute more opinions and information than their colleagues— a plausible scenario, since judges have a special incentive to write a detailed opinion when they suspect that others may disagree with the outcome.

Initially, then, information collection poses formidable challenges for the information sharing model. To succeed, the model depends on written, comprehensive, and representative sentencing information, and generating that kind of information is time-consuming and costly. As explained below, however, even if such information were collected, there would be practical difficulties in making it available to judges.

B. Information Dissemination Challenges

Another set of challenges for the information sharing model relates to the dissemination of sentencing information. Once collected, information about previous cases must somehow be made accessible to sentencing judges. Yet, because sentencing often turns on highly sensitive personal information about offenders and victims, information sharing raises privacy concerns. In addition, there are practical hurdles in making relevant sentencing opinions and statistics accessible to judges.

1. Privacy


Bowman, supra note 104, at 356.

consider the offender’s medical history, mental health, and physical condition, which may serve as mitigating factors at sentencing.\textsuperscript{120} Work history and opportunities for employment may factor into the offender’s prospects for rehabilitation.\textsuperscript{121} The court may also discuss the offender’s home and family life, including the offender’s performance as a parent, obligation to care for young children, or support from relatives.\textsuperscript{122} In some cases, mitigating facts at sentencing include profoundly personal information about the offender, such as a history of sexual or physical abuse.\textsuperscript{123}

Victims and other third parties also have privacy interests at stake. Sentencing courts often emphasize the harm caused to victims of crime, which may require a discussion of physical, psychological, or economic injuries.\textsuperscript{124} Indeed, in some jurisdictions the victims of crime have a legal right to be heard at sentencing.\textsuperscript{125} Witnesses who testify at sentencing, on behalf of the offender or the government, sometimes provide personal information and seek to keep their testimony confidential.\textsuperscript{126} For a host of reasons, evidence at sentencing may be submitted under seal or considered \textit{in camera}.\textsuperscript{127} Yet a written opinion or case record designed to offer future judges a comprehensive picture of the case must, of necessity, disclose that information.

Of particular concern is information about cooperation with the government. Offenders frequently receive a lower sentence because of their assistance to police or prosecutors, and a written sentencing opinion or case record may indicate the nature and extent of that cooperation. Sentencing courts may also rely on the statements of cooperating witnesses who appear at sentencing, at trial, or before a grand jury. That information is potentially explosive because it may expose the offender or family members to violence and retaliation. The risk is chillingly real

\begin{itemize}
\item[\textsuperscript{120}] \textit{Id.} at III-25 to -26.
\item[\textsuperscript{121}] \textit{See id.} at III-27 to -29.
\item[\textsuperscript{122}] \textit{See id.} at II-3 to -4, III-25 to -26.
\item[\textsuperscript{123}] \textit{See id.} at II-4 (instructing probation officers that “[a] very sensitive area that may need to be addressed is whether the defendant has a history of being physically, sexually, or emotionally abused”); \textit{see also} United States v. Corbitt, 879 F.2d 224, 230 (7th Cir. 1989) (discussing facts disclosed in a presentence report, and concluding that “[t]he criminal defendant has a strong interest in maintaining the confidentiality of [the] presentence report”).
\item[\textsuperscript{124}] \textit{Office of Prob. & Pretrial Servs., supra} note 119, at II-23 to -25, III-11 to -12.
\item[\textsuperscript{125}] \textit{See, e.g.,} Crime Victims’ Rights Act of 2004, 18 U.S.C. § 3771(a)(4) (2006); \textit{Fed. R. Crim. P. 32(i)(4)(B)} (“Before imposing sentence, the court must address any victim of the crime who is present at sentencing and must permit the victim to be reasonably heard.”).
\item[\textsuperscript{126}] \textit{See Fed. R. Crim. P. 32(d)(3)(B)} (excluding from presentence reports certain sensitive information, including “any sources of information obtained upon a promise of confidentiality”).
\item[\textsuperscript{127}] \textit{See id. 32(i)(4)(C)} (authorizing courts to hear statements at sentencing in camera upon a party’s motion and “for good cause”).
\end{itemize}
in the Internet age, when websites like Who’s a Rat? make it easier than ever to identify and locate people who cooperate with the government.\footnote{See About Us, WHO’S A RAT, http://www.whosarat.com/aboutus.php (last visited June 9, 2013).}

Courts’ treatment of presentence reports (PSRs) underscores those privacy concerns. A PSR is a written document, usually prepared by a member of the court staff or probation officer, designed to assist the judge at sentencing. In most jurisdictions, a PSR is prepared as a matter of routine in felony cases.\footnote{See FED. R. CRIM. P. 32(c)(1)(A) (requiring a presentence investigation and report, except in a few circumstances).} Like a written opinion or case record, the PSR may contain personal information about the offender, victims, family members, or third parties.\footnote{See id. 32(d) (setting forth required contents of federal presentence reports); OFFICE OF PROB. & PRETRIAL SERVS., supra note 119, at III-1 to -44, (providing guidance to probation officers responsible for preparing presentence reports).} Recognizing those privacy interests, courts have a longstanding practice of maintaining the confidentiality of PSRs.\footnote{U.S. Dep’t of Justice v. Julian, 486 U.S. 1, 12 (1988) (“[C]ourts have been very reluctant to give third parties access to the presentence investigation report prepared for some other individual or individuals.”) (emphasis omitted).} Indeed, for decades, it was controversial to disclose the report even to the offender.\footnote{See FED. R. CRIM. P. 32(c)(2) advisory committee’s note (collecting sources and describing the “heated controversy” over “whether as a matter of policy the defendant should be accorded some opportunity to see and refute allegations made in such reports”).} Disclosure to third parties is almost always forbidden. As one federal judge put it, “I guess I feel strongly that such information should not be made accessible to anyone outside the case.”\footnote{Ian Urbina, New York’s Federal Judges Protest Sentencing Procedures, N.Y. TIMES, Dec. 8, 2003, at B1 (quoting Judge Sterling Johnson, Jr.) (discussing disclosure of sentencing documents to Congress).}

Thus, the dissemination of sentencing information to judges raises important concerns about the privacy of offenders and victims. But the disclosure of sensitive information is not the only concern. As the next section explains, making any sentencing information easily accessible for judges may be equally difficult.

2. Accessibility and Relevance

Another challenge in disseminating sentencing information is ensuring that the relevant information is easily accessible. As noted above, the information sharing model depends on a pool of information that is written, comprehensive, and representative.\footnote{See supra Part II.A.1–3.} Beyond that, however, judges also need some way to wade through the available information and zero in on what is useful. For information sharing to reduce inter-judge disparity, sentencing judges need a way to find similar cases, comparing previous sentences with a new set of facts. Only then can they impose a sentence along the same lines. Likewise, for information sharing to promote rationality, sentencing judges need to consult the reasoning of other courts.
in relevant cases. Only then can they write an opinion accepting that reasoning or distinguishing the case, for the benefit of future judges.

Accordingly, a centerpiece of the information sharing model is some system—an electronic database, a website, or a set of shared files, for example—that allows sentencing judges to search for previous decisions that are similar and relevant. But because of the complexity and volume of sentencing decisions, there are practical challenges in building such a system, and judges may be discouraged from making effective use of the available information.

First, the complexity of sentencing decisions complicates the design of any search system. With a theoretically infinite number of variables in play, no search form or user interface can capture them all. Instead, the designers of the system of necessity must select some search parameters to include, while ignoring others. That process is subjective and controversial.

For example, a system might allow judges to search previous sentences based on the offender’s criminal history. But that kind of search could take many forms. It may be implemented as a yes/no field (i.e., did the offender have any prior criminal history?). Or it might include two yes/no fields, one for juvenile criminal history and another for adult criminal history. Or it might include a single scaled search parameter that captures the total number of prior offenses, the total number of adult offenses, the total number of violent offenses, or all of the above. Or it might include search parameters based on age at the time of the offender’s first offense, at the time of the most recent offense, or a hundred other variations. As the designers of one sentencing information system lamented, “It does not take a mathematical wizard to realize that if there are even as few as three or four levels of each of these [criminal history] variables, there are over 700 combinations of aspects of this one variable—criminal record.” One researcher estimates that the total number of combinations of criminal history parameters alone “would reach into the tens of thousands.” Even the massive datafiles created by the U.S. Sentencing Commission, which are breathtaking in their complexity, could not support searches on many of the criminal-history factors described above.

---

135 As discussed below, courts in Missouri and abroad in Canada, Scotland, and Australia have experimented with providing judges searchable computer databases of sentencing information. See infra Part III.B.1.

136 See Miller, supra note 42, at 134; Schild, supra note 99, at 132.


138 Schild, supra note 99, at 133.

139 For criminal history, the datafiles include a yes/no variable, criminal history points and category scores that indirectly reflect a host of underlying criminal history facts, a count of “incidents” classified according to level of seriousness, and adjustments for committing the instant offense while under various forms of court supervision. See U.S. SENTENCING COMM’N, VARIABLE CODEBOOK FOR INDIVIDUAL OFFENDERS 7–80 (2010). The Commission’s data do not, however, capture which prior incidents were violent, the age of the offender at the time of each incident, the time elapsed since each incident, or the name or type of court that adjudicated the prior incident.
Many other factors relevant at sentencing are equally complex and difficult to operationalize. The problem is that there is no “right” design. Each variation of those search parameters is potentially relevant, and different judges may prefer different options. By picking some limited number for searching and excluding others altogether, system designers risk alienating judges. Why bother with a search system that seems focused on all the wrong issues?

Worse, the subjectivity of search parameters can threaten inter-judge consistency. Suppose, for example, a search system allows judges to search for cases in which the offender’s criminal history included any “violent” prior offenses. Although judges might broadly agree that violent criminal history is highly relevant at sentencing, they may disagree about what offenses qualify as violent or nonviolent. The hypothetical case described above, in which an armed burglar confronts a homeowner at night but never fires or even brandishes the firearm, might be a close case. If judges’ differing views about what qualifies as violent are embedded into the search parameters, the system may lock in inter-judge disparity by concealing the disagreement from future users.140

Nor would full-text searching of written opinions solve the problem. In other contexts, judges (and their law clerks) typically find relevant case law using electronic services like Westlaw and LexisNexis, which offer sophisticated search tools.141 Flexible as they are, however, those services would have significant limitations as a means of reliably identifying similar cases. The language that judges use to describe offense and offender characteristics varies enormously from opinion to opinion, making it easy for factually similar cases to escape notice.142 Innovations in electronic discovery, such as predictive coding, hold out some promise as lower-cost and higher-accuracy alternatives to traditional keyword searches.143 But for now, those techniques remain prohibitively costly and time-consuming for routine use in sentencing decisions.144

140 Uri Schild acknowledges this difficulty, but maintains that it is “not really a problem” because users are interested only in whether the judge thought the offense was violent (or serious or aberrant) and “passed sentence accordingly.” Schild, supra note 99, at 133. The nature of the offense “objectively speaking” does not matter. Id. He is mistaken. The whole point of reducing inter-judge disparity is to ensure that objectively similar offenders receive equivalent sentences.

141 Peter W. Martin, Reconfiguring Law Reports and the Concept of Precedent for a Digital Age, 53 VILL. L. REV. 1, 38–39 (2008). Notably, however, not all state trial courts can afford subscriptions to those services, leaving judges to either rely on the parties or to conduct research using more traditional and less flexible methods. Id. at 27.

142 As a thought experiment, try to devise a search of written opinions that would capture all sentences in which the offender’s criminal history included a violent crime. The terms “violent” or “violence” would not necessarily appear in the opinion, and the possible synonyms are endless: “armed,” “weapon,” “firearm,” “gun,” “knife,” “attack,” “brandish,” and dozens of others. Picking out the relevant results would be time-consuming, and could never guarantee that all relevant cases had been discovered.

143 Adam M. Acosta, Predictive Coding: The Beginning of a New E-Discovery Era, 56 RES GESTAE 8, 8 (2012). Predictive coding “uses a combination of sophisticated
Second, the volume of sentencing decisions makes the search for similar cases difficult. If the system provides too few search parameters, judges may discover that an overwhelming number of cases seem relevant.\(^{145}\) At a high level of generality, each case may be similar to hundreds or even thousands of others.\(^{146}\) Given their time constraints, sentencing judges cannot carefully review so many potentially relevant cases.\(^{147}\) Presented with a daunting volume of matches, judges may give up.

If, on the other hand, the system provides too many search parameters, then searches frequently will yield no results. In sentencing, no two cases are exactly alike—any more than two human beings are exactly alike. Indeed, a key premise of the information sharing model is that judges can develop more rational sentencing law by offering principled reasons to distinguish between dissimilar cases. As a practical matter, however, judges who search for matching cases and constantly come up empty may conclude that the exercise is a waste of time. This level-of-generality problem should not be overstated; as diligent legal researchers know, if one search returns ten thousand results and the next returns zero, continuing to refine the search parameters can make the results manageable. Nonetheless, the volume of sentences creates special challenges in making relevant results readily accessible to judges.

The volume of sentencing decisions also can threaten the accuracy and usefulness of the information. The information sharing model anticipates that sentencing judges will serve both as producers and consumers of information. In that sense, all judges are interdependent, relying on one another to make contributions to a central pool of information. For some judges, however, a crushing case load or other time constraints may result in errors or hurried and unhelpful written explanations. In turn, that inaccurate or inadequate information can frustrate others, who will find the system less useful. A vicious cycle is

\(^{144}\) Id. (noting that, as a first step in predictive coding, “only a few thousand documents need to be reviewed” to prepare a “seed set” of documents that serves as a basis for subsequent searches); \textit{id.} at 8–9 (discussing the costs, in attorney time, of document review to support predictive coding).

\(^{145}\) Schild, \textit{supra} note 99, at 134.


\(^{147}\) \textit{Cf.} Lewis A. Kornhauser, \textit{Adjudication by a Resource-Constrained Team: Hierarchy and Precedent in a Judicial System}, 68 S. CAL. L. REV. 1605, 1623 (1995) (discussing the cost advantages of limiting the number of precedents a trial judge must consult); Stith, \textit{supra} note 111, at 20 (noting that trial courts cannot engage in the “costly, complex” process of reviewing all relevant appellate precedent on a particular legal issue, and therefore may “reduce its costs . . . by considering only a subset of relevant appellate decisions”).
possible, as everyone begins to doubt that painstaking accuracy and thoughtful opinions will actually benefit their colleagues. As computer programmers say, “garbage in, garbage out.”

In summary, the complexity and volume of sentencing decisions create serious challenges in disseminating sentencing information and making it accessible to judges. It is therefore naïve to expect that making sentencing information available to other judges will, in the words of Judge Sweet, “almost automatically” provide “a firmament of reference points and a body of reasoning” that can “alleviat[e] unwarranted disparity” between judges.\textsuperscript{148} Making relevant sentencing information accessible to judges is anything but automatic.

There are further challenges, however, even assuming that the right kind of sentencing information can be collected and disseminated. As explained below, the critical final step involves how judges make use of the information.

\textbf{C. Voluntariness Challenges}

Assuming that written, comprehensive, and representative information has been collected, and the relevant information is accessible, the information sharing model predicts that judges will seek out that information and treat it as persuasive, perhaps even authoritative. By design, the model is voluntary, leaving sentencing judges free to decide whether and how to consult information about previous cases. The idea is not to constrain judges, but to assist them in the exercise of “fully informed discretion.”\textsuperscript{149} Information sharing will reduce inter-judge disparity and promote rationality, the argument goes, because judges who otherwise would have reached a contrary result will instead conform their sentences to those of their colleagues.

The unspoken assumption of the model is that the primary source of inter-judge disparity and irrationality is ignorance. According to this view, judges broadly agree about sentencing, and they reach inconsistent results only because they lack information about how other courts have handled similar cases. Therefore, “sentencing disparity would virtually disappear if judges had access to data about their colleagues’ decisions.”\textsuperscript{150}

That premise is unrealistic. Inter-judge disparity results not merely from a lack of information, but from deep disagreements about sentencing values and priorities. Surveys of judges, for example, have revealed persistent differences of opinion about important sentencing principles and policies.\textsuperscript{151} Judges are

\begin{footnotesize}
\textsuperscript{148} Sweet et al., \textit{supra} note 5, at 943, 946.
\textsuperscript{149} MO. SENTENCING ADVISORY COMM’N, \textit{supra} note 57, at 11–13; Wolff, \textit{supra} note 2, at 97.
\textsuperscript{150} Zenoff, \textit{supra} note 19, at 1451 (describing the view of some observers).
\textsuperscript{151} See, e.g., LINDA DRAZGA MAXFIELD, U.S. SENTENCING COMM’N, SURVEY OF ARTICLE III JUDGES ON THE FEDERAL SENTENCING GUIDELINES, at ES-1 to -5 (2003), \textit{available at} http://www.ussc.gov/Research/Research_Projects/Surveys/200303_Judge_Survey/jsfull.pdf (reporting disagreements about how well the U.S. Sentencing Guidelines accomplish purposes of punishment set out by Congress); U.S. SENTENCING COMM’N,
particularly divided, for example, on hot-button criminal justice topics like child pornography, drug trafficking, and white-collar fraud.\textsuperscript{152} Prior research on sentencing in the federal system has documented a sharp spike in inter-judge disparity following the shift from mandatory to advisory guidelines,\textsuperscript{153} despite the absence of any changes in judges’ access to sentencing information. Other research reveals significant differences in sentencing outcomes between Democratic and Republican appointees.\textsuperscript{154} In light of those basic disagreements, it is unrealistic to expect that sentencing judges in a voluntary system will dutifully fall into line when supplied with information about their colleagues’ decisions.

Instead, when judges disagree in good faith with the actions of their colleagues, there is every reason to believe they will disregard the information and impose a sentence they believe is just and appropriate. The result will be persistent inter-judge disparity. It is possible that, over time, voluntary information sharing could foster a dialogue between judges that resolves the disagreement, as courts settle on one view or the other. But it is equally possible that the disagreement will continue, with dueling courts committed to opposing views. A purely voluntary model is powerless to correct that problem.\textsuperscript{155}

Disregarding the information, moreover, is not the only alternative to falling in line with other judges’ decisions. Consider three other possibilities. First, judges can avoid discovering the information in the first place. When judges know or suspect that their preferred sentence is out of step with that of their colleagues, nothing in a voluntary system prevents them from simply ignoring the available information. Such a “see no evil, hear no evil” impulse would thwart the information sharing model in precisely those cases where it could be most useful.

Second, judges can keep looking. Upon discovering information about past sentences that run contrary to their own preferences, they can continue to mine the available opinions in search of more favorable results. Behavioral literature on confirmation bias has documented similar cognitive errors in other contexts.\textsuperscript{156} Confirmation bias describes the tendency to unwittingly select and interpret evidence in a manner that confirms a previously held belief or hypothesis, while

\textsuperscript{152} U.S. Sentencing Comm’n, supra note 151, tbl.8.
\textsuperscript{153} Scott, supra note 16, at 30–41.
\textsuperscript{155} Cf. Kevin R. Reitz, The New Sentencing Conundrum: Policy and Constitutional Law at Cross-Purposes, 105 Colum. L. Rev. 1082, 1114 (2005) (arguing that advisory sentencing guidelines, as a form of voluntary advice to sentencing judges, are “by definition unenforceable” and provide “no systemic remedy for outlier sentences”).
\textsuperscript{156} See Raymond S. Nickerson, Confirmation Bias: A Ubiquitous Phenomenon in Many Guises, 2 Rev. Gen. Psychol. 175, 189–97 (1998) (discussing confirmation bias in number mysticism, witch hunts, policy making, medicine, science, and judicial reasoning).
minimizing or failing to recognize contrary evidence. Previous research has documented confirmation bias among prosecutors, police, and jurors, and the same tendency undoubtedly exists among judges.

Although concerns about cognitive errors like confirmation bias should not be overstated, the complexity of sentencing decisions creates conditions that may facilitate confirmation bias. At a high level of generality, many offenses and offenders share common characteristics, providing judges with a wealth of “confirming” data points that may reinforce their intuitions. Yet in the details, sentences are as infinitely variable as human beings, allowing courts to draw infinite distinctions that minimize the importance of “disconfirming” data points.

Third, judges can become even more polarized. Judges confronted with information about previous sentencing patterns may not only reject their colleagues’ approach, but may also stake out an even more extreme position. That kind of attitude polarization effect finds some support in psychology literature as well. In a seminal study, for example, test subjects who held opposing views

157 Keith A. Findley & Michael S. Scott, The Multiple Dimensions of Tunnel Vision in Criminal Cases, 2006 WIS. L. REV. 291, 309, 313 (“[S]tudies show that, in some circumstances, people do not respond to information at variance with their beliefs by simply ignoring it, but rather by working hard to examine it critically so as to undermine it.”); Dan M. Kahan, The Cognitively Illiberal State, 60 STAN. L. REV. 115, 121 n.26 (2007) (“[C]onfirmation bias’ . . . refers to the tendency of persons to seek out and assign more weight to evidence that confirms a prior belief or hypothesis than to evidence disconfirming it.”).


159 See David E. Klein, Unspoken Questions in the Rule 32.1 Debate: Precedent and Psychology in Judging, 62 WASH. & LEE L. REV. 1709, 1717–18 (2005) (“J[udges are human, and it is hard to imagine that they escape all the cognitive pitfalls that other people stumble into.”); McGinnis & Mulaney, supra note 108, at 104–05 (comparing confirmation bias between the judicial and legislative branches).

160 Judges usually explain their sentencing decisions in public, and in most jurisdictions the parties can challenge the sentence on appeal. That kind of accountability is often effective in countering cognitive biases. See Klein, supra note 159, at 1718 (observing that “the typical judging experience” involves “the conditions under which accountability has the best chance of reducing cognitive errors”); Jennifer S. Lerner & Philip E. Tetlock, Accounting for the Effects of Accountability, 125 PSYCHOL. BULL. 255, 259, 263 (1999).

161 See supra notes 90–93 and accompanying text.

162 See Charles G. Lord et al., Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence, 37 J. PERSONALITY & SOC. PSYCHOL. 2098 (1979) (concluding that polarization may result when individuals who “hold strong opinion[s] on complex social issues,” such as the death penalty, are exposed to empirical evidence).
about capital punishment grew more polarized after reading identical information about conflicting research on the deterrent effect of the death penalty. As Professor Cass Sunstein explains, at least when people begin with strongly held views, a “balanced presentation[]” in which “competing arguments or positions are laid out side by side” is likely to increase rather than reduce attitude polarization. Information sharing, in other words, does not inevitably result in agreement and uniformity. To the contrary, in some circumstances it can harden individuals’ resolve and make disagreements more pronounced.

Conceptually, then, the information sharing model suffers from several fundamental weaknesses in promoting inter-judge consistency and uniformity at sentencing. The collection of sentencing information is challenging because of the need for written, comprehensive, and representative opinions. The dissemination of that information is challenging because of the complexity and volume of sentencing decisions. And the voluntariness of the model leaves judges free to ignore or disregard the information.

III. EMPIRICAL RESEARCH ON SENTENCING INFORMATION SHARING

These challenges are not merely speculative. New and existing empirical research tends to confirm the fundamental weaknesses of the information sharing model. This Part first reports the results of an original study of federal sentencing data collection. It then examines the history of two reform efforts grounded in the information sharing model, sentencing information systems and sentencing councils. Collectively, the research reinforces the serious challenges related to information collection, information dissemination, and voluntariness.

A. An Empirical Study of the Collection of Sentencing Information

Despite strong support for the information sharing model, to my knowledge no previous empirical research has examined the collection of sentencing information. As explained above, the information sharing model depends on a store of sentencing information that is written, comprehensive, and representative. Promoting rationality in sentencing decisions requires thoughtful and principled explanations and a continuing dialogue between judges, which is impossible without written sentencing opinions. Likewise, comprehensive and representative information is crucial, since a skewed body of previous decisions

163 Id. at 2100–05. But see Arthur G. Miller et al., The Attitude Polarization Phenomenon: Role of Response Measure, Attitude Extremity, and Behavioral Consequences of Reported Attitude Change, 64 J. PERSONALITY & SOC. PSYCHOL. 561, 561–69 (1993) (replicating the Lord study and finding no attitude polarization when subjects reported their views immediately before reviewing the dueling studies).


165 See supra Part II.A.1.
can actually deepen inter-judge disparity. No study has tested the feasibility of that kind of information collection.

To fill that gap, this Article reports the results of an original study of the collection of sentencing information. Drawing on a unique dataset of sentencing documents from a federal district court, the study evaluates the quantity and characteristics of written sentence explanations submitted pursuant to mandatory reporting requirements. Because the documents that form the basis for the analysis are generally nonpublic, the analysis is the first of its kind.

1. Data and Coding

The study is based on a key federal sentencing document called the Statement of Reasons. In federal court, sentencing judges must complete a Statement of Reasons in connection with every criminal sentence. Upon completion, the document is transmitted to the U.S. Sentencing Commission, which extracts and records data about the sentence. The Commission then generates massive datafiles, based on the contents of tens of thousands of documents each year, to serve as the basis for statistical reports about nationwide sentencing practices. To ensure that complete and comprehensive data are available, the chief judge of every federal district court is required by statute to “ensure” that in every case a “written statement of reasons for the sentence imposed” is submitted in a format specified by the Commission.

To understand how the Statement of Reasons works, a brief summary of federal sentencing practice may be helpful. The U.S. Sentencing Commission has promulgated detailed sentencing guidelines that translate offense and offender characteristics into a sentencing range, expressed as a narrow range of months of imprisonment (such as 52–63 months). The guideline range is “advisory,” meaning that judges must consider the advisory guideline range in every case, but they are free to impose any reasonable sentence consistent with a laundry-list statute that sets out broad purposes of punishment. Within-range sentences reflect the sentencing court’s judgment that the guideline range is proper for the “mine-run” of similar cases, and that the particular offense and offender are “not different enough to warrant a different sentence.”

The Statement of Reasons form is adapted to the guidelines regime. It is four pages long and requires that judges provide basic information about the sentence

---

166 See supra Part II.A.2–3.
168 Id.
171 Id.
imposed. Judges must check a box indicating whether the sentence falls within, above, or below the advisory guideline range. They also have the option of checking boxes that correspond to reasons for an out-of-guidelines sentence. The checkboxes express the reasons for the sentence in very general terms, such as “aggravating or mitigating circumstances” or “the nature and circumstances of the offense.”

Most relevant for present purposes, the form provides several spaces in which judges may provide a narrative description of the reasons for the sentence. The form states that a written narrative description is required in two categories of cases: (1) all sentences outside the guideline range, and (2) all sentences within the guideline range that carry a term of imprisonment of “greater than 24 months.” Judges sometimes use the back of the form or attach additional pages or documents, such as a written sentencing opinion or a transcript of the sentencing hearing.

The Statement of Reasons form is usually secret and confidential, pursuant to a policy statement issued by the Judicial Conference of the United States. But one federal district court, the District of Massachusetts, has voted to make Statements of Reasons available to the public on the PACER system (Public Access to Court Electronic Records). That decision affords a rare opportunity to study sentencing practices that remain hidden in every other federal court. The open-access policy reflects the court’s admirable commitment to greater transparency in criminal sentencing.

---


174 STATEMENT OF REASONS FORM, supra note 173, at 2–3.

175 Id. The checkboxes correspond to the titles of guideline departures, see U.S. Sentencing Guidelines §§ 5K2.0, 5K2.13 (2011), and broad purposes of punishment, see 18 U.S.C. § 3553(a)(1) (2006). The checkboxes themselves provide no information about the particular case. Judges can check a box marked “Age,” for example, but the checkbox does not indicate whether the offender was young or old.

176 STATEMENT OF REASONS FORM, supra note 173, at 2 (“Explain the facts justifying the departure.”); id. at 3 (“Explain the facts justifying a sentence outside the advisory guideline system.”); id. at 4 (“Additional facts justifying the sentence in this case . . . .”).

177 See id. at 2, 4.


To assess the kind of information collected using the Statement of Reasons, the study examines a full year (fiscal year 2006\textsuperscript{181}) of cases from the District of Massachusetts. A total of 411 statements were reviewed and coded, representing approximately 80\% of the Statements of Reasons submitted to the Commission that year.\textsuperscript{182} Details concerning document selection are set forth in the Appendix.

Of particular interest is whether the Statements of Reasons supply the kind of written, comprehensive, and representative explanations necessary to support an effective information sharing model.\textsuperscript{183} Accordingly, the study examines the length of any written statement explaining the judge’s reasons, measured by the number of sentences of text contained in the narrative description. Such a sentence count admittedly captures only the length, and not the quality, of the written explanation. But the length of the narrative description is a fairly reliable proxy for the level of factual detail and the thoroughness of the judge’s reasoning. The study classifies a narrative description of one to three sentences as a short explanation. It classifies a narrative description of four to nine sentences, at least a paragraph but less than a page, as a medium explanation. It classifies a narrative description of ten or more sentences of text, roughly one page, as a long explanation.

2. Results

The results are discouraging. As shown in Figure 1, even in cases where a written explanation of the reasons for the sentence is mandatory,\textsuperscript{184} most Statements of Reasons contain no written explanation at all. Only a small fraction of the documents contain a comprehensive explanation.

\begin{figure}
\caption{Figure 1}
\end{figure}

\textsuperscript{181} The analysis includes Statements of Reasons filed in the Commission’s 2006 fiscal year, which runs from October 1, 2005, to September 30, 2006.

\textsuperscript{182} See infra Appendix Part A.

\textsuperscript{183} See supra notes 95–118 and accompanying text (explaining why the information sharing model depends on written sentencing opinions, comprehensive with respect to potentially relevant facts and representative of similar cases in the jurisdiction).

\textsuperscript{184} See supra note 177 and accompanying text.
For cases in which a written narrative description is required, the Statement of Reasons contains no indication whatsoever of the judge’s reasons in 48.6% of cases. In another 9.1% of cases, the only indication of the judge’s reasons is a checkmark in a box corresponding to a broad reason for the decision (such as “aggravating or mitigating factors” or “the circumstances of the offense”). Together, that means that in 57.7% of cases—more than half—the document contains no written narrative description, despite the Commission’s reporting requirements. Such a low rate of written explanations is a major obstacle for the information sharing model, given the importance of written opinions in fostering a dialogue between sentencing courts, and thereby promoting rationality in sentencing law.

185 As discussed infra, the percentage of sentences with no written explanation is much lower for nonguideline sentences (28.6%) than for within-range sentences of more than twenty-four months (78.7%). But because a written explanation is mandatory in both circumstances, Figure 1 reports the level of explanation for both categories combined.

186 See supra Part II.A.1. Checkmarks indicating only a general topic area are insufficient to promote inter-judge consistency and rationality. See United States v. Blackie, 548 F.3d 395, 401 (6th Cir. 2008) (vacating sentence for lack of an adequate...
Nor are the available written explanations comprehensive. In 27.1% of cases, the document provides only a short narrative explanation of up to three sentences of text, and in another 9.1% of cases it provides a medium-length explanation of four to nine sentences. That kind of quick summary represents an improvement over a checklist, but it would be of little use to future courts drawing upon that information to develop a rational body of sentencing law. The explanation can highlight a few factors that the judge deemed most important, but of necessity many relevant facts must be omitted, and there can be room for only a bare-bones explanation of how the judge weighed the competing facts and considerations. Short-shrift explanations pose a problem for the information sharing model because incomplete opinions may fail to capture facts that future judges deem important, masking important similarities or differences between cases and thereby generating inter-judge disparity and undermining rationality.187

There is some good news. In a small number of cases (6.0%), the judge provided a long written explanation consisting of at least ten sentences of text. Indeed, in a tiny subset of cases (1.1% of the total), the judge wrote a formal sentencing opinion or attached a hearing transcript that includes more than fifty sentences of explanation. Those opinions are outstanding, offering future judges not only a full sense of the relevant facts, but also valuable insights into the sentencing judge’s reasons and approach.

Unfortunately, those cases tend to be unusual, resulting in a nonrepresentative body of long written opinions. As shown in Figure 2, sentences outside the guideline range are more likely to produce a long explanation than sentences within the guideline range.

---

187 See supra Part II.A.2.
Figure 2 illustrates the percentage of cases in which the Statement of Reasons contains a long explanation, consisting of ten sentences of text or more, across three categories: sentences within the guideline range, sentences above the guideline range, and sentences below the guideline range. For sentences within the advisory guideline range, judges provided a long explanation in just 2.5% of cases. But for sentences outside that range—those different from the “mine-run” of similar cases—188—the likelihood of a long explanation is roughly four times greater. The documents contain a long explanation for 9.9% of below-range sentences and for 8.3% of above-range sentences. The result is a skewed collection of long written descriptions.189 A store of sentencing information that consists disproportionately of unusual or extreme cases poses a real threat to the

188 In theory, a judge is free to impose a nonguideline sentence even in an “ordinary” case based on policy-driven disagreement with the guideline range. Kimbrough v. United States, 552 U.S. 85, 108–09 (2007). That is exceedingly rare, however; sentencing courts overwhelmingly justify their nonguideline sentences based on the special facts of the case.

189 By way of comparison, in the full set of 411 sentences, 67% are within-range sentences. In the subset of 20 sentences with a long explanation, however, just 35% are within-range sentences.
information sharing model, since a skewed backdrop may mislead judges about sentencing patterns in cases that are ordinary, but effectively invisible.\footnote{\textit{See supra} note 116 and accompanying text.}

The pool of long written explanations is nonrepresentative in another way. Some judges were far less likely to provide full explanations than others. Figure 3 shows, for each judge, the percentage of that judge’s cases in which the Statement of Reasons contains a long written explanation.\footnote{Figure 3 reflects data only for judges with a criminal caseload of at least fifteen sentences during the year of the study. For further discussion of the minimum-caseload requirement, see \textit{infra} Appendix Part B.}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Percentage of Cases with Long Explanation, By Judge, Minimum Caseload Required (13 judges total), District of Massachusetts, FY 2006}
\end{figure}

Not all judges wrote long explanations of their sentences at equal rates.\footnote{In this table and throughout, this Article uses numbers rather than names to identify individual judges. \textit{See Scott, supra} note 6, at 173.} Six of the thirteen judges did not submit any long explanations during the year of the study. By contrast, one judge provided a long explanation for 21.1\% of sentences, and another provided a long explanation for 33.3\% of sentences.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Percentage of Cases with Long Explanation, By Judge, Minimum Caseload Required (13 judges total), District of Massachusetts, FY 2006}
\end{figure}
As a result, the pool of sentences with long written explanations is imbalanced, with some judges far better represented than their colleagues. Judge 13 singlehandedly accounts for 45% of long opinions submitted the year of the study. That is more than Judges 1 through 10 combined. Together the court’s three most prolific authors of long opinions account for 80% of the court’s output of long written explanations. A body of sentencing information in which some judges have a dominant voice, while others remain silent, can exacerbate inter-judge disparity and undermine rationality by creating a distorted impression of actual sentencing patterns.193

Those differences are not the product of chance. As set forth in the Appendix, logistic regression models confirm that the pool of written explanations is significantly imbalanced.194 First, some judges and some kinds of cases are more likely to produce a written explanation than others. Nonguideline sentences are significantly and strongly correlated with written explanations ($b = 3.320$, $p < .001$). Likewise, longer terms of imprisonment are significantly, although more weakly, correlated with written explanations ($b = 0.007$, $p = .001$). Categorical variables capturing the identity of the sentencing judge are also a significant predictor of whether a written explanation is provided ($p = .001$). Surprisingly, however, the fact that a written explanation is mandatory is not itself a significant predictor of a written explanation, after controlling for other factors ($b = 0.719$, $p = .214$).195

Second, some kinds of cases are more likely than others to produce a long written explanation consisting of ten or more sentences of text. Nonguideline sentences are significantly correlated with long explanations ($b = 1.086$, $p = .037$). Neither sentence length ($b = -0.003$, $p = .482$) nor the fact that a written explanation is mandatory ($b = 1.338$, $p = .234$), however, is a statistically significant predictor of a long explanation.196

3. Implications

These findings suggest that challenges in collecting written, comprehensive, and representative sentencing information are formidable. Even in a regime of detailed and mandatory reporting requirements, backed by a federal statute, the available pool of written sentencing opinions is limited.

For three reasons, however, these results should be interpreted with caution. First, Statements of Reasons are not designed as a way for judges to share information with one another. At present, their sole audience is the U.S. Sentencing Commission, which uses the documents as inputs for its statistical reports. No one else has access to the judge’s statements, no matter how insightful

---

193 See supra Part II.A.3.
194 For more information about variables used in the regression models, see infra Appendix Part B.
195 See infra Appendix Part C and Table 1.
196 See infra Appendix Part C and Table 2.
or well reasoned and judges know it. They therefore have little incentive (aside from the commands of federal law) to explain their decisions in detail. Data collection efforts surely would improve if judges believed that their written explanations would reach a wider audience.

Second, the Statement of Reasons form itself is clumsy and cumbersome. Scarcely concealing its intended use as a data-entry tool for the Commission, it consists primarily of a “parade of nearly meaningless check boxes.” That format no doubt discourages judges from providing more thorough narrative descriptions of the reasons for a sentence. With a more open-ended form that encouraged judges to set out their reasons, the contents of the statements might be richer and more useful.

Third, this study examines documents from a single federal district court. Data collection practices in one district may not be representative of practices in other federal courts nationwide or in state courts responsible for the vast majority of criminal sentences. And the District of Massachusetts, admittedly, is far from ordinary when it comes to sentencing issues. Several members of the court are well respected as sentencing experts. Judges Nancy Gertner and William Young have written scholarly articles on sentencing issues, and Judge Patti Saris now serves as Chair of the U.S. Sentencing Commission. Moreover, the same commitment to transparency that led the District of Massachusetts to approve its unique disclosure policy might make its data reporting practices materially different from those of other courts.

On the other hand, because of the distinctive features of the District of Massachusetts, this study likely underestimates the challenges in collecting high-quality sentencing information. Unlike other courts, the District of Massachusetts has consciously chosen to make its Statements of Reasons widely available. The judges know that their reasons will not simply be filed away in a drawer, but ordinarily will be accessible to the public. If anything, the court’s special interest in sentencing likely translates into more frequent and more thorough written opinions, not less. In addition, federal judges enjoy a larger support staff, more law clerks, and a slower criminal docket than their counterparts in state courts. If the

197 Bruce Green, Thinking About White-Collar Crime and Punishment, CRIM. JUST. Fall 2010, at 1, 58.
199 Chanenson, supra note 98, at 147.
200 Id.
203 Although some federal courts are deluged with criminal immigration cases, prompting the development of “fast track” disposition programs, the District of
collection of sentencing information is inadequate in the District of Massachusetts—widely recognized as one of the best sentencing courts in the country—then it is probably even worse elsewhere.

The study thus reinforces the idea that information collection challenges are substantial. Despite explicit reporting requirements, sentencing courts face time pressures and other constraints that prevent them from routinely providing written, comprehensive, and representative information.

B. Research on Previous Reform Efforts

In addition to this original research, insights about the information sharing model can be derived from the experiences of previous sentencing reform efforts grounded in information sharing. Two efforts are particularly salient: (1) sentencing information systems (searchable computer databases of sentencing information constructed in a handful of jurisdictions around the world), and (2) sentencing councils (periodic roundtable discussions of sentencing among judges on the same court). The struggles of those reform efforts, and their occasional successes, underscore serious challenges with information dissemination and voluntariness.

1. Sentencing Information Systems

In the last twenty-five years, a handful of jurisdictions worldwide have experimented with sentencing information systems (SISs), which are searchable electronic databases of sentencing information. Courts in Canada, Scotland, Ireland, and Australia have launched large-scale experiments with SISs. 204 In the United States, Missouri has constructed a web-based system that could be described as an SIS. 205 Although their designs vary, SISs are designed to enable judges to look up statistics, case summaries, written opinions, or other information about previous sentences in similar cases. 206 Using interactive search forms, a judge preparing to impose sentence enters some key characteristics of the case. The SIS responds by reporting information about matching cases. Some systems provide only aggregated information about whole categories of cases, such as the distribution of nationwide sentences for aggravated theft. Others provide specific information about individual cases, such as the case summary or full sentencing opinion in a relevant case. 207

Massachusetts had no such program at the time of the study. See U.S. SENTENCING COMM’N, SOURCEBOOK OF FEDERAL SENTENCING STATISTICS 211 (11th ed. 2006) (reporting no departures imposed pursuant to an “early disposition program”).

204 Miller, supra note 42, at 129; Coulter, supra note 7, at 4.
205 Wolff, supra note 2, at 101–03.
206 Miller, supra note 42, at 129.
The primary goals of an SIS, in the jurisdictions that have adopted them, are to reduce inter-judge sentencing disparity and to promote rationality.\textsuperscript{208} When a judge sentences an offender, an SIS can provide a statistical picture of how similar offenders have been sentenced by other judges in the same court, region, or nation.\textsuperscript{209} Making better information available to judges, the argument goes, makes it possible for judges to align their decisions with those of their colleagues, and thus produces more consistent and better-reasoned sentences.\textsuperscript{210}

Surprisingly, to date no empirical research has examined whether any SIS has succeeded in reducing inter-judge sentencing disparity.\textsuperscript{211} The indirect evidence, however, is disheartening. As explained below, many of the most prominent SISs in foreign courts have been abandoned. Further, in jurisdictions where SISs have survived, they have taken on more of a support role following the implementation of other structured sentencing reforms. Consider the experiences of the three most prominent SISs, launched in Canada, Scotland, and New South Wales, Australia.

In Canada, four provinces experimented with SISs in the late 1980s.\textsuperscript{212} A national sentencing commission had concluded that “detailed information on current practices” would prove valuable to sentencing judges,\textsuperscript{213} and surveys of judges had revealed a widespread appetite for better information about past sentencing outcomes.\textsuperscript{214} Although judges were not required to search the Canadian SIS when imposing sentence, the designers worked closely with judges to design useful and relevant search parameters.\textsuperscript{215}

\begin{footnotesize}
\begin{enumerate}
\item[208] Austin Lovegrove, \textit{Statistical Information Systems as a Means to Consistency and Rationality in Sentencing}, 7 INT’L J. L. & INFO. TECH. 31, 32 (1999) (“The primary aim of these systems is to promote consistency and rationality—the idea is that cases described by similar relevant offence characteristics and offender circumstances should receive similar sentences.”).
\item[209] Miller, \textit{supra} note 42, at 129.
\item[210] \textit{Id.} at 135; \textit{Sentencing Comm’n for Scot., The Scope to Improve Consistency in Sentencing} para. 8.22, at 36 (2006) (“[SISs] can be used to inform sentencing decision making and increase consistency within and between sentencers.”).
\item[211] See Miller, \textit{supra} note 42, at 133–34 (recognizing that current systems have not been subjected to rigorous analysis and evaluation); Schild, \textit{supra} note 99, at 127. The courts that have experimented with SISs apparently do not capture and publicize the kind of data necessary to support such research. See Arie Freiberg, \textit{Australia: Exercising Discretion in Sentencing Policy and Practice}, 22 FED. SENT’G REP. 204, 206 (2010). As a result there is no direct evidence that the availability of an SIS improves inter-judge consistency.
\item[212] See generally Doob & Park, \textit{supra} note 137, at 54 (analyzing the development of SISs for judges in Canada).
\item[214] See Doob & Park, \textit{supra} note 137, at 55 (noting that 79\% of the 414 responding judges believed that better information about past sentences would be helpful).
\item[215] See Miller, \textit{supra} note 42, at 130.
\end{enumerate}
\end{footnotesize}
It did not work. Within a few years, the Canadian SIS experiment was abandoned because judges declined to use it.\textsuperscript{216} The system’s architect attributed its demise to the fact that the reform was strictly voluntary. He was surprised to discover that “[j]udges do not, as a rule, care to know what sentences other judges are handing down in comparable cases.”\textsuperscript{217} Judges were especially reluctant to use the system, he reported, “if they knew (or thought) that these other judges [had] different approaches” than their own.\textsuperscript{218}

Similarly, in Scotland, an SIS was developed for judges of the High Court of Justiciary, which handles sentencing for some of the nation’s most serious criminal offenses. Following a trial period in 1997, the SIS became available to all of the court’s judges in 2002. Judges themselves proposed developing the system, principally as a defensive measure to head off more intrusive reforms (such as presumptive sentencing guidelines) that would have sharply limited their discretion.\textsuperscript{219} As in Canada, judges worked closely with designers of the SIS to define the available search parameters.\textsuperscript{220} The result, according to its creators, was a highly flexible and valuable resource.\textsuperscript{221}

Yet the Scottish SIS collapsed as well. In 2006, just a few years after the system became widely available, the Sentencing Commission of Scotland reported that the system “was not widely used” and had “largely fallen into abeyance.”\textsuperscript{222} Judges rarely used the system to gather information, and rarely took the time to enter narrative information concerning their own sentencing decisions.\textsuperscript{223} The data were also incomplete and at times inaccurate.\textsuperscript{224} The Commission found that “currently the SIS does not have anything other than the most marginal of impacts on the imposition of sentences.”\textsuperscript{225} In 2010, one of the designers of the system proclaimed that the SIS is essentially nonoperational, having been “allowed to atrophy” following years of judicial neglect.\textsuperscript{226}

The Australian state of New South Wales boasts the world’s most successful SIS, at least measured by longevity. The Judicial Commission of New South

\textsuperscript{216} Id.
\textsuperscript{217} Id. (quoting Anthony Doob, Sentencing Aids: Final Report 5 (1989) (unpublished manuscript) (on file with author)) (internal quotation marks omitted).
\textsuperscript{218} Id. at 130–31 (quoting Anthony Doob, Sentencing Aids: Final Report 10 (1989) (unpublished manuscript) (on file with author)) (internal quotation marks omitted).
\textsuperscript{219} See Hutton & Tata, supra note 83, at 275.
\textsuperscript{220} SENTENCING COMM’N FOR SCOT., supra note 210, para. 8.23, at 36.
\textsuperscript{222} SENTENCING COMM’N FOR SCOT., supra note 210, para. 4.17, at 16.
\textsuperscript{223} Id. para. 8.26, at 37.
\textsuperscript{224} Id.
\textsuperscript{225} Id.
\textsuperscript{226} Hutton & Tata, supra note 219, at 275.
Wales has maintained a searchable database of sentencing statistics since 1988.\textsuperscript{227} Strangely, in the twenty-four years of its operation, no effort has been undertaken to evaluate its effectiveness. There is evidence that some judges actually use the system by performing searches,\textsuperscript{228} but no research has been conducted to assess its effects on sentencing outcomes.

Over time, however, the New South Wales system has been relegated to a support role. Beginning in 1998, the Supreme Court of New South Wales began to announce guideline judgments, which specified an advisory sentencing range for offenses.\textsuperscript{229} In 2003, the New South Wales General Assembly enacted standard non-parole periods—essentially a statutory determinate sentencing scheme—that set presumptive sentences for many serious offenses. Judges are now permitted to depart from the legislatively prescribed sentence, but only upon finding facts justifying a departure.\textsuperscript{232} The General Assembly concluded that, despite the availability of the SIS, inter-judge sentencing disparity had reached unacceptable levels. The primary reason for the new regime, according to its supporters, was to reduce that form of disparity.\textsuperscript{233}

Canada, Scotland, and New South Wales are not the only jurisdictions that have experimented with SISs. Promising new systems are underway around the world, and they may provide new insights on the information sharing model. Ireland’s criminal courts launched a new, publicly accessible web-based SIS in 2010.\textsuperscript{234} Australia recently launched a nationwide SIS, modeled on the New South

\textsuperscript{227} Miller, supra note 42, at 129, 133. For a detailed description of the system’s origins and functionality, see Potas et al., supra note 207, at 104–14.

\textsuperscript{228} JUDICIAL COMM’N OF NEW S. WALES, ANNUAL REPORT 2008–09, at 24, fig.7 (2009) (showing some 930,000 pages accessed by judicial officers from 2008 to 2009).

\textsuperscript{229} See, e.g., R v Jurisic, (1998) 45 NSWLR 209, 229–30. Such guideline judgments were “intended to be indicative only” and to preserve “flexibility” for sentencing judges, but reflected the Court’s concerns about “[i]nconsistency in sentencing,” which “offends the principle of equality before the law.” Id. at 216, 220–21.

\textsuperscript{230} Freiberg, supra note 211, at 207.

\textsuperscript{231} For examples of statutory determinate sentencing, see TONRY, supra note 12, at 28.

\textsuperscript{232} See Freiberg, supra note 211, at 207.

\textsuperscript{233} See PATRIZIA POLETTI & HUGH DONNELLY, JUDICIAL COMM’N OF NEW S. WALES, THE IMPACT OF THE STANDARD NON-PAROLE PERIOD SENTENCING SCHEME ON SENTENCING PATTERNS IN NEW SOUTH WALES 3 (2010), available at http://www.judcom.nsw.gov.au/publications/research-monographs-1/research-monograph-33/monograph33.pdf. Preliminary research by the Judicial Commission of New South Wales indicates that presumptive nonparole sentences have succeeded in improving inter-judge consistency. Id. at 59–60. Some observers, however, have suggested that the real motivation was to pander to public criticism that sentences were too lenient. See Freiberg, supra note 211, at 207.

\textsuperscript{234} See Coulter, supra note 7.
Wales system, for federal offenses prosecuted throughout the country. In the United States, Missouri has rolled out a web-based system for disseminating statistics and other sentencing information. And at least preliminary work has begun on SIS-type efforts in Israel and England and in state courts in Oregon. So far, however, those systems have not been the subject of careful research, and their future is uncertain.

Sentencing information systems around the world offer valuable insights about the promise and weaknesses of the information sharing model. A second sentencing reform effort grounded in the information sharing model, this one drawn from federal sentencing practice in the United States, is equally instructive.

2. Sentencing Councils

Sentencing councils were groups of district judges, serving on the same court, who met regularly (typically once per week) to discuss upcoming sentencing decisions. Four federal district courts, in Brooklyn, Chicago, Detroit, and Oregon, experimented with sentencing councils in the 1960s and 1970s. The councils functioned as roundtable discussions of upcoming cases. In advance of the meeting, all participating judges, including the judge responsible for imposing sentence, would review the presentence report and other materials and make an initial recommendation about the appropriate sentence. As a group, judges would then talk about the evidence, share their views about the most important facts and considerations, and try to persuade one another. Nancy Gertner has compared sentencing councils to clinical rounds performed by physicians.

The principal goal of sentencing councils was to reduce inter-judge disparity. Sentencing councils provided the sentencing judge with detailed, case-specific information about how other judges would respond to each new set of facts. Participating judges saw the work of sentencing councils as an “educational process” in which judges could “pool[] [their] knowledge and experience and

---

235 See Wendy Kukulies-Smith, Address at the National Judicial College of Australia and ANU College of Law Sentencing Conference: The Quest for Sentencing Consistency in the Federal System 1–6 (Feb. 6, 2010).
236 See Wolff, supra note 60, at 161.
237 See Miller, supra note 42, at 129.
238 See Michael H. Tonry, Sentencing Councils, in 4 ENCYCLOPEDIA OF CRIME AND JUSTICE, supra note 19, at 1483, 1483–84.
242 See, e.g., Diamond & Zeisel, supra note 239, at 124; Tonry, supra note 238, at 1483.
learn[] from each other." Sharing information about how other judges would handle the case, they predicted, "will likely have the effect of ameliorating the likelihood of sentence disparity."

Consistent with the information sharing model, sentencing councils were voluntary in two ways. First, the recommendations of other judges at the sentencing council were merely advisory. The sentencing judge retained sole discretion to impose the final sentence. "The goal was to assist, not to restrain: "No attempt is made at these Councils to impose the will of one upon another." Second, in some districts, participation in the sentencing council was itself optional, and not all judges elected to attend the meetings.

Participating judges raved. They found the discussions valuable, reporting that they frequently changed their views about the appropriate sentence in response to their colleagues' advice. The discussions, by all accounts, were informal and friendly. Outside observers were uniformly impressed with the seriousness and care with which sentencing councils approached each case.

Yet two major research projects found that sentencing councils did not meaningfully reduce inter-judge sentencing disparity. Shari Diamond and Hans Zeisel studied the effects of sentencing councils in Brooklyn and Chicago by comparing sentencing judges' initial recommendations with their final sentences. They found strong evidence of inter-judge disparity in the initial recommendations, with judges' proposed sentences diverging on average by 36.7% in Chicago and 45.5% in Brooklyn. In their evaluation of cases, the data showed, "some judges are clearly more severe than others." But sentencing councils alleviated only a small fraction of that disparity. Diamond and Zeisel found that final sentences imposed, after full discussion with the sentencing council, diverged on average by

---

243 Smith, supra note 240, at 20.
245 See Comm. on the Fed. Courts, supra note 244, at 881; Diamond & Zeisel, supra note 239, at 117.
246 Smith, supra note 240, at 20.
247 See Diamond & Zeisel, supra note 239, at 139–40 & tbl.22.
249 See Smith, supra note 240, at 19–21 (expressing the view that the sentencing council was "aiding in the reduction of sentencing disparity").
250 See Hosner, supra note 248, at 19.
251 See Tonry, supra note 238, at 1484 (describing and evaluating both studies).
252 Diamond & Zeisel, supra note 239, at 123 & tbl.6.
253 Id. at 123.
35.4% in Chicago and 41.1% in Brooklyn. They estimated, therefore, that sentencing councils reduced inter-judge disparity by just 4% to 10%.255

Another study, commissioned by the Federal Judicial Center (FJC), examined the effects of sentencing councils in Detroit, Chicago, and Brooklyn.256 Focusing on five offense types, the FJC study compared levels of inter-judge disparity in the years before and after the establishment of the sentencing council.257 The results were discouraging. In every district, inter-judge disparity actually increased for some offenses, even as it decreased for others.258 Based on the mixed results, the study concluded that "councils may increase disparity as frequently as they decrease it."259

The experiment did not last. By the mid-1980s, sentencing councils in the federal courts were abandoned.

3. Lessons for the Information Sharing Model

The experiences of SISs and sentencing councils highlight some of the fundamental weaknesses of the information sharing model. Challenges in disseminating sentencing information have resulted in struggles, and even the collapse, of several SISs around the world. Both reform efforts also have suffered from voluntariness problems, with sentencing courts too often disregarding, ignoring, or rejecting the available information.

(a) Information Dissemination

Strikingly, in jurisdictions in which SISs have struggled or even collapsed, the most frequently cited reasons have been practical difficulties with disseminating the information in a useful way. No system can survive for long if its users—judges and their court staff—find it cumbersome, confusing, or unhelpful.

Scotland’s SIS offers the clearest example. Judges themselves proposed the development of the system as a tool to improve inter-judge consistency and

254 Id. at 145 & tbl. 26.
255 See id. at 147. Professor Michael Tonry has urged caution in interpreting the Diamond and Zeisel study because it measured sentences using an “arbitrary point scale” that treated longer probationary sentences as the equivalent of short prison terms, and “[a] different arbitrary scale would have yielded different results.” Tonry, supra note 238, at 1484. In a tautological sense, he is correct; of course using a different scale would produce “different results,” expressed on the new scale. But neither Tonry nor any other researcher has concluded that a different scale would have altered Diamond and Zeisel’s core finding that sentencing councils had a negligible effect on inter-judge sentencing disparity.
257 Id. at 62–63.
258 Id. at 86–95.
259 Id. at 94.
rationality.\textsuperscript{260} They understood that the success of the project depended on a collective effort, since every narrative description they wrote would be added to a searchable database available to their colleagues. And they had strong practical incentives to make it work because the SIS was launched in part to short-circuit proposals for more intrusive reforms like sentencing guidelines.\textsuperscript{261} Yet in just a few years the system fell into disuse, with judges rarely searching for relevant information about past sentencing practice.\textsuperscript{262}

Complexity was one part of the problem. Although the system’s designers worked with judges in selecting search parameters and boasted about the system’s ease of use and flexibility, judges abandoned the system after a short time in operation.\textsuperscript{263} As system designers around the world have acknowledged, an SIS cannot remain neutral with respect to the factors that are most relevant at sentencing.\textsuperscript{264} Privileging some factors while excluding others inevitably discourages some users from accessing the available information.\textsuperscript{265}

Volume was another. In rolling out the Scottish SIS, system designers had to train judges never to enter more than a few parameters at once because the system almost always returned too few cases.\textsuperscript{266} Other systems have encountered the opposite problem. Designers of a fledgling SIS in Israel declined to include “detailed descriptions” or even “summaries” of cases because, given the large number of results, “it was felt that judges would not take the time to read case descriptions.”\textsuperscript{267}

Those problems can compound one another. Following the launch of the Scottish SIS, judges who found the system cumbersome and unhelpful did not always make time to enter detailed and accurate information about their sentences.\textsuperscript{268} Such shortcuts and errors made the system less useful, reinforcing other judges’ sense that the effort of producing detailed opinions was a waste of time and energy.\textsuperscript{269} The result was a vicious cycle in which the system gradually atrophied and ultimately collapsed.

\begin{thebibliography}{99}
\bibitem{261} Hutton & Tata, supra note 83, at 275; Tata, supra note 264, at 209–10.
\bibitem{262} \textit{SENTENCING COMM’N FOR SCOT.}, supra note 210, para. 8.26, at 37.
\bibitem{263} Hutton & Tata, supra note 221, at 44–46.
\bibitem{264} Miller, \textit{supra} note 42, at 134; Schild, \textit{supra} note 99, at 132; Tata, \textit{supra} note 221, at 153 (acknowledging that “no [sentencing] data can ever hope to be value-free” and that judges will not choose to use an SIS if the information lacks “credibility”).
\bibitem{265} \textit{See} Tata, \textit{supra} note 221, at 153–54 (arguing that the “principal determinants” of whether an SIS enjoys credibility and will be used are the selection, relationship structure, and presentation of sentencing information).
\bibitem{266} Lovegrove, \textit{supra} note 105, at 37.
\bibitem{268} \textit{SENTENCING COMM’N FOR SCOT.}, \textit{supra} note 210, para 8.26, at 37; Tata, \textit{supra} note 260, at 210 (noting that sentencing data became unreliable when court clerks took over responsibility for data entry, after which the SIS “quietly withered[ed] away”).
\bibitem{269} \textit{SENTENCING COMM’N FOR SCOT.}, \textit{supra} note 210, para. 8.26, at 37.
\end{thebibliography}
Privacy concerns also have hampered the success of SISs. With few exceptions, SISs have been made available only to judges and court personnel, with no access to defense counsel, prosecutors, or the public. In part, not to the judges’ credit, that kind of secrecy was designed to shield sentencing decisions from public scrutiny and criticism. But in part it also reflected legitimate concerns about disclosing deeply personal information about the lives of offenders, their family members, and victims. In the handful of systems that are publicly accessible, privacy concerns do not arise either because the information is expressed entirely in the aggregate (as in Missouri), or because records are anonymous and little detailed case-specific information is available (as in Ireland).

(b) Voluntariness

The abandonment of sentencing councils and the struggles of SISs also undermine a key assumption of the information sharing model. For information sharing to reduce inter-judge disparity and promote rationality, judges must respond to information about previous sentences in similar cases by conforming their decisions to those of their colleagues. The history of these reform efforts makes clear that judges frequently respond differently. Of particular concern are three patterns of response: (1) judges consult the available information but then disregard it, (2) judges ignore the sentencing information entirely, and (3) judges polarize their views by shifting to an even more extreme position.

First, the experiences of sentencing councils suggest that judges who voluntarily consult sentencing information may simply disregard it. Professor Michael Tonry has concluded that sentencing councils had little effect because “[t]he council recommendations are only advisory,” leaving judges who disagree with their colleagues free to disregard the advice. The FJC study concurred, noting that it should come as no surprise that a strictly voluntary self-reform that preserved broad discretion for sentencing judges did not meaningfully affect sentencing outcomes. Surprisingly, even where judges consciously sought out the advice of their colleagues—in courts where participation in the sentencing

---

270 Tata, supra note 260, at 211 (citing public access to the Scottish SIS as “[t]he key underlying issue, which was never properly resolved” and an important factor in the system’s collapse).

271 Miller, Map and Compass, supra note 1, at 1388 (noting that Scottish judges have been “stingy” about disclosing information from the SIS); Freiberg, supra note 211, at 208 (noting that the New South Wales SIS “remains inaccessible to a broader audience,” although some data on federal sentencing in Australia recently became available because of “external prodding”).

272 Miller, Map and Compass, supra note 1, at 1388.

273 See supra notes 129–133 and accompanying text (discussing the confidentiality of presentence reports).

274 Tonry, supra note 238, at 1484.

275 PHILLIPS, supra note 256, at 100.
council was optional—the effects on inter-judge disparity were negligible.\textsuperscript{276} As long as the sentencing judge had the final word, sentencing councils had little effect.\textsuperscript{277}

Second, the experiences of SISs and sentencing councils make clear that judges in a voluntary system may ignore the information entirely, declining to look it up even when it is available. The Canadian SIS fell into disuse because, in the words of the chief designer, “[j]udges do not, as a rule, care to know what sentences other judges are handing down in comparable cases.”\textsuperscript{278} Particularly troubling is evidence that Canadian judges were especially unlikely to consult the SIS “if they knew (or thought) that these other judges have different approaches” than their own.\textsuperscript{279} That kind of reluctance prevents the information sharing model from reducing inter-judge disparity in precisely the category of cases where it holds the most promise. Similarly, participation in sentencing councils was spotty when judges’ attendance was optional. In the Northern District of Illinois, five of the court’s fourteen judges regularly participated in the sentencing council, bringing more than 60% of their cases before the sentencing council.\textsuperscript{280} But six of the court’s judges elected never to participate, and three others participated only occasionally, bringing less than 50% of their cases before the council.\textsuperscript{281} The result was that only one-third of criminal sentences benefited from the council’s advice.\textsuperscript{282} Some judges resisted the procedure on the ground that it would intrude upon their independent judgment at sentencing, and that it would be a “waste of time.”\textsuperscript{283}

Relatedly, judges may not be fully aware of the ways they ignore available sentencing information. Experience with SISs also highlights the risk of confirmation bias when consulting a large body of sentencing information. Consider the operation of the New South Wales system. Judges can enter a few search parameters and generate a histogram that shows a distribution of prior sentences.\textsuperscript{284} For example, for offenders (1) under the age of twenty-one, who (2) plead guilty, (3) to driving while intoxicated, the SIS may report that 29% of offenders received probationary sentences, 24% received sentences of imprisonment, 21% received intermediate sentences, and 15% received compound sentences.\textsuperscript{285} In theory, judges could review the written opinions in each of the

\textsuperscript{276} Diamond & Zeisel, supra note 239, at 145, 147.
\textsuperscript{277} PIERCE O’DONNELL ET AL., TOWARD A JUST AND EFFECTIVE SENTENCING SYSTEM: AGENDA FOR LEGISLATIVE REFORM 18 (1977).
\textsuperscript{278} Miller, supra note 42, at 130 (quoting Anthony Doob, Sentencing Aids: Final Report 5 (1989) (unpublished manuscript) (on file with author)).
\textsuperscript{279} Id. at 130–31.
\textsuperscript{280} See Diamond & Zeisel, supra note 239, at 139–41 & tbl.22.
\textsuperscript{281} See id.
\textsuperscript{282} Id. at 142–43.
\textsuperscript{283} Comm. on the Fed. Courts, supra note 244, at 881–82 (describing Judge Marvin Frankel’s hypothesis about the degree of resistance revealed by polls of federal judges).
\textsuperscript{284} See Potas et al., supra note 207, at 119 (illustrating possible chart).
\textsuperscript{285} Id. (showing an actual report from the New South Wales SIS with similar results).
approximately 120 cases summarized in the chart, attentive to factors that may be present in a new case that fits those criteria. Equally likely, however, given the volume of matching cases and limited time, judges could focus on cases that confirm the result they already have in mind.

Third, the experiences of sentencing councils suggest judges may respond to information about other judges’ sentences by polarizing their views and adopting an even more extreme position. The FJC study of sentencing councils found troubling evidence that, at least for some offenses, the introduction of the councils coincided with an increase in inter-judge disparity. The study speculated that when judges at a sentencing council “meet, for the first time, opposition to their ideas” about sentencing, the experience “may result in movement to a more extreme position.” Alternatively, judges with more extreme views “may convince moderate judges to follow their more lenient or harsh sentencing patterns.” Better information about other judges’ actions, in other words, did not necessarily bring about consensus and greater inter-judge consistency. It sometimes had the opposite effect, serving as a “catalyst[] for the airing of latent disagreements.”

By revealing the full range of ways that judges may respond to available sentencing information—including disregarding it, avoiding it entirely, or polarizing their views—the experiences of SISs and sentencing councils undermine a key assumption of the information sharing model. As Marc Miller has observed, the power of sentencing information to reduce inter-judge disparity “depends on how judges use the information . . . to guide their own sentencing judgments.” The experiences of SISs and sentencing councils show that judges do not necessarily respond to better information in ways that promote inter-judge consistency and rationality.

IV. MAKING INFORMATION SHARING WORK

Based on new and existing research, there is ample reason for skepticism about the information sharing model. Fundamental weaknesses related to the collection, dissemination, and voluntariness of sentencing information make information sharing highly unattractive as an alternative to other structured sentencing models.

Yet there is also reason for optimism. Information sharing at sentencing may still perform a valuable function as a supplement to other reforms. Despite the mixed track record of previous experiments, it is entirely possible that information sharing can contribute to inter-judge consistency and rationality under the right conditions. In particular, two features of a sentencing system stand a good chance

---

286 See Phillips, supra note 256, at 86–95.
287 Id. at 94.
288 Id.
289 Id. at 100.
290 See Miller, Map and Compass, supra note 1, at 1377–78.
of improving the effectiveness of information sharing: (1) a system of sentencing guidelines, whether advisory or mandatory; and (2) open access to the store of sentencing information for defense counsel and prosecutors. Each of those features helps to address key weaknesses of the information sharing model.

A. Sentencing Guidelines

In light of its weaknesses, the best way to implement the information sharing model may be as an adjunct to a system of sentencing guidelines. As explained above, under a typical set of sentencing guidelines, the judge must make a series of factual determinations about the offense and offender. Based on those facts, the guidelines specify a sentence or a sentencing range. Depending on the jurisdiction, that range may be binding in the absence of extraordinary circumstances, or it may be merely advisory. A set of sentencing guidelines, whether mandatory or advisory, can make information sharing more effective in promoting inter-judge consistency and rationality. That is because guidelines perform several functions— defining terms, operationalizing complex factors, and channeling the attention of sentencing courts—that make information sharing easier.

First, sentencing guidelines define terms and create a shared vocabulary for discussions of sentencing. As Professor Marc Miller has explained, sentencing guidelines “create a language of familiar terms and concepts” among sentencing judges, and that kind of “social language” makes it easier for sentencing judges to understand one another. Guideline systems may define and popularize terms of art like “vulnerable victim,” “substantial assistance,” or “role in the offense,” for example, that offer sentencing judges a succinct way of describing otherwise imprecise categories and concepts.

Second, systems of sentencing guidelines operationalize complex sentencing factors like criminal history. Calculating a guideline sentencing range may depend, for example, on whether the offender has a previous criminal record. Or it may depend on the number of prior convictions or the seriousness of those convictions. Or it may depend on an elaborate scoring system that awards more “criminal history points” for some kinds of offenses (e.g., felonies or violent crimes) than for others (e.g., juvenile offenses or older convictions). In the same manner, guidelines systems operationalize a whole host of other factors, such as the offender’s mental state, role in the offense, and the harm caused to victims. It does not matter, for information sharing purposes, whether the guidelines consider factors in a sensible or principled way. Simply by choosing one method of taking those factors into account, a system of guidelines sets a useful baseline.

291 See supra notes 33–35 and accompanying text.
292 Miller, Map and Compass, supra note 1, at 1372–73.
293 Id.
Third, sentencing guidelines channel the court’s attention to a standard set of facts and considerations. Some systems, like the U.S. Sentencing Guidelines, attempt to account for a dizzying number of factors, forcing judges to engage in extensive and intricate fact finding. Other systems are much simpler, taking into account fewer facts and circumstances and making more general recommendations. But all guidelines systems focus the court’s attention on some set of especially salient offense and offender information. Even if the guideline sentencing range is advisory, the process of determining that range is important.

Those basic functions of a guidelines system help to address some of the major weaknesses of the information sharing model. One set of challenges relates to the collection of sentencing information. The complexity and volume of sentencing decisions makes it difficult to collect written, comprehensive, and representative sentencing information. By channeling the efforts of sentencing courts and requiring that the judge always address some standard set of questions, sentencing guidelines ensure the availability of a basic level of information about every case. That minimum level of information reduces (although it certainly does not eliminate) the risk that noncomprehensive information might mislead future sentencing courts and thereby generate inter-judge disparity. Guidelines also guard against a skewed and nonrepresentative pool of cases because the standard set of findings is required from all judges and in all kinds of cases. It is also possible that routinely answering a standard set of questions may also change sentencing judges’ habits in a way that encourages more written opinions, although overly complex guidelines may have the opposite effect.

Sentencing guidelines also help to address challenges with the dissemination of sentencing information. By operationalizing important sentencing factors, sentencing guidelines provide a baseline set of categories and concepts that judges can use when searching for similar and relevant cases. In addition, by supplying a common vocabulary among sentencing judges, guidelines can make search results easier to understand, reduce errors based on inconsistent terminology and generally reduce the risk of user frustration. Of course, a sentencing information system need not simply replicate the categories and concepts developed in the guidelines. Still, a framework of sentencing guidelines, with its shared language and baseline understanding of key factors, improves the chances that judges can use the information effectively.

Admittedly, to date no research has established that information sharing can reduce inter-judge disparity or promote rationality. At the same time, however, the experiences of SISs and sentencing councils do not foreclose the possibility of successful information sharing within a system of sentencing guidelines. In each of

---

296 See supra Part II.A.
297 See supra Part II.B.2.
298 Cf. Miller, supra note 42, at 143–46 (arguing that SISs need not be merely duplicative of sentencing guidelines).
the jurisdictions where SISs have failed, they were stand-alone systems in which judges received little guidance about sentencing. Similarly, sentencing councils operated in a pre-reform era in federal court, with little external guidance for sentencing judges. Perhaps it is no coincidence that the most successful information sharing experiments, such as the SIS developed in New South Wales, have survived alongside other forms of structured sentencing like guideline judgments and standard nonparole periods. Because of the basic functions they perform, a system of sentencing guidelines likely would improve the chances that information sharing can succeed.

**B. Open Access**

Another way to improve the information sharing model is to publicize the store of sentencing information. The body of written opinions and statistical information, along with any specialized search tools, could be made available to defense counsel and prosecutors—not just the court. Publicly sharing information about sentencing would have obvious benefits for the criminal justice system as a whole. But open access would be of particular value in overcoming challenges related to voluntariness and the dissemination of sentencing information.

As discussed above, one major obstacle to effective information sharing at sentencing is time. Sentencing judges cannot afford routinely to sift through mountains of information about previous cases to ensure that their decisions are compatible with those of their colleagues. The experiences of sentencing councils and SISs reveal when judges decline to consult available information, they frequently cite time constraints as a primary reason. The complexity and volume of sentencing decisions, which make it difficult to design user-friendly search tools, can compound the problem by making judges less inclined to track down relevant cases.

Open access would ease the burden on judges by transferring much of the case-matching legwork to the lawyers. In their briefs and at sentencing hearings, the parties could take the lead in identifying potentially relevant cases. That would reduce the risk of errors in identifying relevant cases, thereby promoting inter-judge consistency. Although open access would mean more work for counsel, there is reason to think that lawyers would avail themselves of the information. A

---

299 Id. at 131–132.


301 See supra notes 229–233 and accompanying text.

302 Miller, supra note 42, at 146–48 (urging the development of “democratic, participatory, transparent” sentencing information systems available to “lawyers, judges, scholars, and reformers” alike).

store of information about previous sentences would be one more weapon in counsel’s arsenal in preparation for sentencing. As an example, a decade ago, the Office of the Federal Public Defender in the District of Massachusetts began compiling a publicly available collection of downward departure decisions, sorted by offense type and judge, as a way to improve advocacy on behalf of criminal defendants.304

At the same time, open access would increase the likelihood that the court will take the information seriously. At sentencing, as in other contexts, judges make a point of responding to the specific contentions raised by the parties. Although the information sharing model by design does not compel judges to give weight to any particular factors at sentencing, even judges who would not independently seek out information about previous sentences might respond to a well-formed argument by the prosecutor or defense counsel.

In addition, open access could promote rationality by improving the quality of the judge’s reasoning. Vigorous advocacy is a great asset to a sentencing court. The parties can draw parallels between cases, debate possible points of distinction, and urge the judge to accept or reject prior judges’ reasoning. It is possible, of course, that the parties might use the store of sentencing information to stake out extreme positions that provide little help to the court.305 Yet adversarial testing of those arguments in principle should help judges reach more thoughtful and principled outcomes.

The major drawback to open access is the privacy of sentencing information. Written opinions and other sentencing documents may disclose deeply personal information about offenders, victims, family members, and witnesses. For offenders and witnesses who cooperate with the government, open access might even create a risk of violent retaliation.306

It may be possible to ameliorate those privacy concerns by partially withholding or redacting sensitive information made available to others.307 But too much redaction could defeat the purpose, undermining the effectiveness of the information sharing model. For example, in an attempted murder case, the judge may impose a severe sentence because of the serious long-term physical or psychological harm to the victim. Withholding that sensitive information would protect the victim’s privacy, but would leave others guessing about the most important reasons for the sentence. Because personal information often plays a

305 In civil actions for punitive damages, for example, plaintiffs and defendants notoriously draw comparisons with cases involving either extremely high awards or no award at all, sometimes leaving trial courts with little discussion of “middle ground” cases that might serve as better comparisons.
306 See supra Part II.B.1.
critical role in sentencing outcomes, concealing that information may render written opinions unhelpful, or even incomprehensible, to future judges and lawyers. Although there is no easy way of determining how to withhold or redact sentencing information, two strategies might prove useful. First, because most sentencing information does not raise serious privacy concerns, case information could be presumed accessible unless the parties or the court request that it be sealed. Defense attorneys and prosecutors would have incentives not to overuse that power, since they stand to benefit from more complete information in future cases. Second, the system could be designed to withhold private information only from the parties, while allowing judges access to complete information. That way, even if counsel does not fully understand the facts and reasoning of a previous decision, the sentencing judge can take them into account.

Neither a system of sentencing guidelines nor open access to sentencing information can guarantee the success of the information sharing model. Its fundamental weaknesses, along with the poor track record of previous reform efforts, provide ample reason for skepticism. Yet sentencing guidelines and open access would help to address challenges related to the collection, dissemination, and voluntariness of sentencing information. They give the information sharing model its best chance to succeed.

CONCLUSION

The information sharing model is often advertised as a method of reducing inter-judge disparity and promoting rationality in sentencing law. Proponents argue that by assembling a body of written opinions and other information about past sentences, judges can align their sentences with those of their colleagues. This Article has identified three fundamental weaknesses in that model. First, there are daunting information-collection challenges. Because of the complexity and volume of sentencing decisions, it is difficult for courts to generate sentencing information that is written, comprehensive, and representative. Second, there are challenges in disseminating sentencing information in a useful way. Privacy interests on the part of offenders and others raise serious concerns. In addition, as a practical matter, it is difficult to make the large volume of relevant information available to judges in a useful format. Third, the voluntariness of the information sharing model is an important drawback. Judges retain the discretion to ignore or reject colleagues’ reasoning, undermining inter-judge consistency and rationality.

New and existing research reinforces each of those weaknesses. The Article contains an original study of information-collection practices in the only federal court that makes key sentencing documents public. It finds that, despite mandatory reporting requirements, the court rarely provides the kind of written explanation needed to support the information sharing model. The history of two other reform efforts, sentencing information systems and sentencing councils, reveals how voluntariness and challenges with information dissemination can frustrate information sharing at sentencing. As the FJC study of sentencing councils
concluded, history is discouraging for “those who see disparity as a problem that can be solved by better communication among judges.”\textsuperscript{308}

More than anything, this “skeptic’s guide” makes a plea for realism about what the information sharing model can accomplish. But information sharing at sentencing is in its infancy, and therefore the claim here is modest. There is reason for skepticism that information sharing can serve as an alternative to other structured sentencing reforms aimed at improving inter-judge consistency and rationality. But information sharing has real potential as a supplement to other efforts. In particular, its odds of success would greatly improve if implemented as part of a system of sentencing guidelines and with open access to the information for defense counsel and prosecutors.

\textsuperscript{308} PHILLIPS, \textit{supra} note 256, at 100.
This Appendix provides additional details concerning case selection and coding for the empirical study, along with detailed regression results.

A. Case Selection

The study examines Statement of Reasons documents (SORs) for sentences imposed in the United States District Court for the District of Massachusetts in fiscal year 2006, which runs from October 1, 2005, to September 30, 2006. As noted above, the District of Massachusetts is the only federal district court that makes the documents public, by posting them on the Public Access to Court Electronic Records (PACER) system as part of the case docket.309

The SORs were collected as part of a related study of inter-judge sentencing disparity, and the Technical Appendix to that article provides background information.310 PACER supports case searches by filing date and closing date, but not by the date of sentencing.311 To identify cases that may include a sentence imposed in fiscal year 2006, the initial search extended to every criminal case filed in the district between January 1, 2000, and June 30, 2006.312 The vast majority of results, including dismissals, jurisdictional transfers, and acquittals, were ignored because they did not produce a sentence during the relevant period.

The result was a body of 411 SORs for sentences imposed in fiscal year 2006. The Sentencing Commission reports that the district as a whole submitted documentation for 512 sentences that year.313 The SORs examined here therefore represent 80.3% of the total. There are several possible explanations for the missing SORs. First, although the District of Massachusetts generally makes SORs available on PACER, in some cases the SOR is unavailable or the docket indicates that the SOR is sealed. Judges retain the power to seal the SOR for case-specific reasons, such as the protection of offenders or witnesses who have cooperated with the government.314 Second, cases filed before 2000 with a sentence in 2006—for example, cases extended by appeal and remand or complex conspiracy cases with many defendants—would have escaped the initial search. Thus, although the available set of 411 documents includes more than 80% of the total, it is not complete.

309 See supra notes 178–180 and accompanying text.
311 The date of sentencing rarely matches the closing date because of multiple-defendant cases, appeals, supervised release hearings, and other post-sentence filings.
312 PACER’s “Reports” tool allows searches by “Case Type,” including criminal cases. The search included pending and terminated defendants, but excluded cases involving fugitive defendants.
313 See U.S. SENTENCING COMM’N, supra note 203, at 211 (reporting a total of 512 Statements of Reasons received from the District of Massachusetts).
Nor are the available SORs necessarily a random sample of the total. It is possible that SORs that do not appear on PACER differ in important ways, such as the level of written explanation provided, from the SORs that are available. Fortunately, as Table A1 indicates, the SORs available on PACER differ only slightly from the total population of SORs submitted to the Commission:

Table A1: Comparison of All Cases and Available SORs
District of Massachusetts, FY 2006

<table>
<thead>
<tr>
<th></th>
<th>All Cases FY 2006</th>
<th>SOR Available on PACER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Cases</td>
<td>516</td>
<td>411</td>
</tr>
<tr>
<td>Imprisonment Ordered</td>
<td>87.6%</td>
<td>81.8%</td>
</tr>
<tr>
<td>Average Prison Sentence</td>
<td>73.0 months</td>
<td>78.1 months</td>
</tr>
<tr>
<td>Within Guideline Range or Government-Sponsored Departure</td>
<td>74.1%</td>
<td>67.4%</td>
</tr>
<tr>
<td>Below Guideline Range</td>
<td>24.8%</td>
<td>29.7%</td>
</tr>
<tr>
<td>Above Guideline Range</td>
<td>1.2%</td>
<td>2.9%</td>
</tr>
</tbody>
</table>

Although modest, potential differences between the available SORs on PACER and the full pool of SORs submitted to the Commission are a reason for caution.

B. Coding

To measure the level of written explanation contained in each SOR, a research assistant counted the total number of sentences of text provided. The count included written explanations from all three of the SOR form’s narrative description fields. It also included written explanations contained in any attachments, such as published sentencing opinions and hearing transcripts. In those cases, however, the count included only portions of the opinion or proceeding in which the judge provided reasons for the sentence imposed. It excluded, for example, discussion of guideline calculations, constitutional challenges, and procedural matters.

In some cases, the SOR contains only a generic statement with no discussion of the particular offense or offender. As an example, some SORs state: “I have considered the factors set forth in § 3553(a) and I have imposed a sentence

---

315 See U.S. SENTENCING COMM’N, supra note 203, at 211.
sufficient, but not greater than necessary, to achieve those purposes.” Because generic statements of that kind merely restate the judge’s task, without providing any explanation of the sentence imposed, they were coded as zero sentences of explanatory text.

In addition, SOR forms were coded for whether any checkboxes were marked to explain the sentence. The checkbox condition was considered satisfied if any checkbox was marked in any section of the SOR form that lists reasons for the sentence. Checkboxes unrelated to the reasons for the sentence, such as those describing whether a fine was imposed or the presentence report was adopted, did not satisfy the checkbox condition.

The study reports the results of two logistic regression models, each based on a different dependent variable. Logistic regression was necessary because both dependent variables are binary rather than normally distributed. The first is whether the SOR contains any written explanation of the sentence. If the SOR contains zero sentences of explanatory text, then the dependent variable is coded as zero. If the SOR contains one or more sentences of explanatory text, then the dependent variable is coded as one. A mark in a checkbox was not coded as a form of written explanation.

The second is whether the SOR contains a long written explanation, consisting of at least ten sentences of text. That cutoff corresponds to roughly one page of narrative description, and was considered a reasonable proxy for the kind of explanation sufficient to provide meaningful information about the reasons for the sentence to another judge. If the SOR contains ten or more sentences of explanatory text, then the dependent variable is coded as one. If the SOR contains fewer than nine sentences, then the dependent variable is coded as zero.

Both models include the following independent variables:

(1) **Sentence length, in months.** Sentence length is coded as length of the term of imprisonment, measured in months. Consistent with the Sentencing Commission’s practice, sentences of probation were coded as zero months of imprisonment.

(2) **Nonguideline sentence, an indicator.** Nonguideline sentences include sentences above or below the guideline sentencing range, whether styled as a “departure,” a post-Booker “variance,” or using another term. Following the Commission’s conventions, government-sponsored below-range sentences were coded as within range.

(3) **Within-range sentences of more than twenty-four months, an indicator.** This dummy variable was coded because the SOR form requires a written narrative explanation for that category of sentences, but not for other within-range sentences.

(4) **Sentence of time served, an indicator.** Sentences of time served present special challenges. In the federal system, offenders may receive credit for time

---

316 The statement closely tracks 18 U.S.C. § 3553(a) (2006), which sets out factors that a judge must consider at sentencing.

317 See STITH & CABRANES, supra note 3, at 62 (employing the same convention).
served in official detention prior to sentencing. 318 It is common for a judge to impose a sentence of time served, allowing the offender to be released immediately. The sentence is not zero months of imprisonment because the offender is credited by statute for serving time. But SOR forms sometimes list the sentence simply as “time served,” with no indication of the length of presentence detention for which the offender has been credited. Lacking more precise information, sentences of time served were coded as zero months of imprisonment, but an additional dummy indicator was added to permit analysis of those cases.

In addition, one model uses dummy variables that capture the identity of the sentencing judge. To avoid the distorting effects of judges with low caseloads, such as those in senior status, those models exclude sentences by judges with fewer than fifteen sentences during the fiscal year. 319 Following standard practice for categorical variables, one judge was omitted as a reference category. The second model, based on long explanations, does not include judge dummy variables because too many judges submitted no long explanations during the year of the study.

C. Regression Results

The first model describes factors that predict whether the Statement of Reasons contains any written explanation. Table 1 reports the results:

Table 1: Logistic Regression Model, Written Explanation Provided

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>-1.153</td>
<td>0.676</td>
<td>.088</td>
</tr>
<tr>
<td>Sentence Length</td>
<td>0.007</td>
<td>0.002</td>
<td>.001*</td>
</tr>
<tr>
<td>Sentence of Time Served</td>
<td>0.409</td>
<td>0.900</td>
<td>.649</td>
</tr>
<tr>
<td>Nonguideline Sentence</td>
<td>3.320</td>
<td>0.392</td>
<td>&lt; .001*</td>
</tr>
<tr>
<td>Within-Range Over 24 Months</td>
<td>0.719</td>
<td>0.579</td>
<td>.214</td>
</tr>
<tr>
<td>Judge Identity (categorical)</td>
<td></td>
<td></td>
<td>.001*</td>
</tr>
</tbody>
</table>

Model significance: < .001*  
* Significant at the .05 level  
Chi-square: 183.866  
n = 398

The model indicates that some kinds of sentences are more likely to produce a written explanation than others. Nonguideline sentences, longer terms of imprisonment, and the identity of the judge are significantly correlated with written explanations. A within-range sentence of more than twenty-four months, however, is not a significant predictor of a written explanation, after controlling for other factors. This casts doubt on the effectiveness of SOR forms’ instructions, which require a written explanation in those circumstances.

The second model describes factors that predict whether the Statement of Reasons contains a long written explanation of ten or more sentences. Table 2 reports the results:

Table 2: Logistic Regression Model, Long Explanation (10+ sentences) Provided

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>-4.495</td>
<td>1.006</td>
<td>&lt; .001*</td>
</tr>
<tr>
<td>Sentence Length</td>
<td>-0.003</td>
<td>0.004</td>
<td>.482</td>
</tr>
<tr>
<td>Nonguideline Sentence</td>
<td>1.086</td>
<td>0.519</td>
<td>.037*</td>
</tr>
<tr>
<td>Within-Range Over 24 Months</td>
<td>1.338</td>
<td>1.124</td>
<td>.234</td>
</tr>
</tbody>
</table>

Model significance: < .001*  
* Significant at the .05 level  
Chi-square: 11.185  
n = 408
Nonguideline sentences are significantly correlated with long explanations. Neither sentence length nor within-range sentences above twenty-four months, however, is a statistically significant predictor of a long explanation.
THE PHANTOM STANDARD: COMPELLING STATE INTEREST ANALYSIS AND POLITICAL IDEOLOGY IN THE AFFIRMATIVE ACTION CONTEXT

Timothy M. Bagshaw*

INTRODUCTION

The degree to which legal doctrine constrains judicial decisionmaking is one of the enduring questions in the study of judicial behavior. Scholars and the public alike have long debated whether and to what extent judges’ personal political ideologies affect their legal decisions. The most influential existing theories have explained judicial holdings as a function of judges’ ideological predispositions. But in conceptualizing holdings as the relevant dependent variable, these theories suffer from one central flaw: they assume that legal tests are ideologically neutral filters through which judges’ policy preferences pass. This assumption obscures the ideological work being done in the creation of particular legal tests in the first instance.

This Note reconceptualizes the role that political ideology plays in the judicial decisionmaking process. Rather than focusing on the outcomes of cases, the discussion below instead redirects attention to the ways in which ideology influences the creation of an important—and ostensibly ideologically neutral—doctrinal test: strict scrutiny. Strict scrutiny analysis is the most rigorous form of judicial review. Under this standard, state action is presumed unconstitutional and will survive only if it is narrowly tailored to achieve a compelling governmental interest. But there is no standard for determining whether an asserted governmental

---

* © 2013 Timothy M. Bagshaw. J.D. Candidate 2014, University of Utah, S.J. Quinney College of Law.

1 See JEFFREY A. SEGAL & HAROLD J. SPAETH, THE SUPREME COURT AND THE ATTITUDINAL MODEL REVISITED 110–14 (2002). Professors Jeffrey Segal and Harold Spaeth find substantial empirical evidence that justices decide cases in light of their sincerely held ideological values:

Attitudinalists argue that because legal rules governing decision making (e.g., precedent, plain meaning) in the cases that come to the Court do not limit discretion; because the justices need not respond to public opinion, Congress, or the President; and because the Supreme Court is the court of last resort, the justices, unlike their lower court colleagues, may freely implement their personal policy preferences as the attitudinal model specifies.

Id. at 111.

2 Id.
interest is compelling for purposes of the analysis; as such, justices have wide
discretion in this field. Under this standardless regime, the justices may credit
particular governmental interests as compelling, to the exclusion of others; in so
doing, their choices may reflect their ideological predispositions. Therefore, even
where an opinion adheres closely to the requirements of strict scrutiny, it may
nonetheless privilege particular ideological understandings.

To illustrate this general theory—that ideological predispositions are
endogenous to the legal standards designed to resolve complex questions of
constitutional law—this Note offers by way of example the field of race-based
affirmative action in higher education. The affirmative action setting is illustrative
and timely. First, over the course of the Court’s jurisprudence in this area, among
the many possible interests that public universities have advanced in support of
their affirmative action, the Court has only recognized diversity as sufficiently
compelling to satisfy strict scrutiny. The choice of diversity, to the exclusion of
other potential governmental interests, reflects particular ideological
understandings about the way race and racism operate in American institutions. In
this respect, affirmative action represents a useful test case to give concrete
meaning to the notion that the Court can craft the strict scrutiny test in such a way
as to privilege its own ideological preferences.

Second, with its recent decision in Fisher v. University of Texas at Austin, the
Court has cast into some doubt the constitutionality of race-based affirmative
action, a question that had been settled law for a decade. In Fisher, the Court
reviewed the constitutionality of the University of Texas’s affirmative action
policy, a policy that differed from the one the Court upheld in Grutter v. Bollinger
only insofar as the university also employed a Top Ten Percent plan in admissions
to the flagship campus. The Court reaffirmed the principle that diversity alone
constitutes a compelling governmental interest for purposes of strict scrutiny, but in doing so put public universities on notice that strict scrutiny would

---

dissenting) (recognizing there is no doctrinal standard guiding compelling state interest
analysis).


5 133 S. Ct. 2411 (2013).

6 See Grutter, 539 U.S. at 343. The Court upheld race-based affirmative action in
higher education when race constitutes a “plus” factor as a part of a holistic review of
candidates, but included a sunset provision. It stated, “We expect that 25 years from now,
the use of racial preferences will no longer be necessary to further the [diversity] interest
approved today.” Id.

7 133 S. Ct. 2411.

general teaching institution” to any student who graduates in the top 10% of her class from
any qualifying high school).

9 Fisher, 133 S. Ct. at 2419.
be strict indeed. Fisher’s emergence presents a timely opportunity to reexamine one of the enduring theoretical problems in the study of judicial politics: how and to what extent does legal doctrine constrain judicial behavior?

The question of whether the law constrains the expression of political ideology in judicial decisionmaking has received significant scholarly attention. Two competing models dominate this debate. First, under the attitudinal model, scholars identify the presence of ideology in the decisionmaking process by measuring justices’ preexisting political preferences and then determining whether their decisions (or votes) in cases match their ideological preferences. In its simplest form, the attitudinal model operates in the following way: ceteris paribus, where a Justice’s vote in a case is consistent with her underlying policy preferences, this is presented as evidence that the law has not constrained the expression of her ideology. Where, however, her vote in a case is inconsistent with her underlying policy preferences, this is presented as evidence that the law has constrained the expression of the Justice’s ideology; that is, but for the constraints imposed by the legal doctrine, her vote would have been consistent with her underlying policy preferences.

From the attitudinalist perspective, legal rules are conceptualized as neutral institutional mechanisms that may constrain the expression of political ideology, but are more likely to operate simply as obstacles around which judges must navigate to express their policy preferences. The attitudinal model commands

---

10 Id. at 2419–22 (reaffirming Grutter but remanding to the Fifth Circuit for reconsideration of whether the university’s affirmative action plan was necessary to achieve its compelling interest in racial diversity in light of racial diversity that is independently achieved under the Top Ten Percent plan).

11 Affirmative action is not the only field that might motivate the general theory that ideology is endogenous to strict scrutiny’s ends-based analysis. Fields ranging from abortion, to lesbian, gay, bisexual, and transgender rights, to physician-assisted suicide, among others, would provide useful test cases as well. Given Fisher’s emergence, however, affirmative action constitutes a timely and interesting example.

12 Jeffrey A. Segal & Albert D. Cover, Ideological Values and the Votes of U.S. Supreme Court Justices, 83 AM. POL. SCI. REV. 557, 559–63 (1989) (finding a correlation between Justices’ ideological values, as determined by newspaper editorials before their confirmation hearings, and the Justices’ opinions). But see Tracey E. George & Lee Epstein, On the Nature of Supreme Court Decision Making, 86 AM. POL. SCI. REV. 323, 328 (1992) (“Although previous research accounts for internal value changes in diverse ways, one acceptable procedure operates under the assumption that as the number of Nixon, Ford, and Reagan appointees increase, so should the number of conservative . . . outcomes.”).

13 See SEGAL & SPAETH, supra note 1, at 86 (“[T]he Supreme Court decides disputes in light of the facts of the case vis-à-vis the ideological attitudes and values of the justices. Simply put, Rehnquist votes the way he does because he is extremely conservative; Marshall voted the way he did because he was extremely liberal.”).

14 This theoretical assumption—that legal rules are ideologically neutral institutional features—is not unique to the attitudinal model. Other behavioral studies evaluating the constraining role of legal rules on courts also retain this key assumption. See Brandon L.
empirical support and has intuitive appeal.\textsuperscript{15} Under this conventional view, where a legal opinion reads as if the court is adhering narrowly to the prescribed legal standard,\textsuperscript{16} it is tempting to conclude that a court has sublimated its policy preferences in fealty to the law. This otherwise reasonable inference depends upon the assumption that political preferences drive ideological decisions, but the doctrinal link—for example, the strict scrutiny test—between them is ideologically neutral.\textsuperscript{17}

New institutionalism represents a second, competing approach to the attitudinal model. From the new institutionalist perspective, institutional norms, rules, and doctrines exert an independent effect on judicial preferences.\textsuperscript{18} Even though justices possess ideological preferences, their behavior is not determined solely, or even primarily, by those preferences; instead, judicial behavior is structured by a complex set of strategic, institutional, and interpersonal norms.\textsuperscript{19} These are valuable insights, but the new institutionalism nonetheless conceptualizes legal rules as independent variables, which deflects attention from the ways in which legal rules are themselves shaped by the values and attitudes that justices bring to the decisionmaking process.

Although the attitudinal and new institutionalist approaches have generated important insights, they are unnecessarily limited and should not exclusively govern how we conceptualize the role of ideological preferences in judicial decisionmaking. Legal rules and doctrines serve as neutral, constraining (or not) forces, but they are also the product of political and ideological contestation. Even where judges adhere to the legal rules prescribed to resolve specific legal problems, the rules may nonetheless privilege particular political understandings. Specifically, the Court has never articulated a test—or any cognizable set of criteria—for determining what constitutes a compelling state interest for purposes

\textsuperscript{15} See Forrest Maltzman et al., \textit{Strategy and Judicial Choice: New Institutionalist Approaches to Supreme Court Decision-Making}, in \textit{SUPREME COURT DECISION-MAKING: NEW INSTITUTIONALIST APPROACHES}, supra note 15, at 476–78 (2009) (modeling how different standards of review constrain the choices of Supreme Court Justices but retaining the assumption that the standards of review are, themselves, ideologically neutral).

\textsuperscript{16} In this context, “adhering narrowly to the prescribed legal standard” refers to settings in which judges’ opinions (or votes) are inconsistent with their perceived political policy preferences.

\textsuperscript{17} This Note treats the terms “ideology” and “political policy preferences” synonymously, each to suggest that justices are predisposed to particular legal positions ranging from politically liberal to politically conservative.


\textsuperscript{19} See id.
of strict scrutiny analysis. This Note highlights how the Court has unconstrained discretion to determine what governmental interests are sufficient to satisfy rigorous judicial review, a process that privileges the justices’ own particular understandings of, for example, race and racial discrimination at the expense of the judgment of political institutions in the race-based affirmative action setting.

To this end, Part I summarizes a survey of the scholarly literature on the role of ideology in the judicial decisionmaking process with an eye toward exploring current understanding of the extent to which legal doctrine constrains the expression of justices’ policy preferences. Existing theoretical approaches are valuable, but ultimately incomplete. As such, Part II advances an alternative approach by emphasizing the role that particular political understandings play in shaping the development of the strict scrutiny test. This section emphasizes how the lack of any cognizable standard for compelling state interest analysis in the strict scrutiny setting creates opportunities for justices to introduce their policy preferences into the test. Part III analyzes the field of race-based affirmative action in higher education to explore this theoretical approach in a concrete doctrinal setting. This section demonstrates how the Court has credited diversity as a compelling state interest in this context, to the exclusion of all other asserted interests. Finally, Part IV argues that the Court’s decision to promote diversity as the only compelling state interest in the affirmative action setting reflects particular ideological understandings about the nature of institutional discrimination.

I. THE SCHOLARLY DEBATE

Rules and norms operate as important constraints on decisionmaking by restricting actors’ range of possible choices within their institutional roles. This is especially true of courts. On the Supreme Court, justices’ decisions are constrained almost entirely by a self-imposed set of norms, rules, and traditions—such as stare decisis—as well as various legal rules and doctrines that govern the outcomes of cases. Scholars in the social sciences and the legal academy debate

---

22 See Lee Epstein & Jack Knight, The New Institutionalism, Part II, 7 LAW & CTS. (Law & Courts Section of the Am. Political Sci. Review, Newark, Del.), Spring 1997, at 4 (arguing that even strategic justices “realize that their ability to achieve their goals depends on a consideration of the preferences of other actors, of the choices they expect others to make, and of the institutional context in which they act”).
23 See March & Olsen, supra note 21, at 740, 744 (describing the constraining impact of institutional norms and rules and how internal institutional processes condition actors’ expectations); David M. O’Brien, Institutional Norms and Supreme Court Opinions: On Reconsidering the Rise of Individual Opinions, in SUPREME COURT DECISION-MAKING: NEW INSTITUTIONALIST APPROACHES, supra note 15, at 91, 91 (“Like other political
whether and to what extent these institutional rules and norms restrict the range of possible decisions that justices make and whether the rules constrain the justices’ own political ideologies in their decisionmaking.24

A. Competing Traditions: The Attitudinal Model and New Institutionalism

A central motivating question for students of judicial behavior is whether and to what extent the ideological attitudes and political preferences of individual justices can predict and explain how justices vote on the central sociolegal issues of the day. Attitudinalists argue that justices’ policy preferences are highly predictive of their votes. New institutionalists, on the other hand, focus their analytic lens on the Supreme Court as an institution, attempting to discover how institutional rules, norms, and relationships shape the justices’ preferences, and thus their behavior. While these attitudinalists and new institutionalists differ in how they approach the question of judicial behavior, each is concerned with one central question: To what extent and in what ways do legal rules constrain judicial decisionmaking on the Supreme Court?

1. The Attitudinal Model

The attitudinal model holds that justices’ votes can be explained almost entirely by their political preferences.25 As a research agenda and theoretical approach, the attitudinal model has its roots in the legal realist movement of the 1920s.26 Legal realism represented a radical shift away from the view “that the state, the economy, and society reflected human nature” and toward the idea that institutions “were constructed according to historically evolved patterns” that were themselves “open to modification.”27 Legal realists “dispatched traditional mechanistic and formalistic conceptions of the law, and instead emphasized the institutions, the Supreme Court of the United States has its own unique set of internal practices, norms, and traditions.”

24 Compare Cornell Clayton, Law, Politics, and the Rehnquist Court: Structural Influences on Supreme Court Decision Making, in The Supreme Court in American Politics: New Institutionalist Interpretations 151, 176 (Howard Gillman & Cornell Clayton eds., 1999) (“[T]hough it is true that law and politics are inescapably linked, it is fanciful to assume that judges vote their policy preferences in the same way as would a member of Congress or a state legislator. Politics enters the judicial process in deeper, more constitutive ways, shaping the way judges think about law.”), with SEGAL & SPAETH, supra note 1, at 86 (describing the attitudinal model as an alternative explanation of judicial decisionmaking and stating that under this model ideology matters).

25 SEGAL & SPAETH, supra note 1, at 110–11.

26 See Clayton, supra note 18, at 16–22 (tracing the lineage between legal realism and the attitudinal model). It should come as no surprise that the legal realist movement developed in the 1920s, as scholars began to reflect upon the jurisprudence of the *Lochner* era.

creativity found in judging. . . . [This shifted the conception towards] the now well known critique that judges did more than simply find law, they also made law.”

Under this view, “law could no longer be understood in isolation but now had to be considered in light of larger political, economic, and social background structures.” Karl Llewellyn, an intellectual pioneer of this movement, discussed the “neutral” doctrine of stare decisis this way:

> What I wish to sink deep into your minds about the doctrine of precedent, therefore, is that it is two-headed. It is Janus-faced. That it is not one doctrine, nor one line of doctrine, but two, and two which, *applied at the same time to the same precedent, are contradictory of each other*. That there is one doctrine for getting rid of precedents that seem troublesome and one doctrine for making use of precedents that seem helpful. That these two doctrines exist side by side. That the same lawyer in the same brief, the same judge in the same opinion, may be using the one doctrine, the technically strict one, to cut down half the older cases that he deals with, and using the other doctrine, the loose one, for building with the other half. Until you realize this you do not see how it is possible for law to change and to develop, and yet to stand on the past.

In focusing on the predictive and explanatory power of justices’ political preferences, attitudinalists share with legal realists deep theoretical assumptions that judges are political actors who use the law to accomplish those particular ends. Each is committed to a political understanding of the legal process.

In their seminal empirical work, *The Supreme Court and the Attitudinal Model*, Professors Segal and Spaeth discovered that justices’ votes on particular
legal issues were stable over time and that their voting patterns were highly correlated with their prior political ideologies, as measured by newspaper accounts of their policy views prior to confirmation to the Court. From these findings, they concluded that institutional norms and legal rules possess little explanatory value; that is, the law does not particularly matter in shaping judicial outcomes, at least not relative to the explanatory power of justices’ political preferences. In another study, Professors Segal and Spaeth found that stare decisis fares poorly in predicting the patterns of judicial decisionmaking. Despite the robust empirical results yielded by the attitudinal model, it has nonetheless been subject to extensive critiques, some of which are empirical and some of which are theoretical; it is the latter category that is of greatest concern here.

modeling how particular justices would vote on cases on the basis of their political ideologies).

33 SEGAL & SPAETH, supra note 1.

34 Clayton, supra note 18, at 22. This measurement strategy was, in some ways, one of necessity, and in another sense one of methodological strategy. As a practical necessity, most justices will keep silent on the important legal issues of the day once they are nominated and then confirmed, so there is no clear way to know a justice’s personal views on matters of politics or law after that point. With respect to the methodological strategy involved in the measurement, looking to preconfirmation public statements was thought to allow for the control of Court norms and rules; thus, in some important ways, the measurement itself concedes at least the possibility that institutional norms—including the law—may exert an independent influence on judicial decisionmaking. Id. at 16.

35 See Jeffrey A. Segal, Supreme Court Deference to Congress: An Examination of the Marxist Model, in SUPREME COURT DECISION-MAKING: NEW INSTITUTIONALIST APPROACHES, supra note 15, at 237, 238. It would be misleading to state that attitudinalists argue that the law is wholly unrelated to decisionmaking because “[n]arrowly speaking, the attitudinal model holds that justices decide cases on the merits in light of the facts of the case vis-à-vis their sincere ideological attitudes and values.” Id.

36 See Jeffrey A. Segal & Howard J. Spaeth, The Influence of Stare Decisis on the Votes of United States Supreme Court Justices, 40 AM. J. POL. SCI. 971, 983 (1996) (finding that 90.8% of justices’ votes on selected landmark cases conformed to their revealed preferences, concluding that justices are infrequently bound by stare decisis).

37 Professors Segal and Spaeth explicitly avoid making claims about generalizability. See id. at 973. Segal and Spaeth limit their conclusions to the Supreme Court because of the special institutional position it occupies. “[B]ecause the Supreme Court sits atop the judicial hierarchy, and because in the type of cases that reach the Supreme Court legal factors such as text, intent, and precedent are typically ambiguous, justices are free to make decisions based on their personal policy preferences.” Id.

38 See, e.g., Lawrence Baum, What Judges Want: Judges’ Goals and Judicial Behavior, 47 POL. RES. Q. 749, 756–57 (1994) (critiquing the attitudinal model for using newspaper editorials as evidence of ideological decisionmaking because newspaper accounts are only indirect measures of justices’ policy preferences, and critiquing the assumption that consistent votes for certain policy positions constitutes evidence of ideological voting, rather than voting based upon a coherent judicial or legal philosophy).

39 See, e.g., Clayton, supra note 18, at 26–30.
2. **New Institutionalism**

New institutionalism is the central competing theoretical approach. New institutionalism acknowledges that political ideology influences justices’ choices, but it emphasizes instead how institutional norms, rules, and doctrines exert an independent influence on voting patterns. According to this approach, the law constitutes a framework structuring judicial decisionmaking. New institutionalism rejects the view that the justices’ policy preferences are the primary motivator of their behavior; rather, it “seeks to understand legal and judicial institutions as independent variables that both constitute and constrain judicial attitudes and motivations.” From the new institutionalist perspective, judicial decisionmaking is conceptualized as a process “in which judicial values and attitudes are shaped by judges’ distinct professional roles, their sense of obligation, and salient institutional perspectives.” New institutionalism’s insights are valuable insofar as they speak to the constraining effect of institutional norms, rules, and relationships on the expression of justices’ preferences. But in conceptualizing judicial institutions—particularly legal rules—as an independent variable, new institutionalism deflects attention away from the ways in which the values and attitudes that justices bring to the decisionmaking process shape the law. Put another way, legal rules are themselves dependent upon the ideological preferences of their creators.

---

40 See id. at 29 (“[N]o serious student of judicial behavior can now doubt that attitudes make a difference to judicial decision-making . . . .”).


42 See, e.g., Clayton, supra note 18, at 30–35.

43 Cornell Clayton & Howard Gillman, Introduction to THE SUPREME COURT IN AMERICAN POLITICS: NEW INSTITUTIONALIST INTERPRETATIONS, supra note 24, at 1, 4.

44 Clayton, supra note 18, at 32.


46 In this respect, the Note advocates resurrecting many of the insights of the “old institutionalism” that evolved concurrently with legal realism. See, e.g., ROBERT EUGENE CUSHMAN, LEADING CONSTITUTIONAL DECISIONS, at iii (1925) (“The Supreme Court does not do its work in a vacuum. Its decisions on important constitutional questions can be understood in their full significance only when viewed against the background of history, politics, economics, and personality surrounding them and out of which they grew.”); see also Clayton & Gillman, supra note 43, at 2–3 (noting the differing emphases of judicial decisionmaking research models); Clayton, supra note 18, at 20–21 (describing the characteristics of old institutionalism).
II. THE ROLE OF LAW IN THE JUDICIAL DECISIONMAKING PROCESS

This section proposes an alternative to the existing conception of legal rules as neutral institutional mechanisms that filter justices’ preexisting ideological preferences. Instead, legal rules, while ostensibly ideologically neutral, can actually reflect sincerely held ideological preferences. In this respect, legal rules depend upon justices’ policy preferences and yield ideologically contingent holdings. This section then takes up the issue of strict scrutiny to evaluate the general theoretical proposition that ideology is endogenous to doctrinal standards. Existing attitudinally based theories hold that strict scrutiny is the most effective standard of review in constraining the expression of justices’ policy preferences. Finally, this section challenges these models by contending that the lack of any explicit standard for determining whether a governmental interest is compelling—and thus sufficient to satisfy the rigorous ends-based analysis under strict scrutiny—means that the strict scrutiny test can be crafted in ideologically contingent ways.

A. Building on the Existing Theories: A New Approach

While rich in their empirical findings, the attitudinal and new institutionalism theories each conceptualize judicial holdings as the central dependent variable in the analysis. The current debate focuses nearly exclusively on the extent to which holdings are explained by justices’ personal political preferences. The approach taken in this Note differs from the attitudinal model, which conceptualizes legal rules and doctrines as neutral institutional mechanisms that filter the justices’ ideological goals. In conceiving of legal rules in this way, the traditional scholarship ignores other important contexts in which political ideology may exert influence on the decisionmaking process. But even when guided strictly by legal doctrine, holdings may nonetheless reflect particular political understandings because legal doctrines and tests are themselves constructed within a politically contested process. The approach taken in this Note differs from new institutionalism as well. This new approach looks beyond new institutionalism’s focus on how the Court’s institutional features—including legal rules—affect
judicial attitudes. Instead, it focuses on how justices’ historically contingent political and social attitudes shape the development of legal rules in the first instance.

Political ideology may be expressed, and located, in the legal tests that the Supreme Court has developed to adjudicate complex questions of constitutional law. In particular, the compelling state interest prong of strict scrutiny analysis is vulnerable to ideological application because the Court has never articulated criteria for determining what governmental interests are compelling. Absent doctrinal constraints, justices have unfettered discretion to determine what interests are sufficient to satisfy rigorous judicial review.

To integrate key insights from each of these theoretical approaches discussed above requires reconceptualizing the role of political ideology in the decisionmaking process. From the attitudinal model, this Note retains the guiding theoretical assumption that justices possess preexisting ideological preferences, though this approach does not depend on the assumption that justices necessarily try to achieve those ideological ends in the decisionmaking process. Instead, this approach assumes that justices can and do sublimate their ideological preferences by strictly following the law, an assumption rooted in the new institutionalism understanding that institutional norms constrain judicial behavior. Indeed, strict adherence to legal rules does not necessarily provide evidence that holdings (or votes) are value free. Thus, from new institutionalism this approach borrows the guiding theoretical assumption that institutional norms, rules, and beliefs can structure the expression of political ideology in judicial decisionmaking.

Existing social and political understandings can become part of the institutional structure of the Court. Armed with the traditional attitudinal analysis, however, this kind of ideological influence remains invisible. The norm toward sublimating political policy preferences in favor of adherence to the law may appear to be value free, but when the legal test itself reflects particular political understandings, then even ostensibly neutral applications of the law will be, in practical effect, ideological. The doctrine of legitimation—the norm toward adhering strictly to the law and sublimating one’s personal political preferences in fealty to the law—is designed to enforce the separation of powers and to guard the legitimacy of the courts. This legitimation constitutes “an organizational

49 Put another way, even where justices are not behaving in an “activist” fashion, in the colloquial sense of the term, the results of a case may nonetheless privilege certain ideological understandings.

50 Lawrence Baum, Judges and Their Audiences: A Perspective on Judicial Behavior 50–157 (2006) (arguing that justices’ interest in approval from other actors and institutions affects their behavior); Clayton, supra note 18, at 21–24 (surveying behavioralist insights regarding how judicial preferences are shaped by institutional understandings of appropriate judicial behavior); Mark J. Richards & Herbert M. Kritzer, Jurisprudential Regimes in Supreme Court Decision Making, 96 Am. Pol. Sci. Rev. 305, 315–16 (2002) (finding that certain seminal precedents exert a significant constraining effect on future cases of the same kind, as compared to cases that preceded the new “jurisprudential regime” in that field).
imperative for courts, serving both as a source of judicial inertia and a requirement for justification of particular judicial decisions and practices.\textsuperscript{51}

Judicial decisions are political acts, not in the sense that justices (or judges) necessarily strategically pursue their own personal political preferences, but rather in the sense that law is itself a process by which political values are constructed and given voice. Armed with this theoretical orientation, scholars are better equipped to understand the ways in which legal tests privilege particular politically, socially, and historically contingent understandings in contested areas of constitutional law.

B. The Constraining Capability of Traditional Standards of Review

Justices devise binding legal rules by issuing holdings in cases. From the perspective of lower federal courts, these precedents are, in fact, binding; lower courts have no discretion to disregard the commands of the Supreme Court.\textsuperscript{52} But Supreme Court precedent is designed to be binding on the Court as well. As such, legal doctrine issued from the Court has the potential to severely restrict the range of possible alternatives available to justices on any given legal issue,\textsuperscript{53} and indeed, it is designed to achieve this end.\textsuperscript{54} Thus, the efficient management of lower courts and a desire to maintain institutional legitimacy provide strong incentives for justices to adhere to existing legal frameworks and precedent.\textsuperscript{55}

\textsuperscript{51} Clayton & Gillman, \textit{supra} note 43, at 4; see also Edward S. Corwin, \textit{The Twilight of the Supreme Court: A History of Our Constitutional Theory}, at xxi–xxvii (1934) (theorizing that judicial commitments to doctrinal norms are themselves sensitive to broader political contexts).

\textsuperscript{52} The reach of Supreme Court precedent extends far beyond lower federal courts, however. While Supreme Court precedent in interpreting state constitutions does not bind state courts, they will nonetheless often use tests and legal doctrines developed by the Supreme Court to do so. See, e.g., Goodridge v. Dep’t of Pub. Health, 798 N.E.2d 941, 959–70 (Mass. 2003) (using Supreme Court jurisprudence addressing discrimination against homosexual persons to strike down state prohibition of same-sex marriage under the state constitution).

\textsuperscript{53} See, e.g., David M. O’Brien, \textit{Constitutional Law and Politics: Civil Rights and Civil Liberties} 125 (6th ed. 2005) (describing stare decisis as a “judicial policy that promotes the certainty, uniformity, and stability of the law” (internal quotation marks omitted)); see also Planned Parenthood of Se. Pa. v. Casey, 505 U.S. 833, 854 (1992) (describing how the decision to overrule a precedent “is customarily informed by a series of prudential and pragmatic considerations designed to test the consistency of overruling a prior decision with the ideal of the rule of law, and to gauge the respective costs of reaffirming and overruling a prior case”).

\textsuperscript{54} See Lee Epstein & Jack Knight, \textit{The Choices Justices Make} 11–12, 46–49, 157–79 (1998) (finding that the stability of Supreme Court jurisprudence promotes the legitimacy of the Court in the eyes of the public and Congress); Richards & Kritzer, \textit{supra} note 50, at 315–16.

\textsuperscript{55} Bartels, \textit{supra} note 14, at 475.
The most common adjudicative framework for assessing the constitutionality of governmental action is the tiers-of-scrutiny analytic set out in footnote four of *United States v. Carolene Products Co.* The tiers-of-scrutiny analytic prescribes different levels of judicial scrutiny for different kinds of governmental action, ranging from highly deferential to highly searching review. Under rational basis review—the most deferential standard—a governmental actor must demonstrate only that its action is rationally related to some legitimate governmental objective. This standard of review is the most common form, and it is used in cases where no fundamental right is at stake or where there is no suspect (or quasi-suspect) classification. Under rational basis review, the governmental action is presumed constitutional until the challenger can demonstrate otherwise.

Heightened scrutiny, on the other hand, is rights protective and prescribes a more searching form of judicial review. Strict scrutiny is the most rigorous form of review, and it is invoked where the governmental actor burdens a fundamental right or employs a suspect classification, notably race. Under strict scrutiny, the governmental actor is presumed to possess a legitimate objective, and the means selected are presumed to be rationally related to the ends.

---

56 See, e.g., United States v. Carolene Prods. Co. 304 U.S. 144, 152–54 & n.4 (1938) ("There may be narrower scope for operation of the presumption of constitutionality when legislation appears on its face to be within a specific prohibition of the Constitution. . . . It is unnecessary to consider now whether legislation which restricts those political processes which can ordinarily be expected to bring about repeal of undesirable legislation, is to be subjected to more exacting judicial scrutiny under the general prohibitions of the Fourteenth Amendment . . . . Nor need we enquire . . . whether prejudice against discrete and insular minorities may be a special condition, which tends seriously to curtail the operation of those political processes ordinarily to be relied upon to protect minorities, and which may call for a correspondingly more searching judicial inquiry.").

57 See ERWIN CHEMERINSKY, CONSTITUTIONAL LAW: PRINCIPLES AND POLICIES 668–77 (3d ed. 2006); see also Roe v. Wade, 410 U.S. 113, 153–67 (1973) (explaining that abortion laws—which entwine rights of privacy—receive strict scrutiny review and can only be upheld based on compelling state interests).


59 See CHEMERINSKY, supra note 57, at 677–89. Each prong of the test is presumed to be satisfied at the outset of the analysis. The governmental actor is presumed to possess a legitimate objective, and the means selected are presumed to be rationally related to the ends.

60 There is considerable debate on the Court regarding what characteristics a right must possess to be considered fundamental, and thus subject to strict scrutiny review. Compare Washington v. Glucksberg, 521 U.S. 702, 720–21 (1997) ("Our established method of substantive-due-process analysis has two primary features: First, we have regularly observed that the Due Process Clause specially protects those fundamental rights and liberties which are, objectively, ‘deeply rooted in this Nation’s history and tradition,’ . . . and ‘implicit in the concept of ordered liberty,’ such that ‘neither liberty nor justice would exist if they were sacrificed.’ Second, we have required in substantive-due-process cases a ‘careful description’ of the asserted fundamental liberty interest." (citations omitted)), with Planned Parenthood of Se. Pa. v. Casey, 505 U.S. 833, 847–48 (1992) ("It is also tempting . . . to suppose that the Due Process Clause protects only those practices, defined at the most specific level, that were protected against government interference by
scrutiny, the governmental action is presumed unconstitutional, a burden that can be overcome only when the government can demonstrate that it seeks to achieve a compelling state interest by means narrowly tailored to accomplish those ends.\(^63\) The final level, known as intermediate scrutiny, is employed when the governmental actor employs a quasi-suspect classification.\(^64\) To satisfy intermediate scrutiny, the classification must serve important interests by means substantially related to those objectives.\(^65\)

On the understanding adopted in *Carolene* and extended since, courts should generally presume that a law is constitutional\(^66\) unless it burdens a fundamental right, undermines the ability of the political process to repeal legislation, or discriminates against a discrete and insular minority.

The legal rules regarding the closeness-of-fit required between the ends the government seeks to achieve and the means it selects to achieve them should create different degrees of judicial discretion in deciding cases on their political policy preferences. And indeed, scholars have posited competing models theorizing the degree to which the different tiers of scrutiny constrain the expression of justices’ ideological preferences.\(^67\)

___

other rules of law when the Fourteenth Amendment was ratified. But such a view would be inconsistent with American law. It is a promise of the Constitution that there is a realm of personal liberty which the government may not enter. Neither the Bill of Rights nor the specific practices of States at the time of the adoption of the Fourteenth Amendment marks the outer limits of the substantive sphere of liberty which the Fourteenth Amendment protects.” (citations omitted)).

\(^{62}\) CHEMERINSKY, *supra* note 57, at 671; see also *City of Richmond v. J.A. Croson Co.*, 488 U.S. 469, 493–94 (1989) (holding that all race-based classifications are subject to strict scrutiny, regardless of whether the classification is invidious or benign); *Adarand Constructors, Inc. v. Pena*, 515 U.S. 200, 237 (1995) (extending *Croson’s* analysis to federal governmental action, stating that the Court wished “to dispel the notion that strict scrutiny is ‘strict in theory, but fatal in fact’”). Strict scrutiny analysis is not limited to racial classifications, however. See *Graham v. Richardson*, 403 U.S. 365, 371–72 (1971) (extending strict scrutiny analysis to classifications involving legal aliens).

\(^{63}\) See CHEMERINSKY, *supra* note 57, at 671.

\(^{64}\) Sex-based classifications are the classic quasi-suspect classification. *Id.* at 752–55; see, e.g., *United States v. Virginia*, 518 U.S. 515, 524 (1996) (requiring an “exceedingly persuasive justification” to satisfy intermediate scrutiny analysis); *Craig v. Boren*, 429 U.S. 190, 197 (1976) (holding that sex-based classifications are subject to intermediate scrutiny, regardless of whether the classification imposes burdens on men or women). The Supreme Court does not limit intermediate scrutiny to sex-based classification, however. See, e.g., *Clark v. Jeter*, 486 U.S. 456, 461–64 (1988) (extending intermediate scrutiny to classifications involving illegitimacy).

\(^{65}\) *Craig*, 429 U.S. at 197.

\(^{66}\) See CHEMERINSKY, *supra* note 57, at 671.

\(^{67}\) It is important to note that each of these models retains key assumptions from both the attitudinal and new institutional theories. They each presume that justices have political preferences that they seek to express through their votes in cases (the attitudinal model) and also that the law can constrain the expression of those preferences (new institutionalism). Professor Bartels’s empirical analysis of each of these theoretical models finds no support
The legal presumptions model posits “that constraint is a function of how strongly a legal rule presumes a certain case outcome.” Thus, under rational basis and strict scrutiny, justices have little discretion in effecting their preferred ideological outcomes because each strongly presumes a particular outcome: rational basis presumes the law is constitutional, while strict scrutiny presumes the law is unconstitutional. With respect to intermediate scrutiny, however, where only a comparatively modest presumption against constitutionality exists, justices have significantly more ideological discretion to decide cases on the basis of those preferences. Despite the fact that presumptions do not guarantee outcomes, they do nonetheless “restrict the range of legally justifiable conclusions for ruling against the presumption.” Thus, put simply, the impact of ideology on judicial decisionmaking should be higher in the intermediate scrutiny setting than it will be in rational basis and strict scrutiny settings because legal rules that presume the constitutionality or unconstitutionality of governmental action will constrain the expression of justices’ ideological preferences in their judicial decisionmaking.

In contrast to the legal presumptions model, the rights-protectiveness model contends that the degree to which the standard of review protects rights and liberties largely determines the outcome. Under this view, “legal rules that prescribe greater constitutional scrutiny of a law will exhibit more constraint on ideological discretion compared to rules that prescribe lower scrutiny.” This model, therefore, prescribes a linear relationship between the standard of review and the degree of ideological discretion available to the justices; as such, as the level of scrutiny increases, from rational basis to strict scrutiny, the level of ideological discretion decreases. Under strict scrutiny, the degree of discretion is low because it is highly protective of individual rights. Rational basis review, however, does not particularly constrain justices because the standard does not

---

68 See id. at 476.
69 Of course, the presumptions only create a likelihood that a justice will follow where they direct. There are many instances where the Court has declared a law unconstitutional under rational basis analysis and constitutional under strict scrutiny analysis. See, e.g., Grutter v. Bollinger, 539 U.S. 306, 334–35 (2003) (upholding affirmative action policy under strict scrutiny analysis); Romer v. Evans, 517 U.S. 620, 634–35 (1996) (striking down state constitutional amendment under rational basis review); see also Adam Winkler, Fatal in Theory and Strict in Fact: An Empirical Analysis of Strict Scrutiny in the Federal Courts, 59 VAND. L. REV. 793, 795 (“[T]his study shows that strict scrutiny is far from the inevitably deadly test imagined by the . . . myth [that strict scrutiny is strict in theory but fatal in fact] and more closely resembles the context-sensitive tool described by O’Connor [in Grutter].”).
70 Bartels, supra note 14, at 477.
71 Id. at 477.
72 Id.
73 Id.
protect rights strongly. Intermediate scrutiny cuts down the middle; it is both less constraining than strict scrutiny but more constraining than rational basis review.\textsuperscript{74}

While these models posit different levels of ideological discretion depending on the standard of review governing the case, a notable feature of both is that they theorize a low level of ideological discretion for justices when operating within the strict scrutiny setting. For the legal presumptions model, strict scrutiny affords low discretion because the unconstitutionality of the governmental action is so strongly presumed. Under the rights-protectiveness model, strict scrutiny constrains the expression of justices’ ideological preferences because it is highly protective of individual rights. Thus, where a justice disagrees with a particular outcome in a fundamental rights case, she is constrained by the rigor of the test; that is, the more rigorous the test, the lower the level of ideological discretion. In his analysis of these theories as applied to the free speech context, political science professor Brandon L. Bartels found that neither the legal presumptions model nor the rights-protectiveness model alone captures the empirical picture particularly well; instead, the “evidence supports the notion that only when constitutional scrutiny is maximized (i.e., strict scrutiny), will ideological discretion be constrained. Anything less than strict scrutiny will not constrain the magnitude of ideological voting.”\textsuperscript{75}

But the capacity of strict scrutiny to constrain justices’ ideological predispositions evidenced by these two competing models assumes that the tests are stable, transparent, and ideologically neutral. But there is good reason to believe that this assumption does not obtain. The next section turns to compelling state interest analysis to suggest that ideological preferences embedded within strict scrutiny analysis can structure the outcomes of cases in ways consistent with those ideological preferences, even where the Court adheres closely to the test.\textsuperscript{76} Because there is no standard for what constitutes a compelling state interest under the strict scrutiny standard of review, what appears under extant theoretical models to be the most ideology-constraining constitutional test may actually provide fertile ground for ideological decisionmaking and outcomes.

\textit{C. Compelling State Interest Analysis}

The existing theories regarding the extent to which legal doctrine constrains the expression of justices’ ideological predispositions in their decisionmaking, if at all, are incomplete. In particular, focusing on strict scrutiny is helpful because

\textsuperscript{74} Id.  
\textsuperscript{75} Id. at 490.  
\textsuperscript{76} This point distinguishes the attitudinal assumptions of the leading theories from the one advanced here. Where existing theories interpret strict adherence to a legal doctrine or test as evidence of the absence of ideological judicial decisionmaking, this conclusion only makes sense by assuming that the legal tests are, themselves, ideologically neutral. If, however, scholars locate the site of ideological contestation within the test itself, what appears to be an ideologically neutral outcome may actually privilege particular historically contingent ideological social and political understandings.
existing attitudinalist models theorize that this test is both ideologically neutral and highly constraining. This section challenges these models by arguing that the lack of any explicit, cognizable standard for determining whether a governmental interest is compelling means that the strict scrutiny test can be crafted in ideologically contingent ways. In this respect, strict scrutiny does not constrain the expression of political ideology in decisionmaking; instead, ideology is endogenous to the strict scrutiny test.

Equal protection analysis is governed by a tiered system of scrutiny designed to balance the government’s regulatory goals against individuals’ constitutional interest in equal treatment. Racial classifications are subject to strict scrutiny—a most rigorous form of judicial review, where the governmental action is presumed unconstitutional and survives only if narrowly tailored to achieve a compelling state interest.

Under the existing theoretical approaches, it would be reasonable to conclude that strict scrutiny imposes meaningful constraints on the expression of justices’ ideological preferences because of how strongly the rule presumes unconstitutionality. If justices disagree politically with the outcome presumed under the standard, for instance, they have little discretion to decide otherwise; that is, strict scrutiny significantly curtails the range of possible legitimate legal outcomes. When the standard itself is the object of analysis, however, a more complete picture of the role of ideology in the decisionmaking process emerges.

The Court has not specified a formal test, or even any informal set of criteria, for determining whether a governmental interest constitutes a compelling state interest for purposes of strict scrutiny analysis. Some opinions have identified governmental interests that are not compelling, but without providing any justification for that status. Others have found governmental interests sufficient to override individual rights, but without justifying their classification as compelling. Lacking any consistent or observable criteria for assessing the

77 See, e.g., Korematsu v. United States, 323 U.S. 214 (1944). Korematsu was the first equal protection case to apply strict scrutiny to racial classifications. The Court stated, “[A]ll legal restrictions which curtail the civil rights of a single racial group are immediately suspect. That is not to say that all such restrictions are unconstitutional. It is to say that courts must subject them to the most rigid scrutiny.” Id. at 216.

78 See, e.g., City of Richmond v. J.A. Croson Co., 488 U.S. 469, 505–06 (1989) (holding that strict scrutiny applies to all racial classifications, regardless of whether they are invidious or remedial).


81 See Buck v. Bell, 274 U.S. 200, 207 (1927) (upholding state statute permitting forced sterilization of incompetent women). While the Supreme Court has never expressly overruled Buck, forced sterilization statutes have been repeatedly challenged on due process and equal protection grounds in many jurisdictions, with varying levels of success.
strength of the governmental interest, the Court has complete discretion in this area. Justices, therefore, have opportunities to express their ideological preferences by crafting the compelling state interest prong of strict scrutiny analysis in ways that privilege their policy preferences and political understandings. And even where justices do not purposively craft compelling state interests with their ideological preferences in mind, which interests are found compelling and which are not is likely to reflect historically contingent social, political, and legal understandings.

The problem with unconstrained compelling state interest analysis is that it implicates federalism and separation of powers concerns, and it risks resurrecting the approach to constitutional analysis characterized by the *Lochner* era, where the Court eschewed principled legal analysis in favor of its value judgments regarding the wisdom and propriety of legislative purposes. The process described here may be even more insidious and enduring because, unlike the explicitly ideological decisions of the *Lochner* era, the ideological work being done in this context is obscured by the legitimacy of the long-established strict scrutiny test. Indeed, just because the Court has rejected the jurisprudence of the *Lochner* era does not mean that it has rejected its methodology. Perhaps in response to *Lochner*, it is now axiomatic in constitutional law that courts exercising judicial review over the elected branches are better equipped to cast judgment on the means that legislatures select to advance their goals, rather than the ends they seek to achieve.

What Professor Bickel called the “counter majoritarian difficulty” is implicated in extant compelling state interest analysis: when unelected actors substitute their political judgments for those of elected officials, the legitimacy of the judicial branch is undermined and the authority of the political branches is unduly harmed. Strict adherence to the law in this context, therefore, should not necessarily assuage Professor Bickel’s concerns; indeed, it may be even more problematic than explicit “judicial activism” because the process is significantly less transparent and occurs under the guise of what appears to be an ideologically neutral constraint.

See generally Jeffrey F. Ghent, Annotation, *Validity of Statutes Authorizing Asexualization or Sterilization of Criminals or Mental Defectives*, 53 A.L.R.3d 960 (1973).


83 See Chemerinsky, supra note 57, at 620–21.


87 Id.
III. COMPELLING STATE INTEREST ANALYSIS IN THE AFFIRMATIVE ACTION CONTEXT

To this point, this Note has suggested that existing explanations of the role of ideology in judicial decisionmaking—and thus the constraining capacity of law—are incomplete because they focus on whether judicial holdings correspond to the ideological predispositions that justices bring to the table. Further, this Note has argued for reorienting the analytical lens to the ways in which ideology can be used to create legal tests. In particular, because of the lack of any explicit standard governing whether an asserted state interest is compelling for purposes of strict scrutiny analysis, justices may craft strict scrutiny’s ends analysis in such a way as to privilege particular ideological understandings in their jurisprudence.

To explore this suggested reorientation in a concrete way, this section now turns to race-based affirmative action in higher education as an illustrative example. Affirmative action jurisprudence is an ideal candidate for refocusing the analytical lens to the ways that compelling state interest analysis is subject to contested ideological views. Its jurisprudential history demonstrates how the justices have used their discretion to credit just one of many possible, complementary state interests as compelling. Thus, this section demonstrates how the Court has used this discretion and chosen diversity in higher education, to the exclusion of all other asserted interests, as the sole compelling governmental interest in this field. Part IV proceeds to discuss what ideological predispositions about the nature of institutional racism are reflected in this choice.

A. The Diversity Rationale

The rationale that diversity is a compelling state interest is a relatively recent innovation; and in this respect, it is a useful example of the kind of compelling state interest divined out of whole cloth. Although Korematsu v. United States settled the question that strict scrutiny would apply to governmental action that classified on the basis of race, it was not until the 1950s that the Court began striking down racially discriminatory laws, and not until the 1960s that it used the

---

88 The rationale did not appear in any remedial context before the Bakke litigation. Indeed, in an earlier affirmative action case, DeFunis v. Odegaard, 416 U.S. 312, 318 (1974), which was ultimately resolved on mootness grounds and not on the merits, the University of Washington law school did not advance a diversity rationale in defense of its affirmative action policy. “[T]he word ‘diversity’ does not appear at all in the record.” Joshua P. Thompson & Damien M. Schiff, Divisive Diversity at the University of Texas: An Opportunity for the Supreme Court to Overturn Its Flawed Decision in Grutter, 15 TEX. REV. L. & POL. 437, 443 (quoting Peter W. Wood, Diversity: The Invention of a Concept 111 (2003)).
89 323 U.S. 214 (1944).
strict scrutiny analytic toward that end.91 It was only by the late 1970s, in the affirmative action setting, that diversity became a compelling state interest for purposes of strict scrutiny analysis. Up to this point, while the Court was moving in the direction of permitting the use of racial classifications in remedial contexts, the only sufficient interest justifying the use of racial classifications was where the governmental actor was seeking to remedy racial discrimination in the particular institution at issue in the case. During the 1960s and 1970s, this was particularly common in the school desegregation context.92 By the 1970s, after the implementation of the Civil Rights Act of 1964,93 the Court began to face novel issues related to race under the equal protection clause. While instances of invidious racial discrimination remained central,94 the Court granted certiorari in cases where local legislative bodies employed benign racial classifications unrelated to past discrimination within the institution utilizing the racial classification. The Court began to address the question of race-based affirmative action in this context.

The diversity rationale has been the subject of considerable legal controversy since the constitutionality of educational affirmative action was first addressed in

---

91 See Loving v. Virginia, 388 U.S. 1 (1967). In Loving, the Court struck down racially restrictive marriage statutes under strict scrutiny analysis. It stated,

Because we reject the notion that the mere “equal application” of a statute containing racial classifications is enough to remove the classifications from the Fourteenth Amendment’s proscription of all invidious racial discriminations, we do not accept the State’s contention that these statutes should be upheld if there is any possible basis for concluding that they serve a rational purpose.

Id. at 8.

92 See Swan v. Charlotte-Mecklenburg Bd. of Ed., 402 U.S. 1, 23 (1971) (“Our objective in dealing with the issues presented by [the desegregation] cases is to see that school authorities exclude no pupil of a racial minority from any school, directly or indirectly, on account of race; it does not and cannot embrace all the problems of racial prejudice, even when those problems contribute to disproportionate racial concentrations in some schools.”).


94 In these years, the Court broadened the scope of its analysis to permit Congress to remedy invidious racial discrimination under the commerce power. The Civil Rights Act, for example, presented important questions about the nature of racial discrimination and the scope of congressional authority under the Commerce Clause. See Heart of Atlanta Motel, Inc. v. United States, 379 U.S. 241, 257–58 (1964) (upholding the Civil Rights Act under the Commerce Clause); Katzenbach v. McClung, 379 U.S. 294, 304 (1964) (extending the reasoning in Heart of Atlanta Motel to regulate racial discrimination in public accommodations under the Civil Rights Act of local, intrastate businesses that have a substantial effect on interstate commerce).
Regents of the University of California v. Bakke.⁹⁵ In its plurality opinion, the Court struck down Davis medical school’s minority set-aside admissions policy on grounds that the set-aside program was not sufficiently narrowly tailored to achieve the University’s interest in diversity.⁹⁶ In doing so, the Court rejected various rationales advanced for the classification, including remedying societal discrimination ⁹⁷ and improving the delivery of health care to underserved communities.⁹⁸ It then accepted the argument that creating a diverse student body is “clearly”⁹⁹ a compelling state interest without advancing a principled reason for this conclusion. The Court justified the diversity rationale on grounds that “[a]n otherwise qualified . . . student with a particular background . . . may bring . . . experiences, outlooks, and ideas that enrich the training of its student body and better equip its graduates to render with understanding their vital service to humanity.”¹⁰⁰ Put another way, the pedagogical benefits conferred on other, presumably white, students constitutes a compelling state interest, whereas promoting the interests of historically disadvantaged communities falls short.

The diversity argument, now embraced in the affirmative action context,¹⁰¹ was a novel approach to justifying racial classifications, one not raised by any litigant prior to the Bakke litigation. But because diversity in higher education was embraced only by Justice Powell’s plurality opinion in Bakke,¹⁰² lower courts retained discretion to reject it as a compelling state interest in the affirmative action setting, and some did precisely that.

In Hopwood v. Texas,¹⁰³ the Fifth Circuit Court of Appeals struck down the race-based affirmative action policy at the University of Texas.¹⁰⁴ In doing so, the

---

⁹⁵ 438 U.S. 265 (1978). Indeed, not until Bakke did the Court address the question of whether a governmental actor could use racial classifications in contexts outside of the institutional-remedial context. Thompson & Schiff, supra note 88, at 443.

⁹⁶ See Bakke, 438 U.S. at 319–20 (stating that the State’s substantial interest may be served by a properly devised admissions program involving the competitive consideration of race and ethnic origin).

⁹⁷ Id. at 307.

⁹⁸ Id. at 310.

⁹⁹ Id. at 311.

¹⁰⁰ Id. at 314.

¹⁰¹ The diversity rationale may be sufficient to satisfy the ends analysis in the strict scrutiny context of race-based classifications only in the higher education setting. But see Metro Broad., Inc. v. FCC, 497 U.S. 547, 567–68 (1990) (applying intermediate scrutiny to remedial racial classifications employed by the federal government and finding that racial diversity in media broadcasting constitutes an important state interest, without addressing the question of whether diversity in that context constitutes a compelling state interest for strict scrutiny purposes), overruled on other grounds by Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 227 (1995) (overruling Metro Broadcasting by requiring strict scrutiny analysis where the national government employs remedial, race-based classifications).

¹⁰² Only Justice Powell embraced the diversity rationale. See Bakke, 438 U.S. at 311–12.

court rejected Justice Powell’s plurality opinion in Bakke, and thus rejected the diversity rationale. In holding that the only sufficient state interest in the remedial racial classification context was remedying past institutional discrimination, the court imported Justice O’Connor’s reasoning from Metro Broadcasting Inc. v. FCC into the affirmative action context, concluding that “[t]he interest in increasing the diversity of broadcast viewpoints is clearly not a compelling interest. It is simply too amorphous, too insubstantial, and too unrelated to any legitimate basis for employing racial classifications.” Indeed, according to the Hopwood court, not only is diversity in higher education not a compelling state interest, it actually undermines the goals of the Fourteenth Amendment.

The diversity rationale remained an open question until the Court held in Grutter v. Bollinger that “[t]he [University of Michigan] Law School’s educational judgment that . . . diversity is essential to its educational mission is one to which we defer.” The Court recognized diversity as a compelling state interest, but as in Bakke, it did not articulate a principled rationale for its holding, but rather a contextual one: “We have repeatedly acknowledged the overriding importance of preparing students for work and citizenship, describing education as . . . [playing] a fundamental role in maintaining the fabric of society.” The Court emphasized that including race as one of several positive factors in assessing applications was a permissible, narrowly tailored, means to achieve the diversity interest, as distinguished from the quota system it struck down in Bakke and in Gratz.

Grutter stands for the principle that the diversity rationale constitutes a compelling state interest for purposes of strict scrutiny analysis in the educational affirmative action field. Fisher has endorsed this view. But what is striking is how few governmental interests the Court has recognized as compelling in the

---

104 Id. at 934.
105 Id. at 945 (“Modern equal protection has recognized only one [compelling state] interest: remedying the effects of racial discrimination [within the institution].”).
107 Hopwood, 78 F.3d at 945.
108 Id. (“Within the general principles of the Fourteenth Amendment, the use of race in admissions for diversity in higher education contradicts, rather than furthers, the aims of equal protection. Diversity fosters, rather than minimizes, the use of race. It treats minorities as a group, rather than as individuals. It may further remedial purposes but, just as likely, may promote improper racial stereotypes, thus fueling racial hostility.”).
110 Id. at 328.
111 Id. at 331.
112 Gratz v. Bollinger, 539 U.S. 244, 270–73 (2003). The University of Michigan’s undergraduate affirmative action policy did not involve an explicit set-aside program, as did the University of California’s in Bakke, but the Court reasoned that Michigan’s plan awarding an automatic 20% of the total points necessary for guaranteed admission to the University nonetheless was not narrowly tailored to achieve its compelling interest in racial diversity at the school. Id. at 270.
113 Fisher, 133 S. Ct. at 2419.
affirmative action field. The Court’s current jurisprudence permits an unduly limited scope of state action. Promoting the interests of underserved communities, remedying societal discrimination, or extending the benefits of an elite education to historically disadvantaged groups would likely promote the ends of the Equal Protection Clause. This is so even under the Court’s truncated state action doctrine, so long as the legislature reasonably attributes some of the harm from those problems to the historical policies and practices of the targeted institution, or if the legislature could reasonably conclude that those institutions could ameliorate past or present discrimination caused by other state action. But, as of now, the Court credits none of these interests as sufficiently compelling to satisfy strict scrutiny analysis. Part IV further explores how the Court’s narrow definition of compelling state interest in this context privileges particular ideological understandings about the nature of institutional racism and silences others.

IV. THE IDEOLOGY OF DIVERSITY

What about the Court’s decision to credit diversity as the only compelling state interest in the affirmative action setting makes it ideological? To demonstrate the utility of this Note’s theoretical contribution—that, in the abstract, the Court’s ability to choose which interests it will credit as compelling means that it can build into strict scrutiny its own ideological predispositions—it is not enough to simply identify how the Court has used its discretion to select one of many possible state interests as compelling. It is also important to show how this choice reflects particular ideological predispositions. This section demonstrates how in conceptualizing diversity as the only compelling state interest in the educational affirmative action setting, the Court has privileged particular, historically contingent, understandings of the way race operates in American life. In particular, the Court’s choice reflects the view that affirmative action has innocent victims. This view, in turn, reflects underlying ideological assumptions about the nature of institutional racism: that it is an individual, rather than a systemic, problem.

Professor Thomas Ross long ago identified that the “rhetoric of innocence” is deeply embedded in the public (and legal) debate over affirmative action. 114 The rhetoric is often explicit, as in Justice Powell’s opinion in Bakke:

The state certainly has a legitimate and substantial interest in ameliorating, or eliminating where feasible, the disabling effects of identified discrimination. . . . We have never approved a classification that aids persons perceived as members of relatively victimized groups at the expense of other innocent individuals.115

The rhetoric of innocence has been a consistent theme in a variety of affirmative action decisions in the years following Bakke. In Fullilove v. Klutznik,\footnote{448 U.S. 448 (1980), abrogated by Adarand Constructors, Inc. v. Pena, 515 U.S. 200 (1995) (requiring strict scrutiny analysis when the federal government employs remedial racial classifications, rather than the intermediate standard of review announced in Fullilove).} for instance, the Court upheld a federal statute requiring that 10% of contracts in federally funded public works projects be awarded to minority-owned businesses. The Court wrote, “When effectuating a limited and properly tailored remedy to cure the effects of prior discrimination, such a ‘sharing of the burden’ by innocent parties is not impermissible.”\footnote{Id. at 484 (emphasis added).} And in Wygant v. Jackson Board of Education,\footnote{476 U.S. 267 (1986).} the Court faced the question of whether a union contract giving Black union members greater protection from layoffs than it did more senior white members offended the Equal Protection Clause. In striking the contract down, the Court opined, “No one doubts that there has been serious racial discrimination in this country. But as the basis for imposing discriminatory legal remedies that work against innocent people, societal discrimination is insufficient and over-expansive.”\footnote{Id. at 276 (emphasis added).}

This discourse is built upon powerful assumptions about the nature of racial discrimination and just remediation. Whites are presumptively perceived as having been displaced by the beneficiaries of race-based affirmative action. This sense of innocence depends on assumptions about the individual, rather than the systemic nature of institutional and societal racism.\footnote{See Ross, supra note 114, at 300–01.} As such, a governmental interest in remedying societal discrimination has never been held to be a compelling state interest. This conceptual assumption about the individual nature of racism undergirds the moral position generating “reverse racism” rhetoric: the displaced white candidate is the victim of reverse racism because she is innocent of any particular, identifiable act of racism, but she is nonetheless being displaced from her rightful place on account of her race. This individualized assessment runs in the other direction as well. Beneficiaries are perceived as undeserving on the presumption that they have not suffered directly from racism; that is, they are not “actual victims.”\footnote{Id. at 301.}

The rhetoric of innocence is further supported by particularized, historically contingent assumptions about merit and deservedness.\footnote{Barbara J. Flagg, Diversity Discourses, 78 Tul. L. Rev. 827, 830 (2004) (critiquing the discourse of affirmative action as a form of racial reparation).} From the perspective of the rhetoric of innocence, in addition to being unfair because affirmative action privileges undeserving beneficiaries (people of color) and punishes innocent victims (white people), affirmative action also eschews cherished norms of merit, without acknowledging that current “objective” measures of merit, thought to be
neutral and to be inherited from time immemorial, are in fact deeply flawed and historically contingent. The *Hopwood* court employed this assumption in concluding that diversity created through affirmative action actually undermines the purposes of the Fourteenth Amendment: where racism is conceptualized as individualistic—a harm perpetrated by one person on another—and merit as neutral and value free, then it is the policy, rather than latent racism, that “promote[s] improper stereotypes, thus fueling racial hostility.”

In *Grutter*, the Court is explicit about the benefits that flow from educational diversity. In Justice O’Connor’s opinion, the Court discusses how having a “critical mass” of minority students confers identifiable benefits for all students by promoting “cross-racial understanding, help[ing] to break down racial stereotypes, and enable[ing] . . . [a] better understand[ing] [of] persons of different races.” The key distinguishing factor linking each of these interests is that they serve the interests of the institution generally by being justified in relation to the benefits they confer on all students, regardless of race. Put another way, diversity alone constitutes a compelling state interest to the exclusion of all other interests originally proposed in *Bakke* because it alone also confers benefits on the nonrecipients of affirmative action.

*Grutter*’s recognition that diversity in higher education constitutes a compelling state interest in the affirmative action setting does not itself raise any particular concerns. But when it is the sole interest that the Court has acknowledged, then the field begins to reflect certain understandings about the nature of race and racism that would not, per se, appear ideologically specific when viewed from the perspective of the attitudinal model and other traditional approaches to the study of ideology in judicial decisionmaking. To the extent that

---

123 *Id.; see also* Charles R. Lawrence III, *Two Views of the River: A Critique of the Liberal Defense of Affirmative Action*, 101 COLUM. L. REV. 928, 942–58 (2001) (critiquing the liberal defense of affirmative action for perpetuating the legitimacy of traditional, numerical indicators of merit); Daria Roithmayr, *Deconstructing the Distinction Between Bias and Merit*, 85 CALIF. L. REV. 1449, 1469–81 (1997) (arguing that traditional, numerical standards of merit disproportionately exclude white women and people of color because merit standards were developed by privileged social groups).

124 *Hopwood v. Texas*, 78 F.3d 932, 945 (5th Cir. 1996) (rejecting the diversity rationale articulated in Justice Powell’s plurality opinion in *Bakke*), *overruled by* *Grutter v. Bollinger*, 539 U.S. 306 (2003); *see also* Metro Broad., Inc. v. FCC, 497 U.S. 547, 603 (1990) (“The dangers of [racial] classifications are clear. They endorse race-based reasoning and the conception of a Nation divided into racial blocs, thus contributing to an escalation of racial hostility and conflict.”), *overruled on other grounds*, Adarand Constructors, Inc. v. Pena, 515 U.S. 200 (1995); City of Richmond v. J.A. Croson, Co., 488 U.S. 469, 493 (1989) (“Classifications based on race carry a danger of stigmatic harm. Unless they are strictly reserved for remedial settings, they may in fact promote notions of racial inferiority and lead to a politics of racial hostility.”).


126 *Id.* at 330.

127 Academic supporters of affirmative action have cited diversity as an important benefit of affirmative action policies. See Flagg, *supra* note 122, at 833–37.
“affirmative action is seen as to transfer goods from innocent white victims to undeserving beneficiaries—people of color—in order to remedy societal and systemic disadvantage and discrimination,” the diversity rationale reinforces the innocence discourse. Under the diversity-alone discourse, a benefit—the educational benefits that flow from racial diversity—conferred on white students is a necessary condition for permissible affirmative action in higher education. This approach reinforces historically, socially, and politically contingent understandings about the role of race and the nature of racism in American society.

CONCLUSION

In conclusion, this Note has applied general theoretical suggestions about how ideology plays a role in judicial decisionmaking to the concrete, doctrinal case of educational affirmative action. Part I suggested existing theories that grapple with the question of whether, and to what extent, the law constrains the expression of justices’ ideological preferences are valuable, but ultimately incomplete. They are incomplete because they focus nearly exclusively on judicial holdings as their dependent variable. In so doing, they assume that legal tests are objective and ideologically neutral, but such assumptions obscure how legal tests are themselves dependent on the Justices’ ideological preferences; that is, shifting the attention from holdings to the legal test permits scholars to relax the assumption that legal tests are neutral and objective.

Part II discussed how doctrinal tests can reflect the Justices’ sincerely held policy preferences. Reconceptualizing the legal test as the dependent variable allows scholars to observe how justices’ policy preferences inform their construction. In particular, attitudinalists hypothesize that strict scrutiny should highly constrain justices’ ideological preferences. This much is true if the focus is solely on judicial holdings. But because there is no explicit, cognizable set of criteria for determining whether an interest is compelling for purposes of the strict scrutiny analytic, justices have complete discretion to determine which interests are compelling and which fall short. Put another way, because of this discretion, the Justices’ ideological preferences can inform how the strict scrutiny test is crafted in any particular doctrinal context. The danger of political ideology becoming endogenous to strict scrutiny’s ends-based analysis is a real one. Therefore, even where an opinion appears to adhere rigidly to the requirements of strict scrutiny—and thus it appears as though the law has constrained the justices’ preferred ideological preferences—the result may nonetheless produce an ideologically contingent result because it was produced by an ideologically contingent test.

Part III moved from the theoretical to the concrete by offering the race-based educational affirmative action field as an illustrative example. This Part discussed the general theory that legal tests—in particular, strict scrutiny—are infused with the justices’ ideological preferences and assumptions: in this case, preferences and

128 Id. at 849.
assumptions about the nature of racial discrimination. Affirmative action provides a useful example of how the justices can use their discretion to credit certain interests as compelling, to the exclusion of all others.

Diversity is the only interest the Court has acknowledged as compelling, and the history of the field provides few explicit indicators of why the justices ultimately made this choice. In any event, the choice to credit only diversity is informed by ideological assumptions about the nature of racial discrimination; and in doing so, the Court has created an ideologically contingent test.

Finally, Part IV suggested some of the ways in which the choice to credit only diversity reflects assumptions about the nature of racial discrimination. In acknowledging only diversity as sufficient to satisfy strict scrutiny, the Court signals a discourse of innocent (white) victims and undeserving (racial minority) beneficiaries, a discourse itself premised on the view that racial discrimination is a harm perpetrated by one individual on another, rather than as a systemic, institutional problem. The utility of this theory is by no means limited to the affirmative action setting. But in this illustrative context, such a theoretical suggestion is useful in uncovering some of the Court’s hidden ideological assumptions. It primes scholars to see how the strict scrutiny test is built upon certain ideological assumptions about the nature of racial discrimination. In so doing, it suggests how even an apparently constraining doctrine such as strict scrutiny can nonetheless produce ideologically specific results.

129 See supra note 11.
I. INTRODUCTION

Cars that drive themselves may seem about as realistic as Marty McFly’s DeLorean in Back to the Future, but this is far from the case. Autonomous technology is available and in use in many new vehicle models. In fact, in 2010 Google announced that the company had successfully built cars capable of driving themselves (“autonomous cars”). Even more amazing, at its 2010 announcement, Google also revealed its driverless cars (customized Toyota Prius models) had logged over 100,000 miles on public roads. As of early 2013, Google’s several autonomous vehicles have traveled over 300,000 miles, and have yet to cause a traffic accident.

Autonomous vehicles are not merely a Google pet project; many automobile manufacturers have been implementing varying degrees of autonomous technology for years. More recently, luxury automakers Lexus and Audi both unveiled their own versions of fully autonomous cars. One of the most fascinating aspects of the growth in autonomous vehicle technology is the way that it could dramatically improve the safety of automobile travel. Google’s self-purported goal is to
eliminate all traffic fatalities. Regardless of whether Google ever reaches that milestone, it is clear that autonomous vehicles can improve the landscape of transportation safety and revolutionize how we travel.\(^8\)

The importance of autonomous vehicles, and largely the purpose for writing this Note, is that they have the potential to drastically reduce the numerous injuries and fatalities caused by automobile accidents.\(^9\) In 2010 there were over five million traffic accidents in the United States.\(^10\) Those accidents resulted in nearly 1.5 million injuries and over thirty-thousand fatalities.\(^11\) While traffic accidents can be caused by a variety of factors, human error is often the root cause. For example, of the thirty-thousand fatalities in 2010, about one-third were caused by an individual driving with a blood alcohol level of over 0.08%.\(^12\) Unfortunately, driving under the influence is only the tip of the iceberg in terms of human error in automobile accidents. It is rare to find an individual who has not engaged in at least some form of distracted driving, whether it is talking on a cell phone, texting while driving, or checking email. Even drowsy driving causes thousands of accidents each year.\(^13\) Autonomous vehicles have the potential to drastically reduce, and possibly eliminate altogether traffic accidents that are caused by human error.\(^14\)

The advantages of autonomous vehicles are not limited to safety. For example, autonomous vehicles are likely more efficient drivers than their human counterparts.\(^15\) Ideally, a standard autonomous vehicle will likely not accelerate out of one intersection and then slam on the brakes at the next. Accordingly, as the number of autonomous vehicles on the road increases, emissions from automobiles

---


\(^9\) See generally Valdes-Dapena, supra note 7 (noting the potential safety benefits of autonomous vehicles).


\(^11\) Id.


\(^14\) See generally Valdes-Dapena, supra note 7.

\(^15\) See NIDHI KALRA ET AL., RAND CORP., LIABILITY AND REGULATION OF AUTONOMOUS VEHICLE TECHNOLOGIES 1 (2009).
could drop dramatically.\textsuperscript{16} In addition to the benefits to the physical environment, autonomous vehicles could also have significant social impact. One of the most profound potential advantages of autonomous vehicles is their accessibility to disabled individuals for whom conventional driving is not possible. In particular, those with visual disabilities could experience unprecedented mobility and freedom with the advent of autonomous vehicles.

With the potential benefits of autonomous vehicles, it is no surprise that some state legislatures have been spurred into action by the rapid progression of the technology. For instance, on September 25, 2012, California followed Nevada’s and Florida’s leads and passed a law to allow autonomous cars on California’s public roadways.\textsuperscript{17}

California’s new law expressly authorizes autonomous vehicles to operate on California roads, and also directs the California Department of Motor Vehicles to draft regulations for the operation of autonomous vehicles.\textsuperscript{18} While California, Nevada, and Florida have taken the first important step towards addressing autonomous vehicles, all of the legislation is very basic, and none of it begins to address the potential liability issues with autonomous vehicles.\textsuperscript{19}

With the implementation of such groundbreaking technology, the automobile industry and transportation as we know it are likely to see significant changes in the coming decades.\textsuperscript{20} To address these changes, the law needs to be ready to adapt. Technology advances at a break-neck pace with new inventions and advances emerging constantly. The law, sometimes rightly so, does not progress at

\textsuperscript{16} Id.


such a fast clip. While law and technology may evolve to different degrees, they are interconnected in a variety of ways. Technology has significantly changed the substance and practice of the law. Advances in DNA matching and ballistics are just a handful of technological advances that have changed the law. Similarly, changes in the law can have significant impacts on the development of technology. If a law disfavors a particular product, either because it is unsafe or for some other reason, it can often cripple the producer of said product. In this manner, the legal landscape can drastically affect the way in which new technology influences society.21

The implementation of autonomous vehicles is a perfect example of the relationship between the law and innovative technology. The concept of cars that drive themselves raises a whole host of potential legal concerns. While there are countless criminal law questions that arise with such vehicles, this Note will focus on what may become the most significant legal hurdle for the implementation of autonomous vehicles: products liability. To determine liability in a traditional accident, conventional theories of liability primarily apportion liability based on manual error by the driver. Accordingly, automobile manufacturers are not held liable for the vast majority of the traffic accidents that occur.22 While manufacturer liability is still present in cases where a failure of the vehicle causes an accident or enhances the damage, the majority of the liability in traffic accidents remains with the driver. Traditional liability becomes exponentially more confusing and difficult to apply, however, when the “driver” of a vehicle is not a human, but rather a complex system of interconnected machinery. The question becomes: Who is liable when autonomous cars eventually crash?

The ultimate goal of this Note is to provide the best solution to this question. To most effectively answer the question of liability, this Note is divided into several distinct sections. Part II provides a basic description of how autonomous vehicles function so as to illustrate the complexity of the legal questions these vehicles pose. Next, an overview of products liability law is presented to detail the relevant components of the current state of the law. Part III provides an application of current products liability law to potential autonomous vehicle claims. Such an application will showcase how ill-prepared products liability law is and the potential consequences to both manufacturers and potential plaintiffs. To address the concerns with current products liability law, Part IV suggests a no-fault compensation system that can promote the interests of manufacturers and plaintiffs alike.

22 See NAT’L HIGHWAY TRAFFIC SAFETY ADMIN., supra note 12 (detailing traffic accident causes, notably those accidents that were caused by alcohol-impaired drivers).
II. BACKGROUND

Varying degrees of autonomous technology have been present in the automobile industry for decades.\(^{23}\) Depending on how one defines “autonomous,” traditional devices such as air bags and seatbelts can be qualified as autonomous technology. In this basic sense, items that do not require human engagement to act are the basic examples of autonomous technology.\(^{24}\) In relation to autonomous vehicles, however, the true definition of “autonomous” still varies. For the purposes of this Note, autonomous vehicle technology is defined according to a more accepted modern definition: “technologies and developments that enable a vehicle to assist, make decisions for, and, ultimately, replace a human driver.”\(^{25}\)

The following sections provide a foundation that is necessary to understand the unique legal concepts that autonomous vehicles raise. Section A provides a high-level summary of how autonomous vehicles function and also highlights the development of the technology through the example of Google’s autonomous vehicles. Section B goes on to provide an overview of current products liability law that would apply to autonomous vehicles—namely, manufacturing and design defect claims.

A. Autonomous Safety Technology and the Development of the Google Car

For driverless cars to function, multiple components of autonomous technology have to work in sync to direct the vehicle. “Autonomous vehicles,” however, is a term that is commonly used to refer to very different technologies. For example, two of the more common and currently used examples of autonomous vehicle technology are Adaptive Cruise Control (ACC) and Lane Departure Warning Systems (LDWS).\(^{26}\) Both automated features assist the human driver and reduce the potential for human error.\(^{27}\) ACC is similar to traditional cruise control, except that sensors and radar work in correlation to automatically adjust the speed of the vehicle based on the activity of the surrounding traffic without any driver input.\(^{28}\) Notably though, for ACC to begin acting it must be enabled by the driver and can likewise be disabled at the driver’s request.\(^{29}\) LDWS

---

\(^{23}\) See Daniel Bartz, Autonomous Cars Will Make Us Safer, WIRED (Nov. 16, 2009, 8:00 AM), http://www.wired.com/autopia/2009/11/autonomous-cars (discussing early predictions and hopes of autonomous technology as well as a General Motors group that attempted to develop automated highway technology in 1950).

\(^{24}\) See NIDHI KALRA ET AL., supra note 15, at 3.

\(^{25}\) Id. at 1.


\(^{27}\) Id.

\(^{28}\) Id.

\(^{29}\) Id.
use cameras or sensors that monitor the location of the vehicle with respect to other traffic and alert drivers if they begin drifting out of their lane. LDWS became a priority after a 1999 National Highway Traffic Safety Administration (NHTSA) study noted that 62,000 traffic accidents occurred as a result of drifting. If the system detects drifting or changing lanes without signaling, the system will alert the driver through audible and visual warnings.

While both ACC and LDWS are examples of autonomous vehicle technology, they do not spark the same legal concerns as fully autonomous vehicles because, while ACC and LDWS are autonomous technologies in the sense that they act automatically, they do not take full control of all vehicle functions (notably steering), and they require driver input. As such, they are more aptly characterized as “partial control systems." Partial control systems are certainly examples of autonomous technology, but because of the difference noted above, their presence in a car is insufficient, on its own, to consider the car an autonomous vehicle.

The diagram from the RAND study shown below clarifies the varying degrees of autonomous technology in vehicles. Starting on the left with conventional automobiles, as the diagram moves to the right, the automated systems begin to take more control of the vehicle as human input decreases.

For purposes of consistency, any autonomous vehicle mentioned in this Note refers to “autonomous vehicles with driver-in-the-loop,” as referenced in the table. Stated more simply, this means that the vehicle is navigating without human input, but still with a human in the driver seat who can take control if a failure in automation were to occur. The distinction between such an arrangement and a truly driverless car is that in a driverless car, as the name implies, there would be no

30 Id.
31 Id.
33 Carley, supra note 26.
34 Id.
36 Id. at 4.
human participant in the driver seat. While truly driverless cars are an eventual possibility, early legislation shows that the more likely scenario is that there will need to be a licensed human in the vehicle for the purpose of a safety backup.37

As noted in the Introduction, there are numerous automakers invested in the development of autonomous vehicles. Google has been at the forefront of the development in the United States, and its history of autonomous vehicle development has been more widely publicized. Thus, this Note will focus on autonomous vehicles as developed by Google. While the technology that powers autonomous vehicles varies among developers, Google’s model of autonomous vehicles provides enough of a general representation to address the legal issues presented by autonomous vehicles as a whole.

In traditional vehicles, the human driver handles virtually all of the interaction and response between the vehicle and the surrounding environment. The driver evaluates the surroundings (speed limit signs, lanes, weather, etc.), determines how the vehicle will perform (change lanes, speed up, slow down, etc.), and then ultimately decides how to execute that performance (accelerating through a turn, how far to turn the steering wheel, etc.)38 Further, the driver must also provide relevant information to the environment, such as waving another vehicle through an intersection or engaging the blinker to signal a turn.39 In a traditional automobile, the only autonomous response that the car exhibits occurs when the environment impacts the vehicle in some manner (e.g., airbags that deploy upon contact).40

For autonomous vehicles to function, different types of advanced technologies, working in concert with each other, perform all of the tasks detailed above.41 In the case of Google’s fleet of autonomous vehicles, the company took conventional Toyota Prius models and equipped them with a variety of advanced equipment that allows the car to execute the functions of a traditional driver.42 One of the essential components that the autonomous car uses is a rotating Light Detection and Ranging (LiDar) unit on the roof.43 The LiDar system scans two

---

39 Id.
40 Id.
41 Id.; Thrun, supra note 2.
hundred feet around the vehicle in each direction and creates a type of three-dimensional schematic of the car’s surroundings. To locate the vehicle within the environment, the left rear wheel is equipped with a position estimator. The LiDar and position estimator work in conjunction with a mounted video camera and radar sensors to “see” what is going on in the car’s environment. This allows the car to respond when it notes identifying features in the surrounding environment, such as traffic lights and pedestrians. The last component is a sophisticated computer located in the back of the vehicle. This device processes all of the real-time data and directs the car as to how it should respond and safely interact with the environment.

In an oversimplified description, the computer acts similarly to a brain as it combines and analyzes the information in order to ultimately direct the actions of the vehicle. The ability to analyze and direct the vehicle is an extremely complicated process that involves complex communication between advanced technology features, an aspect that has significant implications for the legal issues regarding autonomous vehicles. In traditional vehicles, the human driver analyzes the surrounding environment and makes decisions about how to direct the vehicle based on that analysis. In autonomous vehicles, however, the computer makes this “decision” based on data collected by the various other hardware features. This computer-based decisionmaking raises several questions: What happens when this decisionmaking process results in an accident? What if Google’s self-driving car swerves into oncoming traffic to avoid a toddler crossing the street?

B. Overview of Current Products Liability Law

In modern products liability cases, there are four primary claims under which a plaintiff may seek remedy from the manufacturer or seller of a product: negligence, breach of warranty, tortious misrepresentation, and strict liability. In any products liability case, a plaintiff must show that (1) the defendant sold a defective product, and (2) the defect was the proximate cause of the plaintiff’s injury. While breach of warranty and tortious misrepresentation claims may be

---

45 Id.
46 Id.
47 Id.
49 See id.
50 See NIDHI KALRA ET AL., supra note 15, at 8.
51 Id.
somewhat interesting when applied to autonomous cars, products liability claims seem to require a more robust discussion. Further, a discussion of strict liability largely parallels suits brought under negligence, and thus a large part of the analysis is similar. Moving to the different claims brought in automobile cases, crashworthiness does not, at first glance, present any uniquely difficult issues. While autonomous technology may modify vehicle design, the doctrine of crashworthiness seems to be equipped to address any new issues. Thus, for purposes of impact and brevity, this Note will primarily address autonomous cars under a strict liability scheme and will detail liability for automobile defects that cause accidents.

Strict products liability is considered the prominent legal theory by which plaintiffs can recover for losses caused by defective products. The roots of strict liability are found in Justice Traynor’s concurring opinion in the case of Escola v. Coca Cola Bottling Co. Justice Traynor lists several reasons supporting the implementation of a strict liability scheme with regard to products, notably that the manufacturers and sellers of products are better suited than the public to bear the burden when accidents occur. Traynor’s opinion later became embodied in the Restatement (Second) of Torts (hereinafter Second Restatement) which ascribes liability to one who sells a product in a “defective condition unreasonably dangerous to the user,” even if the seller has “exercised all possible care in the preparation and sale of [the] product.” While the Restatement (Third) of Torts: Products Liability (hereinafter Third Restatement) has attempted to push back on the concept of strict liability, it has not been widely recognized and the majority of courts follow an approach more analogous to that listed in the Second Restatement.

---

54 KRAUSS, supra note 52, at 41.
55 1 LOUIS R. FRUMER & MELVIN I. FRIEDMAN, PRODUCTS LIABILITY § 8.01(1) (Cary Stewart Sklaren ed., 2012).
56 150 P.2d 436, 440 (Cal. 1944) (Traynor, J., concurring).
57 Id. at 440–41.
58 RESTATEMENT (SECOND) OF TORTS § 402(A) (1965). The full section reads:

(1) One who sells any product in a defective condition unreasonably dangerous to the user or consumer or to his property is subject to liability for physical harm thereby caused to the ultimate user or consumer, or to his property, if (a) the seller is engaged in the business of selling such a product, and (b) it is expected to and does reach the user or consumer without substantial change in the condition in which it is sold. (2) The rule stated in Subsection (1) applies although (a) the seller has exercised all possible care in the preparation and sale of his product, and (b) the user or consumer has not bought the product from or entered into any contractual relation with the seller.

Id.
59 KRAUSS, supra note 52, at 72 n.3.
Recovery under a strict liability scheme similar to that detailed in the Second Restatement requires that a product be defective in some manner.\textsuperscript{60} There are three primary defect claims available to plaintiffs in products liability cases: a manufacturing defect claim, a design defect claim, and a warning defect claim (sometimes referred to as an informational defect).\textsuperscript{61} With respect to autonomous vehicles, manufacturing defect and design defect claims seem to present the most profound issues. Warning defect claims generally focus on the safety information that a seller gives a buyer and disclosure of potential risks.\textsuperscript{62} For purposes of brevity, the following paragraphs will focus solely on manufacturing and design defect claims.

1. Manufacturing Defect Claims

In general, strict liability in tort has been, and currently is, the preferred method for recovery in manufacturing defect cases.\textsuperscript{63} The backbone of manufacturing defect claims is that “while consumers may abstractly comprehend the practical necessity of allowing imperfect production, their actual expectation when purchasing a new product is that its important attributes, including safety, will match those of other similar units.”\textsuperscript{64} In many cases, there has been no specific test applied to manufacturing defect claims.\textsuperscript{65} In recent years, however, courts have begun to primarily use two different tests to apportion liability.\textsuperscript{66} The first is a “departure from intended design” test, which functions as the name implies: when a design departs from the original, liability may be imposed.\textsuperscript{67} Second is the malfunction doctrine, which is used when circumstances suggest a product defect, but there may be no direct evidence of one.\textsuperscript{68} Such an approach requires a more circumstantial means of demonstrating liability.\textsuperscript{69}

The departure from the intended design test stems from the Third Restatement, which states that a product “contains a manufacturing defect when the product departs from its intended design even though all possible care was exercised in the preparation and marketing of the product.”\textsuperscript{70} Since the publishing of the Third Restatement in 1998, an increasing number of courts have used some variation of this approach when defining manufacturing defects,\textsuperscript{71} and several

\begin{flushleft}
\textsuperscript{60} Restatement (Second) of Torts § 402(A).
\textsuperscript{61} See Krauss, supra note 52, at 73.
\textsuperscript{62} Id. at 41.
\textsuperscript{63} David G. Owen, Manufacturing Defects, 53 S.C. L. Rev. 851, 863 (2002).
\textsuperscript{64} Id. at 862.
\textsuperscript{65} Id. at 865.
\textsuperscript{66} Id.
\textsuperscript{67} Id. at 865–75.
\textsuperscript{68} Id.
\textsuperscript{69} Id. at 865.
\textsuperscript{70} Restatement (Third) of Torts: Products Liability § 2(a) (1998).
\textsuperscript{71} See, e.g., Perez-Trujillo v. Volvo Car Corp., 137 F.3d 50, 53 (1st Cir. 1998) (holding that a manufacturing defect exists if the product differs from the manufacturer’s
\end{flushleft}
states have enacted statutes defining manufacturing defects in a similar fashion.\textsuperscript{72} In statutes and cases that address manufacturing defect claims under the departure from design theory, plaintiffs establish defectiveness either by comparing the product that caused the incident to the manufacturer’s formal design specifications, or by comparing the faulty product with the dimensions and other parameters of some otherwise identical product.\textsuperscript{73}

In many manufacturing defect cases a plaintiff can use direct evidence that a product had an identifiable production defect and that the defect caused the injury.\textsuperscript{74} Sometimes a product malfunctions, however, and evidence of a defect is not apparent, or perhaps does not exist at all.\textsuperscript{75} In cases where direct evidence is seemingly unavailable, plaintiffs can still make a prima facie products liability case under the malfunction doctrine.\textsuperscript{76} The malfunction doctrine is similar to a \textit{res ipsa loquitur} approach, and a court may allow the inference of a defect if (1) the product malfunctioned, (2) the malfunction occurred during regular and proper use of the product, and (3) the product was not altered or misused in such a way that could cause the malfunction.\textsuperscript{77} While the malfunction doctrine and \textit{res ipsa loquitur} share several aspects in common, scholars and courts are quick to note that a \textit{res ipsa loquitur} approach is different because it still raises the question of negligence, whereas the malfunction doctrine works under strict liability where negligence is not applicable.\textsuperscript{78}

In applying the malfunction doctrine, courts generally take one of two approaches. Some courts treat the malfunction doctrine as a strict test, whereas others treat it as more of a principle of evidence that allows circumstantial

\begin{itemize}
\item \textsuperscript{72} See LA. REV. STAT. ANN. § 9:2800.55 (2009); N.J. STAT. ANN. § 2A:58C -2 (West 2000); OHIO REV. CODE ANN. § 2307.74 (LexisNexis 2010).
\item \textsuperscript{73} Owen, supra note 63, at 870.
\item \textsuperscript{74} \textit{Id.} at 872.
\item \textsuperscript{75} \textit{Id.}
\item \textsuperscript{76} \textit{Id.} at 873. The doctrine is used by name in several jurisdictions, including Pennsylvania, which has the most developed history on the theory. \textit{See}, e.g., Rogers v. Johnson & Johnson Prods., Inc., 565 A.2d 612, 616 (Pa. App. Ct. 1995) (finding a manufacturing defect “because the measurements of the parts of the wrench were shown not to comply with the manufacturer’s specifications”); Allstate Ins. Co. v. Ford Motor Co., 772 So. 2d 339, 344 (La. Ct. App. 2000) (holding a manufacturing defect exists where a “product deviated in a material way from the manufacturer’s specifications or performance standards for the product or from otherwise identical products manufactured by the same manufacturer”).
\item \textsuperscript{77} Owen, supra note 63, at 870.
\item \textsuperscript{78} 1 DAVID G. OWEN ET AL., MADDEN & OWEN ON PRODUCTS LIABILITY § 7:12, at 431 (3d ed. 2000).
\end{itemize}
evidence to be used to infer a defect. The malfunction doctrine as embodied in the Third Restatement is frequently applied in cases involving automobiles. Automobile accident cases fit well within the malfunction doctrine because in many cases the evidence showing a manufacturing defect is entirely destroyed or damaged beyond use. In particular, plaintiffs tend to use the malfunction doctrine in cases of steering failures, brake failures, and cases where inexplicable acceleration causes an accident. While the malfunction doctrine offers an attractive option for plaintiffs in automobile products liability cases, it is not limitless. Courts note that while the doctrine does provide for the admission of circumstantial evidence to show a defect, “the law will not allow plaintiffs or juries to rely on guess, conjecture, or speculation.”

2. Design Defect Claims

Design defect claims generally highlight defects that have the potential to be much more widespread and catastrophic while manufacturing defect claims usually address a single incident where an anomaly in the manufacturing process has caused an unsafe alteration to one product. Design defect claims generally

---

79 Id.
80 Section 3 states that a defect can be inferred when it “a) was of a kind that ordinarily occurs as a result of a product defect; and b) was not, in the particular case, solely the result of causes other than the product defect existing at the time of sale or distribution.” Restatement (Third) of Torts: Products Liability § 3 (1998).
81 Owen, supra note 63, at 875.
82 See, e.g., Stewart v. Ford Motor Co., 553 F.2d 130 (D.C. Cir. 1977) (involving a vehicle that was only twelve days old and had traveled just over one thousand miles); Stewart v. Budget Rent-A-Car Corp., 470 F.2d 240 (Haw. 1970) (involving a vehicle with nearly three thousand miles); Caprara v. Chrysler Corp., 423 N.Y.S.2d 694 (App. Div. 1979), aff’d on other grounds, 417 N.E.2d 545 (N.Y. 1981).
83 See, e.g., Tweedy v. Wright Ford Sales, Inc., 357 N.E.2d 449, 450 (Ill. 1976) (involving brakes that failed to perform on a vehicle that had been driven only 7,500 miles); Vernon v. Stash, 532 A.2d 441 (Pa. Super. Ct. 1987) (involving the failure of a parking brake); Darryl v. Ford Motor Co., 440 S.W.2d 630, 632 (Tex. 1969) (involving the failure of brakes on a truck that was only months old and had fewer than one thousand miles).
84 See, e.g., Wakabayashi v. Hertz Corp., 660 P.2d 1309, 1311 (Haw. 1983) (involving a vehicle that was nearly two years old that had been driven over twenty-two thousand miles); Juris v. Ford Motor Co., 752 So. 2d 260, 262 (La. Ct. App. 2000) (involving the failure of cruise control to correctly disengage); Phipps v. Gen. Motors Corp., 363 A.2d 955, 956 (Md. 1976) (involving the failure of an accelerator in a truck).
86 Krauss, supra note 52, at 81.
87 Id.
involve numerous products and sometimes even an entire line of automobiles. When plaintiffs make a design defect claim, they are not asserting that an individual product was defective because it was not built according to the design specifications—as would be the case in a manufacturing defect claim—but that the design as a whole was defective. This presents a much larger issue for manufacturers and retailers because, if the design defect claim is valid, then defects exist in their entire line of products, subjecting them to potentially massive liability costs. Automobile design defects are often the cause of major product safety recalls, and if not quickly remedied, could bring about massive class action lawsuits.

Nearly all states follow some variation of the Second Restatement’s embodiment of design defect, which ascribes strict liability to manufacturers or retailers if the design is defective and unreasonably dangerous. While the Third Restatement seemingly has attempted to push back at the notion of strict liability for design defect, courts still tend to follow a rule that is more in line with the Second Restatement. When deciding design defect claims, courts use two primary tests to determine the defectiveness of a design: (1) the consumer expectations test, and (2) the risk-utility test.

The consumer expectations test evaluates the defectiveness of a design based on whether “the danger posed by the design is greater than an ordinary consumer would expect when using the product in an intended or reasonably foreseeable manner.” Under the consumer expectations test, consumers must have “sufficient knowledge or familiarity with the design of the product to have reasonable expectations about its safety and performance.” Although such a definition is

---

88 Id.
89 Id.
91 KRAUSS, supra note 52, at 83; RESTATEMENT (SECOND) OF TORTS § 402A (1965).
92 See KRAUSS, supra note 52, at 83. The Third Restatement states that a design is defective when the foreseeable risks of harm posed by the product could have been reduced or avoided by the adoption of a reasonable alternative design by the seller or other distributor, or a predecessor in the commercial chain of distribution, and the omission of the alternative design renders the product not reasonably safe.
93 KRAUSS, supra note 52, at 84–89 (discussing consumer expectation test); OWEN ET AL., supra note 78, § 8:5, at 463 (discussing risk-utility test).
95 Id. at 138–39.
applicable in some cases, the consumer expectations test is often difficult to apply to complex products such as automobiles.96

While the consumer expectations test is still in use in certain jurisdictions, there has been some discontent with its application.97 As a result, the risk-utility test has emerged as the dominant test in design defect claims.98 Under the risk-utility test, plaintiffs can succeed in proving a design defect if they demonstrate “the product’s design proximately caused [their] injury and the defendant fails to establish, in light of the relevant factors, that, on balance, the benefits of the challenged design outweigh the risk of danger inherent in such design.”99

Analysis under the risk-utility test falls in line with principles first set forth in an article by Dean John Wade, the American Law Institute’s Reporter for the Second Restatement.100 Dean Wade lists seven factors to weigh in considering the risk-utility test:

(1) The usefulness and desirability of the product—it’s utility to the user and to the public as a whole.
(2) The safety aspects of the product—the likelihood that it will cause injury, and the probable seriousness of the injury.
(3) The availability of a substitute product which would meet the same need and not be as unsafe.
(4) The manufacturer’s ability to eliminate the unsafe character of the product without impairing its usefulness or making it too expensive to maintain its utility.
(5) The user’s ability to avoid danger by the exercise of care in the use of the product.
(6) The user’s anticipated awareness of the dangers inherent in the product and their avoidability, because of general public knowledge of the obvious condition of the product, or of the existence of suitable warnings or instructions.
(7) The feasibility, on the part of the manufacturer, of spreading the loss by setting the price of the product or carrying liability insurance.101

These factors are used to aid the court in deciding when there is a defect in design, but due to the nature of their composition, often result in a subjective analysis of defect. The ultimate goal of the risk-utility test is to determine if an accident could

96 See id. at 139.
97 KRAUSS, supra note 52, at 87, 89.
98 See John W. Wade, On the Nature of Strict Tort Liability for Products, 44 Miss. L.J. 825, 837–38 (1973) (providing an early example of the discontent with the consumer expectations test and noting the need for an alternative test in design defect cases).
100 KRAUSS, supra note 52, at 88; Wade, supra note 98, at 837–38.
101 KRAUSS, supra note 52, at 88; Wade, supra note 98, at 837–38.
have caused less damage or been avoided altogether if a reasonable alternative design had been used.102

Both the consumer expectations test and risk-utility test have been applied to cases involving automobile design defects. Each theory presents its own advantages and disadvantages, and as such, some states use a combination of the two tests.103 Regardless of which theory is applied, or even if a combination of the two is used, each approach can ultimately require a subjective analysis by the decisionmaker. Especially in cases of new technology, it becomes increasingly difficult to determine reasonable expectations about a product and its safety features. Further, when a technology is novel, it is more difficult to determine if an alternative design—if there is one—would have provided better safety.

III. PRODUCTS LIABILITY APPLIED TO AUTONOMOUS VEHICLES: SQUARE PEG, ROUND HOLE

According to U.S. census data, from 2006 until 2009 in the United States alone, there were over thirty million automobile accidents resulting in over one hundred thousand fatalities.104 Of the thousands of traffic accidents that occur each day, human error is involved in many of those cases.105 If early estimates about autonomous vehicles prove to be even close to accurate, their widespread implementation could lead to one of the greatest safety advances in decades.106 Especially in cases where distracted driving causes an accident, such as with drowsy driving or use of a cell phone, autonomous cars have serious potential to make the roadways a safer place for everyone involved.

While the potential safety benefits of autonomous vehicles are fairly clear, the legal course for them is not. There are thousands of vehicle accidents each day, and while autonomous vehicles are likely to be much safer than their human counterparts, they will inevitably be involved in accidents.107 The following sections will address autonomous vehicle accidents and two potential forms of liability. Section A shows that applying conventional products liability law to autonomous vehicles will be difficult and inconsistent, will likely harm consumers and potential plaintiffs, and will delay the implementation of a beneficial safety technology. To avoid the problems with conventional products liability, Section B

---

102 See Owen et al., supra note 78, § 8:5, at 463 (quoting McCarthy v. Olin Corp., 119 F.3d 148, 155 (2d Cir. 1997)).
105 See generally Nat’l Highway Traffic Safety Admin., supra note 10, at 2 (noting with particularity the accidents, such as those caused by drunk driving where human error is a component).
106 See Thrun, supra note 2 (suggesting that autonomous vehicle technology has the potential to reduce traffic accident casualties by as much as half).
107 See U.S. Census Bureau, supra note 104.
suggests a federal no-fault compensation system, similar to the system used for vaccine product claims, as a more efficient and stable regime for liability claims relating to autonomous vehicles.

A. Problems Applying Current Products Liability Law to Autonomous Vehicles

Products liability cases involving traditional automobiles are difficult enough as it is. Manufacturers may be subjected to liability even in cases where no definitive evidence of defect can be shown. On the flip side, injured plaintiffs may suffer because of the difficulty involved in proving a defect in complicated technology like automobiles. The consequences to both manufacturers and plaintiffs will likely only become more difficult with the advent of autonomous vehicles. The following sections address the problems associated with bringing a claim involving an autonomous vehicle. Section 1 provides an analysis of autonomous vehicle crashes for manufacturing defect claims, and Section 2 does the same for design defect claims.

1. Autonomous Vehicles Under a Manufacturing Defect Claim

Autonomous vehicles, when implemented, will likely cause a significant shift in liability away from customers and onto manufacturers. The extent to which this will occur will largely depend on the development of the law relating to autonomous vehicles. Currently, states have been hesitant to give autonomous vehicles full control, and the early state statutes that address autonomous vehicles require a licensed driver to be able to take control of the vehicle in the event of some system failure. Under such statutes, it seems possible that manufacturers could deflect some liability for accidents based on principles similar to human failures in traditional accidents.

---


111 NIDHI KALRA ET AL., supra note 15, at ix.

Nevertheless, autonomous vehicles are generally designed to replace human intervention with a safer alternative.\textsuperscript{113} By having the vehicle perform the tasks of a human driver, however, manufacturers risk the vehicle incurring the same liability that a human driver would. How long it will take for autonomous cars to replace human drivers is uncertain,\textsuperscript{114} but it is inevitable that these vehicles are going to cause accidents, and plaintiffs will inevitably seek recourse. Of the products liability claims discussed in this Note, manufacturing defect claims are the most natural means to address autonomous vehicles. While manufacturing defect claims may currently be best suited to address autonomous vehicles, they are not without significant problems. Namely, this Note argues that manufacturing defect claims could provide significant obstacles to plaintiffs seeking redress, enhance the uncertainty of liability costs to manufacturers, and ultimately slow the implementation of autonomous vehicles.

While autonomous vehicles will present unique issues and hurdles under manufacturing defect claims, it is important to note that there are many instances where such claims may be well equipped to deal with autonomous vehicles. If a radar sensor is not installed according to specifications and that faulty installation causes an accident, the manufacturer will be liable. In general, it seems apparent if autonomous vehicles are manufactured incorrectly or components fail, liability will fall to the manufacturer. In such clear situations, the manufacturing defect test is completely capable of addressing autonomous vehicles defects. While these claims will afford future plaintiffs redress, they will likely not occur with any type of frequency. Advances in manufacturing and fabrication technology have reduced the rate of defects and will likely continue to improve in the future.\textsuperscript{115}

Autonomous vehicles are likely to improve the safety of vehicle travel, but it will be many years, if ever, before human drivers are ultimately replaced. What that means is that there will continue to be drivers who drive while drowsy or are distracted by their cell phones, and ultimately there will continue to be human error in driving. While autonomous cars will likely enhance vehicle travel safety, it is seemingly impossible that they will be able to navigate the roads entirely without incident.

In addition to cases where a defect in production is clear, the malfunction doctrine can also offer an avenue for recovery in cases where autonomous vehicles crash, but no direct evidence of a defect is available.\textsuperscript{116} Recall that the malfunction doctrine contains three basic elements: (1) the product malfunctioned, (2) the malfunction occurred during a regular and proper use of the product, and (3) the product was not altered or misused in a way that could lead to the malfunction.\textsuperscript{117} By allowing circumstantial evidence, the malfunction doctrine allows a

\textsuperscript{113} For a more detailed description of how autonomous vehicles perform the tasks of a traditional human driver, see supra Part II.A.
\textsuperscript{114} See text accompanying and sources cited supra note 20.
\textsuperscript{115} NIDHI KALRA ET AL., supra note 15, at 28.
\textsuperscript{116} See Owen, supra note 63, at 873 (“[A] product defect may be inferred by circumstantial evidence . . . ”).
\textsuperscript{117} See supra Part II.B.1.
decisionmaker to infer the presence of a defect. While this aspect of the malfunction doctrine can be applicable to autonomous vehicles, it presents problems for both manufacturers and plaintiffs.

The malfunction doctrine would likely negatively affect manufacturers of autonomous vehicles because it could subject them to liability in cases where no defect was present. Driving is a notoriously unsafe practice, and while autonomous vehicles can improve that, it is still a risk-involved activity. With traditional human drivers behind the wheel there are occasionally no-fault accidents. Cars crash, but responsibility is not necessarily pinned on one specific individual. With autonomous cars, that would likely change. The malfunction doctrine permits manufacturing defect claims where evidence allows the inference of a defect. At a very basic level, autonomous vehicles take control away from the driver and place it with the vehicle. In cases where autonomous vehicles crash, the very fact that the vehicle crashed seems to be evidence that supports a defect. Autonomous vehicles are not intended to crash, and so when one does, it would seem that plaintiffs would be incentivized to bring a products liability suit under the malfunction doctrine.

It will likely take considerable time for courts to develop a predictable jurisprudence with respect to the malfunction doctrine as applied to autonomous vehicles. If plaintiffs are incentivized to bring a manufacturing defect claim every time an autonomous vehicle crashes, then during such a period the costs of defending constant product liability claims would place a heavy burden on manufacturers. While predictions are that autonomous vehicles will be safer than their human driver counterparts, it is impossible to predict how frequently they will be involved in accidents. Regardless of the circumstances of a potential accident, because autonomous vehicles take control away from a human driver, any crash draws the manufacturer’s liability into question. Thus, it is certainly rational for manufacturers to be reluctant to subject themselves to such speculative liability. If the economic incentive is overcome by the costs of potential litigation, society may be deprived of the significant safety benefits that autonomous vehicle could provide.

The malfunction doctrine also presents problems for plaintiffs. It is entirely conceivable that an autonomous vehicle may crash while operating in a fully functional capacity. For instance, imagine a toddler runs across the street in front of an autonomous car and the car slams on the brakes and swerves to avoid the toddler, ultimately crashing into another vehicle. The car did not have a defect that caused the accident, but performed according to design when swerving to avoid the toddler. With traditional vehicles, when a driver overreacts to an environmental circumstance and causes an accident, the driver is liable.

When human drivers operate a vehicle, they analyze their surroundings, make decisions, and take action. Autonomous vehicles however, require a computer to make such decisions based on the environment around the vehicle. As a result,

---

118 Owen, supra note 63, at 873.
119 NIDHI KALRA ET AL., supra note 15, at xi.
the malfunction theory gets significantly more complex when applied to autonomous vehicles. Everything on the car could be functioning according to specifications, but the vehicle just interprets the data in a way that ultimately causes an accident. Effectively then, when no direct evidence of a defect is present, plaintiffs will have to question the “decision” made by the technology. Such an allegation does not necessarily imply the vehicle malfunctioned according to the traditional use of the malfunction theory.

A computer incorrectly interpreting data is not the same as a steering wheel that locks, an accelerator that gets stuck, or a vehicle that catches on fire. The malfunction theory, at its most basic level, requires that plaintiffs allege some type of defect and allege the accident would not have happened but for that defect.\textsuperscript{120} The hypothetical situation posed above shows that it is not a stretch to believe that an autonomous vehicle could cause an accident without having some defect that causes a malfunction. Driving is filled with a multitude of variables that play into a human driver’s choice of action, and the same applies to autonomous vehicles. Thus, a plaintiff seeking to establish a manufacturing defect under the malfunction theory would have to prove the decision made by an autonomous vehicle would likely not have occurred but for some abnormality created by a failure in the manufacturing process.

Manufacturing defect claims as applied to autonomous vehicles could have two vastly different effects. On one hand, it could open up manufacturers to liability for basically any crash that involves an autonomous car. Autonomous vehicles take control from the human driver, and thus any crash could imply a defect in the production of the vehicle. On the other hand, a plaintiff could have an exceedingly difficult time proving that an autonomous vehicle’s response to its environment is a result of an abnormality in design. Regardless of who manufacturing defect claims ultimately benefit, the effect on the development of autonomous vehicles would be negative. If manufacturers are subjected to increased liability, then development will likely not be as aggressively pursued. But if plaintiffs are denied recovery, then there may be less incentive to buy, and thus a decreased market for autonomous vehicles. With the huge potential safety benefits that autonomous vehicles could provide, any such scenario would ultimately be of consequence to the general public.

2. Autonomou s Vehicles Under a Design Defect Claim

Advances in manufacturing technology make it unlikely that a large number of manufacturing defect suits would ultimately occur.\textsuperscript{121} Thus, the majority of claims relating to autonomous vehicles would likely be brought under design


\textsuperscript{121} NIDHI KALRA ET AL., supra note 15, at 28.
defect. But design defect claims are also fraught with problems and are likely less equipped to address autonomous car accidents than are manufacturing defect claims. The following paragraphs will show that both the consumer expectations test and the risk-utility test present significant hurdles for plaintiffs pursuing design defect claims for autonomous cars. Such hurdles would essentially make design defect claims unavailable to any plaintiffs injured by autonomous cars.

The first and less frequently used test for design defect is the consumer expectations test. At its core, this test seeks to prove the existence of a design defect based on evidence that a product does not conform to the reasonable expectations of a consumer. In some cases the consumer expectations test is a usable means of identifying a design defect. If a hammer is designed in such a way that the head flies off, the consumer expectations test fits well. A reasonable consumer would expect a hammer to perform its function of impacting nails without the head becoming dislodged and injuring someone. In such a scenario, consumers have common general expectations about the use of the product and are familiar with the functionality of a hammer. Further, because hammers are a simple product, users’ safety expectations can be articulated with a fair degree of consistency.

The difficulty with applying the consumer expectations test is that not every product is as simple as a hammer. As products become increasingly more complex, a great deal of uncertainty can arise with the consumer expectations test, primarily because the test largely stems from the negligence principle of reasonableness on the part of the ordinary consumer. The difficulty arises in determining where to draw the line for the basis of reasonableness. In the case of the traditional automobile, a very complex product, it becomes more difficult to apply the principle of reasonableness. A mechanic’s reasonable expectations about automobiles will likely be significantly different than a grandmother who only drives to church and back on Sundays. The challenge then becomes how specific a court will be when defining a reasonable consumer. Is an ordinary consumer an educated consumer, one with a mechanical engineering degree, or one who has a mechanical engineering degree who works for Ford? Depending on where courts draw the line, the expectations that courts deem “reasonable” may differ widely.

Autonomous vehicles only exacerbate the present problems with the consumer expectations test. A basic tenet of the consumer expectations test is that a consumer needs to have “sufficient knowledge or familiarity with the design of the product to have reasonable expectations about its safety or performance.” The technology that allows these vehicles to function is extremely advanced and novel, thus there are likely few individuals who have the requisite knowledge to have a reasonable expectation about the safety of these vehicles. This basically

122 KRAUSS, supra note 52, at 84–86.
123 See KIELY & OTTLEY, supra note 94, at 135.
124 Id. at 138.
125 See id. at 138–39 (noting that the complexity of the automobile “frequently tests the limits of the consumers’ expectations”).
eliminates for many years the consumer expectations test for plaintiffs injured by autonomous vehicles. Consumers will certainly have expectations about how autonomous vehicles will function, but considering that some courts choose not to apply the consumer expectations test to devices such as airbags, it will likely be a considerable time until courts are comfortable applying the test to autonomous cars.126

The second test commonly applied to design defect claims, the risk-utility test, would likely prove as difficult to apply as the consumer expectations test. The risk-utility test focuses on the safety benefits of a proposed design compared to alternative models in the same category.127 Early in the implementation of autonomous vehicles, the risk-utility test will likely not afford any remedy to plaintiffs. Dean Wade proposed seven factors to consider in the risk-utility test.128 Of those seven factors, three could be of particular interest when applied to autonomous vehicles. First, is “[t]he availability of a substitute product which would meet the same need and not be as unsafe.”129 Second, is “[t]he usefulness and desirability of the product—its utility to the user and to the public as a whole.”130 Third, is “[t]he user’s anticipated awareness of the dangers inherent in the product and their avoidability, because of general public knowledge of the obvious condition of the product, or of the existence of suitable warnings or instructions.”131

The first factor at issue focuses on the availability of alternate products that would perform the same functions and be less unsafe.132 It will likely be considerable time until there are a sufficient number of manufacturers involved in the production of autonomous vehicles. Accordingly, potential plaintiffs would be limited in their ability to show that there are safer alternate designs available.

Another interesting issue under the risk-utility test involves the social utility that autonomous vehicles serve. Analyzing autonomous vehicles under Dean Wade’s second factor could show there is a strong case in support of the argument that autonomous vehicles’ utility to the public as a whole is considerable. Specifically, autonomous vehicles are likely to be significantly safer than human drivers,133 and thus are of substantial utility to public safety.134 Furthermore, they will be of great utility to people with disabilities who cannot drive standard vehicles. Accordingly, manufacturers would seem to largely benefit from such a

126 See Pruitt v. Gen. Motors Corp., 86 Cal. Rptr. 2d 4, 6 (Ct. App. 1999) (noting that air bags are too complex to use the consumer expectations test).
127 Wade, supra note 98, at 837–38.
128 Id.
129 Id. at 837.
130 Id.
131 Id.
132 Id.
133 See Vijayenthiran, supra note 7 (noting that studies of early autonomous vehicle technologies demonstrate a reduction in accidents).
134 See Valdes-Dapena, supra note 7 (noting the potential safety benefits of autonomous vehicles).
consideration under the risk-utility test. When the utility of a product is considerable, it lends evidence that there may not have been a design defect.

The factor that becomes increasingly interesting when applied to autonomous vehicles is “the user’s anticipated awareness of the dangers inherent in the product and their avoidability, because of general public knowledge of the obvious condition of the product, or of the existence of suitable warnings or instructions.”\textsuperscript{135} In particular, it is interesting to hypothesize what potential autonomous vehicle purchasers will expect as the inherent danger of the cars. This question because more difficult in light of the current legislation regarding autonomous vehicles, which requires a licensed driver to be present as a backup to the technology.\textsuperscript{136} The fact that the law requires a driver to be present as a backup seems to support the idea that the car may fail, in which case the driver would be forced to take over. If this remains the case, then individuals whose autonomous vehicles crash may be without recourse. Specifically, they would lack evidence that they were not aware of the “dangers inherent in the product.”\textsuperscript{137}

Whether it is the consumer expectations or risk-utility test that is used, plaintiffs will likely have a difficult time bringing design defect claims against autonomous vehicle manufacturers. The technology is novel and complex, and it could be years before consumers can develop consistent safety expectations about the vehicles. The consumer expectations test would be difficult to apply in the early years of implementation. Further, the factors presented by Dean Wade would seem to support autonomous vehicle manufacturers. Absent a clear defect in design, autonomous vehicles present such a profound potential safety benefit that courts may be swayed to rarely find design defects. The limited alternate models and the clear possible risks would also lessen plaintiff’s chances of success.

\textbf{B. The Need for a No-Fault Scheme for Autonomous Vehicles}

Current products liability claims will be difficult to apply to autonomous vehicles and will ultimately disadvantage both plaintiffs and manufacturers.\textsuperscript{138} Plaintiffs will be limited in the claims that they can bring relating to autonomous vehicles. Those remaining claims will involve complex legal issues that will likely result in expensive and elaborate lawsuits. For manufacturers, the implementation of autonomous vehicles presents a significant shift in the layout of liability for accidents.\textsuperscript{139} Further, the jurisprudence on the matter will likely take some time to stabilize.\textsuperscript{140} The combination of the two makes cost predictions about autonomous

\begin{itemize}
  \item \textsuperscript{135} Wade, \textit{supra} note 98, at 837–38; see Valdes-Dapena, \textit{supra} note 7 (noting the potential safety benefits of autonomous vehicles).
  \item \textsuperscript{136} \textit{CAL. VEH. CODE} § 38750(b)(2) (West Supp. 2013); see \textit{FLA. STAT. ANN.} § 316.85(1) (West Supp. 2013); \textit{NEV. REV. STAT. ANN.} § 482A.200 (LexisNexis Supp. 2011); \textit{supra} text accompanying note 37.
  \item \textsuperscript{137} Wade, \textit{supra} note 98, at 837–38.
  \item \textsuperscript{138} \textit{See supra} Part III.A.
  \item \textsuperscript{139} \textit{Id.}; see also \textit{NIDHI KALRA ET AL.}, \textit{supra} note 15, at ix.
  \item \textsuperscript{140} \textit{See supra} Part III.A.
\end{itemize}
vehicles uncertain, which will likely delay the widespread availability of autonomous vehicles.\textsuperscript{141}

For many products, the problems detailed above would not be of significant concern to the public as a whole. Many would argue that delaying the time that a new ladder or new child’s toy takes to hit the market would be of no substantial concern. However, autonomous vehicles require a different sort of analysis. In 2010 there were over 5.4 million traffic accidents in the United States.\textsuperscript{142} Those accidents resulted in over two million injuries and over thirty thousand fatalities.\textsuperscript{143} Of the thirty thousand fatalities in 2010, about one-third were caused by an individual driving with a blood alcohol level of over 0.08\%, and many more were also caused by human error.\textsuperscript{144} It is undeniable that too many Americans have been injured or killed in automobile accidents, many of which could have been avoided if the driver had not been distracted or made a poor decision. The sum total of these accidents likely results in millions of dollars in costs for healthcare, insurance, and damage repair. Overall, the damage caused by automobiles, in particular human error in automobiles, takes a heavy toll on society.

Autonomous cars have the ability to change all of that. If fully implemented, accidents caused by a drunk driver could be completely eliminated. An individual could leave the bar, have his car pick him up, and never put other drivers at risk.\textsuperscript{145} Further, the disabled and the elderly could enjoy unprecedented levels of mobility.\textsuperscript{146} The environment would benefit as travel becomes more efficient. Productivity and individual freedom could skyrocket as individuals gain significantly more free time. The benefits of autonomous vehicles are profound, and their implementation is of the upmost importance for all of the above reasons and more.

While such a utopian world created by autonomous cars is likely far in the future, society cannot afford to delay the use of these vehicles in our day-to-day lives. Thirty-thousand people die each year in traffic accidents, and this technology has the potential to change that.\textsuperscript{147} For these reasons, a no-fault compensation scheme, similar to that created by the National Childhood Vaccination Injury Act, should be implemented to address autonomous vehicles. Such a no-fault scheme would create a roadmap for the future of autonomous vehicles, ease manufacturer concerns, and ultimately encourage autonomous vehicle development. Further, such a scheme would provide consistent and available remedies for plaintiffs who are injured when autonomous vehicles inevitably crash.

\textsuperscript{141} See id.
\textsuperscript{142} NAT’L HIGHWAY TRAFFIC SAFETY ADMIN., supra note 10.
\textsuperscript{143} Id.
\textsuperscript{144} See NAT’L HIGHWAY TRAFFIC SAFETY ADMIN., supra note 12.
\textsuperscript{145} It is important to note that this would be in violation of the current laws addressing autonomous vehicles.
\textsuperscript{146} Jeremy A. Kaplan, Amazing! Google’s Self-Driving Car Allows the Blind to Drive, FOXNEWS.COM (Mar. 28, 2012), http://www.foxnews.com/tech/2012/03/28/amazing-google-self-driving-cars-allow-blind-to-drive/.
\textsuperscript{147} See NAT’L HIGHWAY TRAFFIC SAFETY ADMIN., supra note 12.
1. Overview of the National Childhood Vaccination Injury Act (NCVIA)

The NCVIA came about due to two primary issues during the 1980’s: the “inadequate and inconsistent nature of existing state tort remedies and the instability of vaccine supplies due to prior litigation.”\(^{148}\) Vaccine makers had become the target of considerable liability, which caused several manufacturers to leave the market.\(^{149}\) Further, there was significant pressure from pro-plaintiff groups who were concerned about the inconsistency of the tort system with respect to vaccine injury claims.\(^{150}\) A culmination of these concerns spurred Congress to pass the NCVIA.\(^{151}\)

The NCVIA created a no-fault compensation system as an alternative to the ineffective state tort system.\(^{152}\) The federal system eliminated the need for plaintiffs to demonstrate that there was a defect in the vaccine.\(^{153}\) Plaintiffs seeking redress could file a petition in the United States Court of Federal Claims and provide service on the Secretary of Health and Human Services.\(^{154}\) Their petition could contain one of two claims: an on-table claim or an off-table claim.\(^{155}\) An on-table claim refers to a table that contains a list of vaccine-related injuries and relevant timelines for pursuing those claims.\(^{156}\) If a petitioner cannot demonstrate an on-table claim then they must pursue an off-table claim, which requires a higher standard of proof and evidence of causation.\(^{157}\) It is important to note the NCVIA does not entirely remove the availability of the state tort system, but rather just requires plaintiffs to assert claims through the NCVIA first.\(^{158}\) If petitioners are not satisfied with the adjudication under the NCVIA, they can move into state or federal court.\(^{159}\) The NCVIA, however, does remove remedies in tort for injuries that are the “unavoidable side effects” of vaccinations.\(^{160}\)

Petitioners who successfully bring a claim are rewarded damages from the Vaccine Injury Compensation Trust Fund.\(^{161}\) A $0.75 excise tax on recommended


\(^{150}\) Breen, supra note 148, at 316.

\(^{151}\) Id. at 1856.

\(^{152}\) Id. at 1860.

\(^{153}\) Breen, supra note 148, at 317.

\(^{154}\) Id. at 318.

\(^{155}\) See id.

\(^{156}\) Id.; see 42 C.F.R. § 100.3 (2011) (providing the most recently revised Vaccine Injury Table).

\(^{157}\) Breen, supra note 148, at 318.

\(^{158}\) Cantor, supra note 149, at 1861.

\(^{159}\) Id.

\(^{160}\) Id.

vaccines is used to generate the funds for awards given to successful claimants. Administration of awards under the NCVIA is generally less than that of the sometimes excessive damages awarded under tort law. By standardizing the awards, the NCVIA provided vaccine manufacturers with more certain liability. Further, the NCVIA provided a more efficient process for plaintiffs to seek remedy.

2. A No-Fault Compensation System for Autonomous Vehicles

The NCVIA has several components particularly relevant to the legal landscape surrounding autonomous vehicles. A federal no-fault system would provide a clear picture of liability for manufacturers seeking to develop and implement autonomous vehicles. By accurately gauging their potential liability costs, manufacturers could better price and develop autonomous vehicles. Further, plaintiffs would ultimately benefit from a standardized and streamlined process similar to that created under the NCVIA. The potential pitfalls associated with current products liability would be mitigated, and plaintiffs could seek recourse without the complex and costly litigation they would otherwise be forced to pursue.

While the reasoning for the enactment of the NCVIA is in some ways unique to vaccinations, there are several similar principles relating to autonomous vehicles. Vaccines focus on prevention and public safety. Autonomous cars do the same. Human error while operating an automobile accounts for thousands of American deaths each year, and autonomous vehicles have the potential to change this. Further, similar to vaccinations, there will be unavoidable side effects from using autonomous vehicles, and injuries are sure to eventually result. Autonomous vehicles will inevitably crash, albeit hopefully less frequently, and injuries will result. An additional similarity is that state tort systems are likely ill-equipped to deal with autonomous vehicle liability. A similar pattern is likely to emerge if manufacturers cannot ascertain a more certain road ahead for their potential liability costs. Like vaccines, though, a lack of development in autonomous vehicles could result in a significant consequence to public safety.

While the exact makeup of a no-fault compensation system is beyond the scope of this Note, there are a few points to consider going forward. First, the types of vehicles that are included under a no-fault scheme will be of the utmost importance and should be clearly and narrowly defined. Second, both manufacturing and design defect claims should be covered under the system, but not informational defects. One of the most important aspects of implementing autonomous vehicles is that the public must understand what the vehicles are truly

---

162 Id.
163 See Valdes-Dapena, supra note 7 (noting the potential safety benefits of autonomous vehicles).
164 See supra Part III.A (noting the difficulties associated with applying current products liability law to autonomous vehicles).
capable of and how the vehicles can be safely used. Lastly, the program should be enacted for a set time. Autonomous cars are novel technology with profound benefits but they are not the last stop. An endless no-fault compensation scheme may deter advancements in technology by encouraging the development of technology that only fits under the system.

IV. CONCLUSION

Autonomous vehicles have the potential to change the future of transportation, enhance travel safety, improve the environment, help the disabled, and ultimately bring more freedom to our daily lives. For autonomous vehicles to reach their full potential, the legal system must be ready to adopt new policies that will advance the interests of manufacturers and potential plaintiffs alike. The current liability system, namely strict products liability, does not provide such an environment. Manufacturing and design defects are ill-suited to address the unique issues raised by autonomous vehicles. This uncertain and ill-fitting legal landscape will likely deter manufacturers from producing autonomous vehicles and preclude those injured by such technology from receiving an adequate remedy. A no-fault compensation system, similar to that set up under the NCVIA would solve both problems. Manufacturers would be able to accurately gauge liability costs and would be encouraged to produce this new technology. Plaintiffs would benefit as well from access to a consistent and efficient system of remedies. While a no-fault compensation regime is a unique solution, autonomous vehicles are a unique technology. By fostering development while maintaining legal redress, a no-fault system will benefit society as a whole by furthering this transformative new technology.