ARTICLES

RECONSTRUCTING FORESEEABILITY

W. Jonathan Cardi

[pages 921–988]

Abstract: This Article is the second of a two-part endeavor assessing the use of foreseeability in negligence law, and arguing against its ever-expanding role in the element of duty. The first article, Purging Foreseeability, urged courts to adopt the general duty provisions of the proposed Restatement (Third) of Torts—provisions that would purge duty of foreseeability-based considerations. Courts and scholars have resisted the Restatement’s attempts to rid duty of foreseeability out of a jurisprudential view that foreseeability simply “belongs” there. This Article urges that the conceptual work done by foreseeability also might fit wholly and seamlessly within the elements of breach and proximate cause. In proving that foreseeability’s conceptual fit is thus indeterminate, the Article aims to refocus the debate on whether the court or the jury is the better arbiter of foreseeability—a matter that courts are reluctant to discuss and scholars have largely ignored.

OUTSOURCING POWER: HOW PRIVATIZING MILITARY EFFORTS CHALLENGES ACCOUNTABILITY, PROFESSIONALISM, AND DEMOCRACY

Martha Minow

[pages 989–1026]

Abstract: Private contractors have played key roles in recent high-profile scandals. These scandals hint at the degree to which the U.S. military has increased the scope and scale of its reliance on private security companies in recent decades. This trend offers many advantages, including nimbleness in the deployment of expertise and geographic flexibility. But it also departs from conventional methods of accountability through both public oversight and private market discipline. The lack of transparency
in the use of private contractors compounds the problem of assessing the impact of their increasing role. Failures of basic governmental oversight to ensure contract enforcement by the Department of Defense are well-documented. Departures from conventional government contracting procedures exacerbate these failures and obscure whether inherently governmental functions are in effect privatized. The large sums of money involved contribute to risks of corruption and a scale of private lobbying that can distort the legislative process. These developments jeopardize the effectiveness of military activities, the professionalism of the military, the integrity of the legislative process and foreign policy decision making, public confidence in the government, national self-interest, and the stability of the world order.

NOTES

OH CANADA!: ANTITRUST GEOGRAPHIC MARKET DEFINITION AND THE REIMPORTATION OF PRESCRIPTION DRUGS

Maryan M. Chirayath

[pages 1027–1068]

Abstract: In recent years, public attention has focused on the need for affordable prescription drugs. Although Congress has recently enacted a Medicare Prescription Drug Plan, many private citizens and state and local governments continue to reimport prescription drugs from Canada to take advantage of the lower drug prices available in Canada. Many pharmaceutical companies have responded to this phenomenon by cutting off supplies of their drugs to Canadian pharmacies engaging in reimportation. As a result, some state and local governments have initiated litigation alleging state antitrust violations. This Note addresses one of the first questions raised in U.S. antitrust litigation under the Rule of Reason—the definition of the geographic market. First, this Note surveys the current caselaw standard and the academic approaches to geographic market definition. This Note then applies these approaches, concluding that Canada may be included in any geographic market when addressing the legality of reimportation in an antitrust context.

TAKING THE RELIGION OUT OF RELIGIOUS PROPERTY DISPUTES

Elizabeth Ehrlich

[pages 1069–1094]
Abstract: The current trend of local parish suppressions raises fundamental questions regarding ownership of gifted property. Disputes regarding ownership arise when a donor’s intent conflicts with the institution’s hierarchy claiming ownership of the gifted property. Hierarchical religious institutions like the Roman Catholic Church can assert such claims under two legal principles. First, because the First Amendment requires the separation of church and state, civil courts cannot interpret religious law. Second, because hierarchical religious institutions are often organized as corporations sole, local parishes are agents of the archdiocese. This Note argues for extending neutral principles of law currently used to settle intra-church property disputes to property disputes between religious institutions and individuals. In applying neutral principles of law, including the law of wills and trusts, courts can successfully adjudicate property disputes to ensure a testator’s intent is effectuated.

MEDICAL-MALPRACTICE REFORM: IS ENTERPRISE LIABILITY OR NO-FAULT A BETTER REFORM?

Kristie Tappan

[pages 1095–1130]

Abstract: This Note compares two medical-malpractice reforms: enterprise liability and no-fault. The Note compares the reforms for their relative ability to compensate injured patients and deter malpractice. The Note also examines the reforms’ economic and sociopolitical feasibility. The Note concludes that a no-fault medical-malpractice system would better compensate patients and deter malpractice, but enterprise liability is a more feasible reform that policymakers should pursue more aggressively.
RECONSTRUCTING FORESEEABILITY

W. JONATHAN CARDI*

Abstract: This Article is the second of a two-part endeavor assessing the use of foreseeability in negligence law, and arguing against its ever-expanding role in the element of duty. The first article, Purging Foreseeability, urged courts to adopt the general duty provisions of the proposed Restatement (Third) of Torts—provisions that would purge duty of foreseeability-based considerations. Courts and scholars have resisted the Restatement’s attempts to rid duty of foreseeability out of a jurisprudential view that foreseeability simply “belongs” there. This Article urges that the conceptual work done by foreseeability also might fit wholly and seamlessly within the elements of breach and proximate cause. In proving that foreseeability’s conceptual fit is thus indeterminate, the Article aims to refocus the debate on whether the court or the jury is the better arbiter of foreseeability—a matter that courts are reluctant to discuss and scholars have largely ignored.

Introduction

The concept of foreseeability is fast devouring the negligence cause of action. Foreseeability of a risk of injury has for centuries rested at the heart of court determinations of whether a defendant breached its duty of care.1 More recently, the foreseeability of a particular plaintiff’s injury has become central to the element of proximate cause.2 Foreseeability’s most aggressive advance of late, however,

* Assistant Professor of Law, University of Kentucky. B.A. 1991, Harvard University; J.D. 1998, University of Iowa. I am much indebted to Professors Mike Green, Ken Kress, Vincent Cardi, Albert Yoon, and Michael Healy for their gracious contributions. My sincere thanks also to John Goldberg and Benjamin Zipursky, whose correspondence and conversations sparked the research and thoughts that led to this Article.


2 DAN B. DOBBS, THE LAW OF TORTS § 180, at 444 (2000); MARC A. FRANKLIN & ROBERT L. RABIN, TORT LAW AND ALTERNATIVES 399 (7th ed. 2001). Many identify the emerging importance of foreseeability in proximate cause decisions with the Australian case, Overseas Tankship (U.K.), Ltd. v. Morts Dock & Eng’g Co. (The Wagon Mound), [1961] A.C. 388 (P.C.), although foreseeability had almost certainly played a role in proximate cause for some time before that case.

921
has been into the realm of duty. Emboldened by a series of decisions in the California Supreme Court culminating in *Tarasoff v. Regents of the University of California*, foreseeability now threatens to swallow the duty calculus whole, to become duty’s “unified theory.”

This Article is the second phase of a two-part endeavor to assess the use of foreseeability in negligence law and to make the case against its expansive role in duty. The first article reviewed foreseeability’s various doctrinal incarnations and urged courts to adopt the general duty provisions of the proposed *Restatement (Third) of Torts*—provisions that would largely purge duty of foreseeability-based considerations. As that piece argued, ridding duty of foreseeability would result in a number of significant benefits for negligence law. It would untangle negligence doctrine, which at present is confusingly redundant with regard to foreseeability. It would encourage judges to be overt about the public policy choices that underlie foreseeability-based duty decisions. Finally, and perhaps most importantly, it would place initial authority over the necessarily fact-dependent issue of foreseeability properly in the hands of the jury rather than the judge.

---

3 551 P.2d 334, 342 (Cal. 1976) (holding that among several factors to be considered in analyzing duty, “[t]he most important of these considerations . . . is foreseeability”). For a discussion of *Tarasoff*, see infra notes 271–305 and accompanying text.


5 The term “unified theory” is a reference to the search in theoretical physics for a single umbrella theory that explains the properties and relationships of the entire physical world. See Brian Greene, *The Elegant Universe* 424 (1999) (defining “unified theory” as “[a]ny theory that describes all four forces and all of matter within a single, all-encompassing framework”).

6 The Section reads: “An actor ordinarily has a duty to exercise reasonable care when the actor’s conduct creates a risk of physical harm.” *Restatement (Third) of Torts: Liab. for Physical Harm (Basic Principles)* § 7(a) (Tentative Draft No. 4, 2004).

7 W. Jonathan Cardi, *Purging Foreseeability*, 58 VAND. L. REV. 739, 769, 773–74 (2005). The *Restatement (Third)*’s duty approach will, however, only affect negligence cases that involve physical injury. Negligence cases involving economic or psychic harm are to be governed by another set of Restatement provisions yet to be drafted. *Restatement (Third) of Torts: Liab. for Physical Harm (Basic Principles)*, at *Introduction* (Tentative Draft No. 2, 2002).

8 Cardi, *supra* note 7, at 774–87.

9 Id. at 787–90.

10 Id. at 794, 799–804.
The arguments in favor of a foreseeability-free duty standard are largely normative. Despite the merits of such an approach, foreseeability remains a pervasive consideration in many courts’ duty analyses. There are several possible explanations for foreseeability’s popularity. First, judges sometimes find respite in easily comprehensible, if imperfect, solutions to complex legal dilemmas. Foreseeability provides such a solution to difficult duty questions—one that is not only intuitively sensible, but also safely woven into the historical fabric of common-law negligence. Furthermore, foreseeability has the sound of doctrine and yet is a malleable enough concept to serve as a vessel through which courts might render jurisprudentially palatable many of the policy decisions inherent to tort law. In addition, by infusing foreseeability into the element of duty—the only element of negligence decided by the court without deference to the jury—courts gain a large measure of control over juries. Much of the doctrinal work now accomplished by foreseeability as a part of duty has been traditionally relegated, in the first instance, to the jury in the context of breach and proximate cause. By co-opting as part of duty significant portions of the breach and proximate cause inquiries, courts avoid having to decide such matters pursuant to the “no reasonable jury standard,” thereby tightening their reins on the perceived caprice of juries. A final reason for courts’ increasing reliance on foreseeability in deciding duty—and the reason central to this Article—is that many feel that foreseeability simply “belongs there.” That is, courts and scholars often theorize duty such that foreseeability is a natural conceptual fit, and they explain foreseeability in ways that tie it necessarily to duty. This understanding of foreseeability’s relationship with duty can be traced in large measure to the decisions of Justice Cardozo in such watershed cases as MacPherson v. Buick Mo-

11 Id. at 762–67, 787–90; see also Thomas C. Galligan, Jr., A Primer on the Patterns of Negligence, 53 La. L. Rev. 1509, 1523 (1993) (suggesting that “judges should not rely on, or hide behind, words like . . . foreseeable, unforeseeable, . . . and whatever other magic mumbo jumbo courts could use to obfuscate the policies that were really at the heart of their decisions”); Patrick J. Kelley, Restating Duty, Breach, and Proximate Cause in Negligence Law: Descriptive Theory and the Rule of Law, 54 Vand. L. Rev. 1039, 1046 (2001) (describing foreseeability as “so open-ended [that] [it] can be used to explain any decision”).

12 Restatement (Second) of Torts § 328B (1965); Dobbs, supra note 2, § 149, at 355.

13 Cardi, supra note 7, at 744–50.

14 See infra notes 48–76, 98–324, and accompanying text.

tor Co.\textsuperscript{16} and \textit{Palsgraf v. Long Island Railroad Co.},\textsuperscript{17} although the movement toward this view has almost certainly been broader.\textsuperscript{18}

This Article refutes these conceptualizations of duty and foreseeability, and it questions the assertion that foreseeability necessarily plays a role in duty decisions. The Article urges that instead, the conceptual work done by foreseeability also might fit wholly and seamlessly within the elements of breach and proximate cause. In proving that foreseeability's conceptual fit is thus indeterminate, the Article aims to refocus the debate on the only determinative factor: whether court or jury is the better arbiter of foreseeability—a matter that courts are reluctant to discuss and that scholars, at least since Leon Green,\textsuperscript{19} have largely ignored.

The Article proceeds as follows: Part I offers a brief overview of the current doctrinal roles played by foreseeability in negligence cases.\textsuperscript{20} Part II examines the nature of a “cause of action” generally and the purpose served by defining a cause of action according to a series of elements.\textsuperscript{21} Part III then addresses the various conceptual roles played by foreseeability in negligence cases.\textsuperscript{22} With regard to each role save categorical foreseeability (that is, judicial determinations of foreseeability for entire categories of risks or injuries or plaintiffs), the subsections of Part III demonstrate that the conceptual work done by foreseeability might be accomplished equally well in the context of duty or as part of the breach or proximate cause analyses. With respect to categorical foreseeability, Part III.D urges that negligence would simply be better off without it.\textsuperscript{23} Finally, Part IV summarizes the primary points of the Article and briefly explores a general statement of duty that might lead courts toward a more lucid understanding of the concepts of foreseeability and duty.\textsuperscript{24}

\textsuperscript{16} 111 N.E. 1050 (N.Y. 1916).
\textsuperscript{17} 162 N.E. 99 (N.Y. 1928).
\textsuperscript{19} See generally Leon Green, \textit{Judge & Jury} (1930) (discussing at length the judge versus jury question in negligence).
\textsuperscript{20} See infra notes 25–76 and accompanying text.
\textsuperscript{21} See infra notes 77–97 and accompanying text.
\textsuperscript{22} See infra notes 98–324 and accompanying text.
\textsuperscript{23} See infra notes 271–305 and accompanying text.
\textsuperscript{24} See infra notes 325–26 and accompanying text.
I. THE DOCTRINAL ROLES OF FORESEEABILITY

As background to the discussion in Part III of the conceptual work accomplished by foreseeability, the following presents a descriptive overview of foreseeability’s current doctrinal roles.25

A. Foreseeability’s Role in Breach

Perhaps the most straightforward doctrinal function of foreseeability in negligence law is to aid the factfinder in determining breach. Once the judge has determined that the defendant owed a duty and has delimited that duty in a standard of care, the jury must then decide,26 in the context of breach, whether the defendant’s conduct failed to conform to that standard.27 The near-universal standard of care in negligence cases is the duty to act as would have a reasonable person under the circumstances.28 Reasonableness often turns on (1) the degree of foreseeable likelihood that the defendant’s actions might result in injury,29 (2) the range in severity of foreseeable injuries, and (3) the benefits and burdens of available precautions or alternative manners of conduct.30 Together, the range of likelihood and severity of foreseeable injury constitutes the foreseeable “risk” created by an actor’s conduct.31 As a general matter, the higher the risk—that is, the more foreseeable it was that injury might result from particular behavior and the more severe the range of foreseeable injuries—the more careful the defendant is required to have been.32

25 For a more detailed account, see Cardi, supra note 7, at 743–67.
26 Dobbs, supra note 2, § 115, at 270.
28 RESTATEMENT (SECOND) OF TORTS § 283 (1965).
30 See, e.g., United States v. Carroll Towing Co., 159 F.2d 169, 173–74 (2d Cir. 1947) (enshrining these factors in the mathematical formula in which liability lies where B (burden of precautions) < P (probability of loss) x L (magnitude of loss)); Dobbs, supra note 2, §§ 143–146, at 334–48 (explaining the interplay of foreseeability and reasonableness); Prosser & Keeton, supra note 27, § 65, at 453–54 (“The unreasonableness of the risk[,] which [a reasonable person of ordinary prudence] incurs is judged by the . . . process of weighing the importance of the interest he is seeking to advance, and the burden of taking precautions, against the probability and probable gravity of the anticipated harm . . . .”).
31 See, e.g., Zettle v. Handy Mfg. Co., 998 F.2d 358, 360 (6th Cir. 1993) (“[A] showing of the magnitude of the foreseeable risks includ[es] the likelihood of occurrence of the type of accident . . . and the severity of injuries sustainable from such an accident.”).
32 See, e.g., Lollar v. Poe, 622 So. 2d 902, 905 (Ala. 1993) (“The degree of care required of an animal owner should be commensurate with the propensities of the particular ani-
This form of foreseeability is one of general focus. That is, it examines not the foreseeability of the particular injury suffered by the plaintiff, but the foreseeable likelihood and severity of injuries that might have resulted from the defendant’s conduct.\(^{33}\)

**B. Foreseeability’s Role in Proximate Cause**

Once it has been determined that a defendant owed and breached a duty, and that the breach in fact caused the plaintiff’s injury, the jury\(^{34}\) must decide what is known as “proximate cause,”\(^{35}\) “legal cause,”\(^{36}\) or, as in the proposed *Restatement (Third)*, the “scope of liability.”\(^{37}\) This element serves as a limitation on the consequences of an actor’s conduct.\(^{38}\) Through proximate cause, courts recognize that although “the consequences of an act go forward to eternity, . . . . any attempt to impose responsibility upon such a basis would result in infinite liability for all wrongful acts, and would ‘set society on edge and fill the courts with endless litigation.’”\(^{39}\) Proximate cause thus focuses on the connection between a defendant’s unreasonable conduct and the plaintiff’s injury,\(^{40}\) and limits liability after the point at which “the harm that resulted from the defendant’s negligence is so clearly

---

\(^{33}\) See *Dobbs*, *supra* note 2, § 143, at 335. In the context of a discussion of breach, Dobbs offers the following example:

So if a speeding driver crashes into your living room, the fact that a reasonable person would not have specifically recognized a risk of harm to living room furniture will not assist the driver to avoid liability. It is one of the cluster of harms in a generally foreseeable category, and that is enough.

*Id.*

\(^{34}\) *Prosser & Keeton*, *supra* note 27, § 45, at 321.

\(^{35}\) *Id.* § 41, at 263.

\(^{36}\) *Restatement (Second) of Torts* § 281(c) (1965).

\(^{37}\) *Restatement (Third) of Torts: Liab. for Physical Harm (Basic Principles)* § 29, Special Note on Proximate Cause, at 1 (Tentative Draft No. 3, 2003).

\(^{38}\) *Prosser & Keeton*, *supra* note 27, § 41 at 264.

\(^{39}\) *Id.* § 41, at 264 (quoting North v. Johnson, 59 N.W. 1012, 1012 (Minn. 1894)).

\(^{40}\) *Id.*; see also Galligan, *supra* note 11, at 1513 (explaining that proximate cause is “really a way of deciding whether society ought to hold this defendant, whose negligent acts were a cause-in-fact of the plaintiff’s damages, liable under these circumstances, to this plaintiff . . . [or to] sever the chain of causation”).
outside the risks created that it would be unjust or at least impractical to impose liability."

Explanations and tests for proximate cause abound. A common thread among proximate cause cases, however, is that most explicitly or implicitly consider some notion of foreseeability. In contrast to foreseeability’s role in breach, the foreseeability inquiry in the context of proximate cause is not general but specific to the particular injury suffered by the particular plaintiff at hand. Thus, even where injury of some kind to some person was foreseeable, proximate cause may fail where the defendant’s actions resulted in (1) an unforeseeable type of injury, (2) injury occurring in an unforeseeable manner, or (3) injury to an unforeseeable plaintiff. Furthermore, foreseeability in the context of proximate cause does not help to decide whether the defendant acted unreasonably, as in the context of breach; rather, foreseeability here aids in the decision of whether the actual consequences of the defendant’s conduct were so bizarre or far-removed from the risks that made the conduct negligent that the defendant, though blameworthy, should not be held liable for them.

41 Dobbs, supra note 2, § 180, at 443.
42 See, e.g., Tetro v. Town of Stratford, 458 A.2d 5, 7–8 (Conn. 1983) (“The test for finding proximate cause ‘is whether the harm which occurred was of the same general nature as the foreseeable risk created by the defendant’s negligence.’” (quoting Coburn v. Lenox Homes, Inc., 441 A.2d 620, 627 (Conn. 1982))).
43 For example, the approach of the draft Restatement (Third) of Torts might be described as little more than a “foreseeability of type of harm” standard. Restatement (Third) of Torts: Liab. for Physical Harm (Basic Principles) § 29, cmt. j (Tentative Draft No. 3, 2003). But see John C. P. Goldberg, Rethinking Injury and Proximate Cause, 40 San Diego L. Rev. 1315, 1332–43 (2003) (suggesting an alternative view of the Restatement (Third) “risk rule”).
44 Dobbs, supra note 2, § 181, at 444; Franklin & Rabin, supra note 2, at 399.
45 See, e.g., Baltimore City Passenger Ry. Co. v. Kemp, 61 Md. 74, 79–80 (1883) (holding that it is within the purview of the jury to decide whether it was foreseeable that a speeding driver would hit another car, that the collision would bruise its driver, and that the bruise would later become cancerous); Hines v. Morrow, 236 S.W. 183, 187–88 (Tex. Civ. App. 1921) (holding that it was foreseeable, as a matter of law, that a pothole left by defendant in a highway would stall a car, that a good Samaritan attempting to pull it out would get his wooden leg stuck in the mud, and that a loop in the tow rope would lasso his good leg and break it).
46 See, e.g., Bunting v. Hogsett, 21 A. 31, 32 (Pa. 1891) (holding that the injury of a railroad passenger was foreseeable, as a matter of law, where a collision threw a railroad engine out of control and the engine then ran around a circular track and struck the passenger in a second collision).
47 See, e.g., In re Guardian Cas. Co., 2 N.Y.S.2d 232, 234 (N.Y. App. Div. 1938) (holding the plaintiff was foreseeable, as a matter of law, where a collision forced a taxi into a building and loosened a stone, which fell and killed the plaintiff, a bystander, while the taxi was being removed twenty minutes after the initial accident).
In contrast to breach and proximate cause, duty is the province of the court. Courts’ analysis of duty is a two-step process. First, the court must decide whether to impose a duty on the defendant at all. Second, the court must define the scope of that duty in the form of a standard of care.

Most courts adhere to a general structure for duty in negligence cases. The foundation of this structure is the principle that one generally owes a duty to avoid affirmatively causing physical harm to others. The flip side of this universal duty is that one generally does not owe a duty to warn, protect, or rescue a person from risks created by another source. There are, however, a number of commonly held exceptions to this “no duty to rescue” rule. These “affirmative duties” include, for example, the duty to rescue persons with whom one has a judicially recognized special relationship and the duty to continue, under some circumstances, a rescue effort voluntarily undertaken. Finally, courts have carved out a number of other special duty rules that turn either on the nature of the injury alleged or on certain characteristics of the defendant’s identity. For example, a defendant

---

48 Restatement (Second) of Torts § 328B (1965); Dobbs, supra note 2, § 149, at 355.
49 Id.; Dobbs, supra note 2, § 149, at 355.
50 See, e.g., Weirum v. RKO Gen., Inc., 539 P.2d 36, 39 (Cal. 1975) (“[E]very case is governed by the rule of general application that all persons are required to use ordinary care to prevent others from being injured as the result of their conduct.”); Heaven v. Pender, (1883) 11 Q.B.D. 503, 509 (“[W]hen one person is by circumstances placed in such a position with regard to another that every one of ordinary sense who did think would at once recognize that if he did not use ordinary care and skill in his own conduct with regard to those circumstances he would cause danger of injury to the person or property of the other, a duty arises to use ordinary care and skill to avoid such danger.”); Dobbs, supra note 2, § 227, at 578; 3 Harper et al., The Law of Torts, supra note 1, § 18.2; Prosser & Keeton, supra note 27, § 53, at 356–59; see also Restatement (Third) of Torts: Liab. for Physical Harm (Basic Principles) § 37, Reporters’ Note to cmt. b (Tentative Draft No. 4, 2004) (citing a string of cases that recognize the general duty not to create a risk of harm).
52 Restatement (Second) of Torts § 314 (1965); Dobbs, supra note 2, § 314, at 853.
53 See, e.g., Methola v. County of Eddy, 629 P.2d 350, 353–54 (N.M. Ct. App. 1981) (holding that jailors have a duty to protect and rescue inmates from other abusive inmates); Restatement (Second) of Torts § 314A (explaining that common carriers, innkeepers, land possessors, and those in custodial roles owe to customers, invitees, or those in their custody a duty to protect, warn of dangers, and offer first aid).
54 Generally, a voluntary rescuer must use reasonable care to continue a rescue effort if failure to do so would leave the rescuee in imminent peril of serious bodily harm. E.g., Atkinson v. Stateline Hotel Casino & Resort, 21 P.3d 667, 672 (Utah Ct. App. 2001).
generally (though with significant exceptions) does not owe a duty to avoid causing purely emotional or economic harm, and where the defendant is a branch of government or a landowner, limited-duty rules may attach.

The conceptual substance of duty is a matter of considerable debate. Generally, duty is seen to be “an obligation, to which the law will give recognition and effect, to conform to some standard of conduct toward another.” Regarding the method by which courts impose and define legal obligations, their analysis appears to focus on five major considerations: community notions of obligation, a broad sense of social policy (including a court’s judgments regarding substantive rights and critical morality), concern for the rule of law, administrative capability and convenience, and, not least of all, foreseeability.

55 See, e.g., Dillon v. Legg, 441 P.2d 912, 916–20 (Cal. 1968) (analyzing emotional harm rule generally); Dobbs, supra note 2, § 452, at 1282 (discussing the general rule that there is no duty for negligent economic harm).

56 See Dobbs, supra note 2, §§ 231–237, at 587–620 (discussing duties commonly owed by landowners to those on their land); id. § 271, at 723–27 (describing the public duty doctrine and its limitations on governmental liability).


58 See, e.g., Davis v. Westwood Group, 652 N.E.2d 567, 569 (Mass. 1995) (“In determining whether the defendant had a duty to be careful, we look to existing social values and customs, as well as to appropriate social policy.”); John C.P. Goldberg, Note, Community and the Common Law Judge: Reconstructing Cardozo’s Theoretical Writings, 65 N.Y.U. L. Rev. 1324, 1334–35 (1990) (discussing tort law’s incorporation of social norms and expectations); Prosser, supra note 57, at 15 (“In the end the court will decide whether there is a duty on the basis of the mores of the community . . . .”).

59 See Prosser & Keeton, supra note 27, § 33, at 358 (stating that duty is “only an expression . . . of policy which lead[s] the law to say that the plaintiff is entitled to protection”); Oliver W. Holmes, Jr., The Path of the Law, 10 Harv. L. Rev. 457, 466–68 (1897) (discussing the idea generally).

60 Because duty is the only prima facie element in negligence decided by the court, it provides judges their primary means of ensuring that like cases are decided alike and different cases differently. Because determinations of negligence are overwhelmingly fact-specific, however, duty rules most often set general standards rather than particularized codes of conduct. See Dobbs, supra note 2, § 227, at 579 (describing duty decisions as “expressions of ‘global’ policy rather than evaluations of specific facts of the case” and explaining that “no-duty rules should be invoked only when all cases they cover fall substantially within the policy that frees the defendant of liability . . . . [R]ules of law having the quality of generality . . . . should not be merely masks for decisions in particular cases.”).

61 See Prosser, supra note 57, at 15 (“In the decision whether or not there is a duty, many factors interplay . . . [including] the convenience of administration of the rule . . . .”). For example, the rules governing claims for emotional distress owe their origin to such concerns. See Galligan, supra note 11, at 1511 (noting that in the context of claims for emotional distress, “a court may decide there is no duty owed . . . . [for] administrative convenience”). Claims for emotional distress pose several unique administrative challenges: scientific and legal shortcomings in the ability to determine accurately the existence and extent of emotional harm, the problem of approximating emotional harm in
ity. Often, foreseeability is cited as a reason to impose a duty where one otherwise would not exist—for example, due to the rescue rule.\textsuperscript{62} Courts also sometimes cite a lack of foreseeability as grounds for denying a duty, even where the defendant’s conduct created a risk of harm.\textsuperscript{63} Indeed, foreseeability has become so central a concept in many courts’ duty analyses that a ruling on foreseeability is outcome-determinative.\textsuperscript{64}

In ruling on questions of duty, courts employ each of the doctrinal forms of foreseeability also used in deciding breach and proximate cause.\textsuperscript{65} Imposition of a duty often turns upon a court’s view of the foreseeability of some risk attendant to the defendant’s conduct.\textsuperscript{66}

dollar awards, and the threat of a floodgate of litigation over claims of trivial emotional injury. See Dobbs, supra note 2, § 302, at 823–24. In response to these challenges, courts have imposed only limited duties on defendants to avoid causing purely emotional harm. See Metro-N. Commuter R.R. Co. v. Buckley, 521 U.S. 424, 429–36 (1997) (outlining the common law’s limited duties to avoid causing emotional distress and explaining the policy reasons for such limitations).

\textsuperscript{62} See, e.g., Tarasoff v. Regents of the Univ. of Calif., 551 P.2d 334, 342 (Cal. 1976) (imposing duty on mental health professionals to warn foreseeable third parties of dangers posed by their patients and noting that among several factors to be considered in analyzing duty, “[t]he most important of these considerations . . . is foreseeability”); Murdock v. City of Keene, 623 A.2d 755, 757 (N.H. 1993) (explaining that a jailer may be liable for injuries sustained from an inmate’s suicide attempt if the attempt was foreseeable).

\textsuperscript{63} See, e.g., Herrera v. Quality Pontiac, 73 P.3d 181, 187 (N.M. 2003) (“For our duty analysis, ‘it must be determined that the injured party was a foreseeable plaintiff—that he [or she] was within the zone of danger created by [the defendant’s] actions . . . .’”) (alteration in original) (quoting Calkins v. Cox Estates, 792 P.2d 36, 38 (N.M. 1990)).

\textsuperscript{64} See, e.g., Harper v. Remington Arms Co., 280 N.Y.S. 862, 868–69 (Sup. Ct. 1935) (finding that a gun manufacturer had no duty and thus was not liable for injuries resulting from the use of highly charged ammunition, meant only for use in firearms testing and distributed only to testing parties, because the injured party was an unforeseeable user who was given the shells by a third party).

\textsuperscript{65} For a description of the roles of foreseeability in breach and proximate cause, see supra notes 26–47 and accompanying text.

\textsuperscript{66} See, e.g., Remsburg v. Docusearch, Inc., 816 A.2d 1001, 1006 (N.H. 2003) (“All persons have a duty to exercise reasonable care not to subject others to an unreasonable risk of harm. Whether a defendant’s conduct creates a risk of harm to others sufficiently foreseeable to charge the defendant with a duty to avoid such conduct is a question of law . . . .”) (citations omitted); Brennen v. City of Eugene, 591 P.2d 719, 723 (Or. 1979) (stating that a defendant owes a duty where the defendant “creat[ed] a foreseeable risk of harm to others”); Greater Houston Transp. Co. v. Phillips, 801 S.W.2d 523, 525 (Tex. 1990) (“In determining whether the defendant was under a duty, the court will consider several interrelated factors, including the risk, foreseeability, and likelihood of injury weighed against the social utility of the actor’s conduct, the magnitude of the burden of guarding against the injury, and the consequences of placing the burden on the defendant. Of all these factors, foreseeability of the risk is ‘the foremost and dominant consideration.’”) (citation omitted) (quoting El Chico Corp. v. Poole, 732 S.W.2d 306, 311 (Tex. 1987)); Galligan, supra note 11, at 1511 (noting that some courts hold that “a person has a duty to
Where injury was not a sufficiently foreseeable consequence of the defendant’s conduct, or where the foreseeable severity of injury was not particularly great, the judge will dismiss the case on the ground that the defendant did not owe a duty of care.\textsuperscript{67} The converse is also true.\textsuperscript{68} Courts also frequently condition duty on the foreseeability of the type of harm or the manner in which harm occurred. That is, even where a court finds that the defendant’s conduct created some risk of harm, the court will decline to impose a duty where the type or manner of harm was unforeseeable.\textsuperscript{69} Finally, as demonstrated by the famous case of \textit{Palsgraf v. Long Island Railroad Co.},\textsuperscript{70} courts also consider the foreseeability of a particular plaintiff to be an important factor in deciding whether to impose a duty.\textsuperscript{71}

One last introductory observation regarding foreseeability’s doctrinal place in courts’ duty determinations: some courts and academics explain foreseeability’s seemingly redundant roles in duty by urging that courts decide foreseeability in the context of duty categorically, whereas they decide foreseeability as part of breach or proximate cause strictly according to the particular facts of the case. Pursuant to this view, for example, plaintiff-foreseeability in the duty context involves asking whether the class or category of persons of

\begin{footnotesize}
\textsuperscript{67} See, \textit{e.g.}, Albert v. Hsu, 602 So. 2d 895, 898 (Ala. 1992) (holding that the defendant restaurant owner did not owe a duty to a restaurant patron to protect the patron from the unforeseeable event of a car backing across a parking lot, over a curb, and through the wall of the restaurant); Washington v. City of Chi., 720 N.E.2d 1030, 1033–34 (Ill. 1999) (holding that where the plaintiff was struck by a truck whose driver had decided to skirt traffic by driving onto the shoulder, striking a median and planter box installed by the defendant city, the city had no duty because “the accident . . . was not a reasonably foreseeable consequence of the condition of the median”).

\textsuperscript{68} See, \textit{e.g.}, Snyder v. Am. Ass’n of Blood Banks, 676 A.2d 1036, 1048–49 (N.J. 1996) (holding the severity of harm caused by AIDS to be dire, the possibility that patients might contract AIDS via contaminated transfusions to be foreseeable, and that the defendant blood bank thus owed a duty to use reasonable precautions to avoid causing such infections).

\textsuperscript{69} See, \textit{e.g.}, Bryant v. Glastetter, 38 Cal Rptr. 2d 291, 295 (Ct. App. 1995) (declining to impose a duty—where the decedent tow truck driver was killed by a third party while impounding the defendant’s vehicle after the defendant’s arrest for drunk driving—because “[t]he harm suffered by decedent . . . was not a ‘harm of a kind normally to be expected’ as a consequence of negligent driving”) (quoting George A. Hormel & Co. v. Maez, 155 Cal. Rptr. 337, 339 (Ct. App. 1979)).

\textsuperscript{70} 162 N.E. 99 (N.Y. 1928).

\textsuperscript{71} Id. (holding that although railroad workers created some risk in dislodging a package from the grip of a boarding passenger, they did not create a risk to the plaintiff, who stood some distance away on the station platform).
\end{footnotesize}
which plaintiff is a member was foreseeable to the category of persons of which defendant is a member. To the extent that foreseeability in duty is indeed decided as a categorical matter—and it is certainly not uniformly so decided—such decisions are normatively unwise. As I have elsewhere argued in greater detail, foreseeability of any doctrinal stripe is a particularly fact-dependent question, the answer to which might turn on even slight differences in the facts of a case. Deciding foreseeability as a categorical matter—for a category of defendants, a category of injuries, or in the context of a category of risk-creating activities—is thus nothing more than the announcement of a broad rule that may not make sense under any number of unforeseen future circumstances. A categorical decision of foreseeability thus suffers from the same faults as dictum. There is, however, a different sense in which a form of categorical foreseeability might provide insight into certain foreseeability cases. This conceptual use of categorical foreseeability, and categorical foreseeability in general, is discussed at length in Part III.D below.

II. The Purpose of Elements of a Cause of Action

The foregoing explanation highlights the similar, perhaps overlapping doctrinal appearances of foreseeability in the primary elements of negligence. I have elsewhere argued that the law’s redundant use of foreseeability is confusing, illogical, and normatively undesirable. Regardless of one’s opinion regarding the proper doctrinal place of foreseeability, however, it is helpful to consider why the question of to which element of negligence foreseeability is best suited is important. More generally, why does it matter that a particular legal concept is associated with one element of a cause of action or another? Indeed, what is the purpose of dividing a cause of action into elements at all?

72 See Goldberg & Zipursky, supra note 18, at 1818–20, 1828.
73 See generally Mussivand v. David, 544 N.E.2d 265 (Ohio 1989) (in deciding whether an adulterer owed a duty to the spouse of his adulteress not to transmit to him a sexually transmitted disease, the court did not speak in terms of a generalized, class-based foreseeability, but imposed a duty of due care in light of specific foreseeability-related facts).
74 Cardi, supra note 7, at 792–93, 801–03.
75 Id. at 802.
76 See infra notes 271–305 and accompanying text.
77 See generally Cardi, supra note 7.
The cause of action is a descendent of the writ system, imported to colonial America from England. Under the writ system, in order to survive immediate dismissal, a plaintiff had to allege facts that fit one of a limited number of particularized forms of proceeding at law or in equity. Each form of proceeding limited the scope of relief it provided by requiring distinct procedural incidents, limiting its application to a particular type of factual scenario, offering only a specific form of remedy, and setting the means of judgment and execution. If a plaintiff’s pleading failed in any one of these departments, it was dismissed without regard to the potential that the plaintiff had indeed been wronged and had suffered injury as a result.

Division of the early forms of proceeding into elements provided a formal means by which to test whether the plaintiff had properly satisfied the requirements for access to the courts and to a particular remedy. The cause of action as we know it today is analogous to its ancestor in the sense that it sets forth the legal requirements according to which a plaintiff’s suit is judged. It is different, however, in several important respects. Today’s causes of action are of broader applicability than were the early forms of proceeding—the general negligence action, for example, has replaced more specific writs such as trespass on the case. In addition, the chosen cause of action no longer dictates the remedy available. Indeed, a plaintiff who has been wronged is no longer limited to existing causes of action for re-


79 Bellia, supra note 78, at 784–85.

80 See id. at 784, 789.

81 See 1 Joseph Chitty, A Treatise on the Parties to Actions, the Forms of Action, and on Pleading 572 (1828) (“When the declaration . . . appears on the face of it, and without reference to extrinsic matter to be defective, either in substance or form, the opposite party may in general demur . . .”); Joseph Story, A Selection of Pleadings in Civil Actions 23 (2d ed. 1829) (noting that “[a] mistake of the proper form of action” or “a defect of form in the count or declaration” was grounds for dismissal); Bellia, supra note 76, at 789.

82 See Horwitz, supra note 78, at 89–94 (describing the movement from trespass and trespass on the case to general negligence).

83 See Bellia, supra note 78, at 794–98 (detailing the 1930s debate, culminating in the drafting of the Federal Rules of Civil Procedure, which divorced the concept of a cause of action from the existence of a remedy).
lief at all. At least ostensibly, a plaintiff need only convince a court that she has been wronged, and the court is empowered to grant relief. The cause of action is, in this sense, no longer formalistic: a court is not required to dismiss a plaintiff’s claim solely for having failed to plead the elements of a cause of action. Lastly, and most saliently, the elements of a cause of action focus no longer on formalistic hoops through which a plaintiff must leap, but instead describe the conceptual elements of proof of which the plaintiff must convince the judge or jury in order to prevail.

Courts have thus relaxed the formalistic aspects of the forms of proceeding, adopting instead an approach that focuses directly on the plaintiff’s rights and the defendant’s putative wrong. The modern division of a cause of action into a series of elements serves this approach in at least three important ways. First, a list of elements guides the liability inquiry by providing a structured, ordered method according to which a court or jury may reason through the substantive requirements. Because each court follows the same pattern of elements, such blueprints serve the rule of law by helping to ensure that like cases will be decided alike.

84 See id. at 799 (“[I]t seems clear that today courts conceive of the concept of the cause of action in more functionally adaptive terms than they did before the merger of law and equity: a plaintiff may be said to have a cause of action notwithstanding the fact that it is uncertain upon what legal grounds the plaintiff ultimately will prevail if successful; . . . and the plaintiff may prevail upon a legal theory without either pleading or proving all of its elements.”).

85 See id. at 792 (“The Federal Rules of Civil Procedure do not require that a complaint set forth in a particular form the facts and events necessary to attain a certain form of relief.”); Patrick Kelley, Infancy, Insanity, and Infirmity in the Law of Torts, 48 Am. J. Juris. 179, 182 (2003) (“At first, the legal question raised was whether, on these real facts, the case was properly brought under the pleaded form of action. Inexorably, however, these procedures invited the litigants to ask the substantive question: on these real facts, should the defendant be held liable?”).

86 Although many cases are still dismissed for failure to allege facts sufficient to state a claim upon which relief may be granted, courts’ focus now is not on searching through a pile of writs for a proper fit, but on determining whether the plaintiffs’ facts implicate the concepts on which substantive justice turns.

87 See John Norton Pomeroy, Code Remedies: Remedies and Remedial Rights § 347 (4th ed. 1904) (defining a cause of action as “the facts from which the plaintiff’s primary right and the defendant’s corresponding primary duty have arisen, together with the facts which constitute the defendant’s delict or act of wrong”).

Second, division of a cause of action into elements facilitates decisionmakers’ analysis of the legal concepts upon which liability rests. As explained above, the elements of a cause of action typically represent the legal concepts of which the plaintiff must convince the judge or jury in order to obtain relief. By considering such concepts distinctly and in ordered sequence, courts more capably identify the characteristics of each legal concept, isolate the analysis of each concept from that of others, and determine the relationship between concepts. These acts, in turn, coalesce courts’ understanding of the prerequisites to liability and lead to a more correct and consistent application thereof. For example, courts’ understanding that the element of duty is distinct from and antecedent to the element of breach allows courts to parse more carefully those cases in which a defendant unreasonably failed to rescue another from a risk created by some third source.\(^8^9\) As another example, consideration of the relationship between the concepts of breach and factual causation is a necessary step in the proper resolution of many toxic tort cases.\(^9^0\)

Third, setting forth a theory of recovery as a series of elements provides a useful means by which to segregate issues to be decided by the court from those left in the first instance to the jury. It is helpful to consider the proper doctrinal placement of foreseeability in light of the foregoing discussion. What is the proper home (or homes) for foreseeability in light of the first purpose, the provision of a structured analytical blueprint? At one level, the need for an ordered inquiry would seem to be met by virtually any logical placement of foreseeability. So long as courts have a uniform schematic, they will “build the same house.” Of course, one might argue that the existing multifaceted, potentially redundant doctrinal role of foreseeability—even if ultimately logical—proves confusing to courts and therefore cannot cultivate the desired methodological consistency.\(^9^1\) On the other

\(^8^9\) That is, separate consideration of the concepts of duty and breach ensures in such cases that the unreasonableness of the defendant’s actions does not necessarily lead to an imposition of legal blame. See, e.g., Randi W. v. Muroc Joint Unified Sch. Dist., 929 P.2d 582, 587–93 (Cal. 1997) (deciding first whether a duty existed pursuant to misfeasance or nonfeasance before turning to the issue of reasonableness).

\(^9^0\) See, e.g., Zuchowicz v. United States, 140 F.3d 381, 390 (2d Cir. 1998) (explaining that it is not enough that the plaintiff prove that the drug called Danocrine caused the plaintiff’s injury, but that the plaintiff also must prove that it was the defendant’s excessive—and hence, wrongful—prescription of Danocrine that caused the injury).

\(^9^1\) See Cardi, supra note 7, at 790–94.
hand, one might counter that courts’ confusion about foreseeability need not be linked necessarily to structural defects—it may be that although courts have not yet come around to a consistent understanding of foreseeability’s structure, they will eventually. Although I do not concede the point, this Article proceeds on the assumption of a neutral resolution of this first purpose.

Regarding the second purpose, in which element or elements of negligence would foreseeability best aid courts’ analysis of the legal concepts on which negligence liability rests? This question is the subject of Part III and the primary focus of this Article. Part III demonstrates that although a compelling case can be made that certain conceptual roles of foreseeability fit nicely within the context of duty, these same concepts fit equally well as part of breach and/or proximate cause.92 Ultimately, this Article asserts that this second purpose is therefore indeterminate as a means of deciding foreseeability’s proper place in negligence doctrine.

Insofar as these are the most important considerations, if the first and second purposes are both indeterminate, resolution of foreseeability’s proper home must lie with the third purpose. The debate should therefore be resolved by candid discussion as to whether each of foreseeability’s various incarnations would best be decided, in the first instance, by the court or by the jury. I have elsewhere argued in favor of the latter,93 and portions of Part III below supplement this contention.

Before moving to a discussion of foreseeability’s conceptual roles in negligence, I wish to make one aspect of this inquiry transparent. Some scholars and courts speak of duty, breach, and proximate cause as if they exist as a matter of natural law, as if they hang “in the ether” awaiting our discovery.94 Those who think about the elements of negligence from such a vantage often place considerable importance on the elements’ historical development, a development that reflects our gradual honing-in on the concepts’ “true nature.”95 This natural law-

92 See infra notes 98–324 and accompanying text.
93 See Cardi, supra note 7, at 794–804.
94 See generally Stephen R. Perry, The Moral Foundations of Tort Law, 77 Iowa L. Rev. 449, 477–78 (1992) [hereinafter Perry, Moral Foundations] [suggesting that Ernest Weinrib views tort law as “having an essence” or “intrinsic ordering,” versus an understanding of tort law as a “conventional ordering, the constituent elements of which reflect ‘the contingencies of social practice and linguistic usage’”) (quoting Ernest J. Weinrib, Understanding Tort Law, 23 Val. U. L. Rev. 485, 496 (1989)).
95 See generally H.L.A. Hart, The Concept of Law 182 (1961) (describing as a core value of natural law that “there are certain principles of human conduct, awaiting discov-
type view of negligence seems also to lead some to wish for robust conceptual meaning on the constituent elements of negligence simply for its own sake. The bias of this Article is rather more pragmatic. The following analysis works from a presumption that the conceptual building blocks of negligence spring from the minds of lawyers, judges, legislators, and academics. They exist simply as a means of embodying and perhaps formalizing courts’ normative choices regarding who should be liable to whom and under what circumstances.96

This approach does not deny that tort law has, or is capable of having, a coherent structure and meaning; hence, the approach does not embody a form of “reductive instrumentalism.”97 It is instrumental, however, in the sense that it works from the premise that any feasible structure and meaning of tort law, even if coherent, represents a set of broad policy choices. For example, the general rule that one must not act unreasonably to the detriment of another represents a set of policy choices regarding the circumstances under which we are willing to coerce the payment of damages. This approach de-mystifies tort doctrine just enough that we might find a structure and meaning that more accurately reflects the policy choices that we have made or wish to make.

Perhaps for this reason, my argument regarding the proper doctrinal place for foreseeability is not purely descriptive, but largely normative. I concede that the approach of many courts is at odds with my urgings and that the accounts of duty offered by many of my colleagues are descriptively accurate. The exposition below thus offers an alternative approach to duty and to foreseeability.

---

96 See Tarasoff v. Regents of the Univ. of Calif., 551 P.2d 334, 342 (Cal. 1976) ("[B]ear in mind that legal duties are not discoverable facts of nature, but merely conclusory expressions that, in cases of a particular type, liability should be imposed for damage done."); Holmes, supra note 59, at 458 (explaining that “a legal duty so called is nothing but a prediction that if a man does or omits certain things he will be made to suffer in this or that way by judgment of the court;—and so of a legal right").

III. THE CONCEPTUAL ROLES OF FORESEEABILITY

Part I described the doctrinal types of foreseeability as they appear in the elements of the negligence cause of action: foreseeability of risk, foreseeability of type or manner of injury, and foreseeability of plaintiff. Each doctrinal type serves, in its various incarnations, five conceptual purposes. Foreseeability serves as a constituent of moral responsibility and as an instrument of behavioral modification and economic efficiency. Foreseeability also provides a means of limiting the range of consequences for which a defendant will be held liable and of gauging the merit and extent of court action. Finally, foreseeability often operates as a proxy for decisions of policy that have little to do with foreseeability’s other conceptual purposes. The following subsections examine foreseeability’s various conceptual roles in negligence in an attempt to reveal that foreseeability might serve each of these purposes equally well, whether its doctrinal vessel is duty or the elements of breach or proximate cause.

As an initial matter, it is helpful to consider foreseeability’s conceptual purposes in light of the means by which foreseeability generally must be determined. A brief explanation of this process—drawn primarily from the insightful work of Stephen Perry—follows.

“Reasonable foreseeability,” as the term is commonly used, is a function of two separate effects: (1) the objective probability of an event occurring, and (2) a reasonable person’s knowledge and beliefs about that probability. The objective probability that an event will occur is best understood as the event’s relative frequency within a reference class of events—for example, the relative frequency of a car crashing when one drives thirty miles per hour over the speed limit. Reasonable foreseeability is not simply a reflection of this objective

98 See supra notes 25–76 and accompanying text.
99 See infra notes 115–86 and accompanying text.
100 See infra notes 187–215 and accompanying text.
101 See infra notes 216–70 and accompanying text.
102 See infra notes 271–305 and accompanying text.
103 See infra notes 306–24 and accompanying text.
105 Perry, Responsibility for Outcomes, supra note 104, at 97.
106 Id.
probability, however. Rather, foreseeability measures the fragment of objective probability that a reasonable person could have or should have—depending on the context of the decision—foreseen under the circumstances. Thus, foreseeability is often referred to as epistemic (or knowable) probability.

The determination of foreseeability requires two levels of judgment. First, the relevant decisionmaker must properly describe the subject event and frame that event within its proper reference class of events. Continuing the example from the prior paragraph, an event might be described as “injury,” “injury from a car crash,” “shattered pelvis from a car crash,” or “shattered pelvis from a car crashing into a tree.” The spectrum of possible descriptions, from general to specific, is quite broad. Similarly, the reference class of events might be described as “while driving,” “while speeding,” “while driving thirty miles over the speed limit,” “while driving thirty miles over the speed limit on a rainy day,” or any number of other possible variations. One’s choice of description of the event and of the reference class of events strongly influences the event’s relative frequency (and likely its epistemic probability). For example, the relative frequency of an injury occurring while driving thirty miles over the speed limit on a rainy day is undoubtedly higher than that of suffering a shattered pelvis from crashing one’s car into a tree while driving. Yet the law provides no guide for determining the appropriate breadth of description.

H.L.A. Hart and Tony Honoré have proposed that the indeterminacy associated with such decisions is lessened, across a wide range of cases,

---

107 See id. Whether “could have” or “should have” is the proper condition depends on the context of the foreseeability inquiry. See infra notes 140–45 and accompanying text (distinguishing the two inquiries).

108 Perry, Risk, Harm and Responsibility, supra note 104, at 322.

109 Perry, Responsibility for Outcomes, supra note 104, at 98; see also Weinrib, supra note 94, at 521 (explaining that creating a link between the defendant’s action and the plaintiff’s suffering requires “viewing the plaintiff’s suffering from the standpoint of an appropriately general description of the risk created by the defendant”).

110 Perry, Responsibility for Outcomes, supra note 104, at 98–99. Indeed, as Ernest Weinrib explains:

The most that the courts can accomplish through abstract prescription is point out that foreseeability of ‘the precise concatenation of events’ is irrelevant, while also cautioning against setting up excessively broad tests of liability. The description of the risk can be formulated only case by case in terms of what is plausible in any given fact situation as compared with analogous fact situations.

Weinrib, supra note 94, at 523.
by the “common knowledge of ordinary persons” who make them.\textsuperscript{111} Nonetheless, some level of indeterminacy remains.

According to Hart and Honoré, this same “common knowledge of ordinary persons” also shapes the second judgment necessary to determine foreseeability. Once the decisionmaker frames the relevant event within a reference class of events, it must then gauge the event’s epistemic probability.\textsuperscript{112} Again, in so doing, the decisionmaker is not attempting to uncover the event’s objective probability (although revelation of such a fact might influence the decisionmaker’s thought process), but rather to decide what an ordinary person could have or should have foreseen under the circumstances.\textsuperscript{113} As I explain more thoroughly in Part III.A, this decision is part fact-finding—determining what the ordinary person would foresee under these circumstances—and part philosophical exercise—deciding what level of epistemic probability should open the door to liability.\textsuperscript{114}

This two-part reasoning plays some role in decisions regarding foreseeability regardless of the conceptual purpose of such decisions and no matter their doctrinal context. I now turn to an examination of the five primary conceptual purposes served by foreseeability.

A. Foreseeability as a Constituent of Moral Responsibility

Each element of negligence reflects, in some respect, a dimension of blameworthiness. The element of breach is commonly thought to be the chief instrument in this regard.\textsuperscript{115} Even where a defendant acted unreasonably, however, such dangerous conduct is not deemed blameworthy unless the defendant owed a duty to have acted reasonably.\textsuperscript{116} Furthermore, even where a defendant unreasonably defaulted on an obligation of care, one cannot blame a plaintiff’s injury on the default without proof of a causal connection between the

\textsuperscript{111} Perry, Responsibility for Outcomes, supra note 104, at 100–01 (citing H.L.A. Hart & Tony Honoré, Causation in the Law (2d ed. 1985); Clarence Morris, Duty, Negligence, and Causation, 101 U. Pa. L. Rev. 189 (1952)).

\textsuperscript{112} Perry, Responsibility for Outcomes, supra note 104, at 100.

\textsuperscript{113} See Arthur Ripstein, Equality, Responsibility, and the Law 94 (1999) (“[F]oresight is neither a matter of what is in fact foreseen nor of what could ideally be foreseen. Instead, it is a matter of what a reasonable person would foresee.”).

\textsuperscript{114} See infra notes 155–64 and accompanying text.

\textsuperscript{115} See, e.g., Prosser & Keeton, supra note 27, § 85, at 608 (discussing fault as the basis for compensation); Weinrib, supra note 94, at 517–19 (analyzing the role of fault in negligence).

\textsuperscript{116} Restatement (Second) of Torts § 282 (1965); Prosser & Keeton, supra note 27, § 30, at 164.
two.\textsuperscript{117} Proximate cause might also be described in terms of blame: although a defendant’s unreasonable conduct may have caused the plaintiff’s injury, other factors—the passage of time, the conduct of others, serendipity—might play such a significant part in causing the injury that the defendant’s role seems insignificant and therefore not blameworthy. Alternatively, proximate cause might stand for the proposition that a defendant may be blamed only for those injuries the risk of which made the defendant’s conduct wrongful.\textsuperscript{118}

A series of scholars have sought to explain the imposition of blame in tort law in moral terms and to justify the circumstances under which moral responsibility, in the form of legal liability, arises.\textsuperscript{119} The most prominent form of this pursuit, known as corrective justice theory, views tort liability solely as a matter of justice between the parties to an action, without regard to law’s capacity as a distributive force, maximizer of societal wealth, or other social instrument.\textsuperscript{120} Although corrective justice encompasses a variety of approaches, recent scholarship on fault-based tort liability has focused largely on the theory of outcome-responsibility first proposed by Tony Honoré.\textsuperscript{121}

Generally, outcome-responsibility proposes a moral link between conduct and the injury caused by it. It describes a set of circumstances under which a defendant can be said to be morally responsible for the consequences of his or her conduct.\textsuperscript{122} According to Honoré, this link

\textsuperscript{117} See Ernest J. Weinrib, The Idea of Private Law 56–144 (1995) (explaining that if a defendant’s conduct results in harm to another, the defendant owes a moral duty to compensate the plaintiff for such harm). See generally Richard Epstein, A Theory of Strict Liability, 2 J. Legal Stud. 151 (1973) (proposing the same as a unitary theory of tort liability).

\textsuperscript{118} See Restatement (Third) of Torts: Liab. for Physical Harm (Basic Principles) § 29 (Tentative Draft No. 3, 2003) (“An actor is not liable for harm different from the harms whose risks made the actor’s conduct tortious.”); Ripstein, supra note 113, at 69 (“Those who fail to exercise appropriate care with respect to particular risks act at their peril with respect to those risks.”).

\textsuperscript{119} Aristotle is the earliest western philosopher associated with this pursuit. See James Gordley, Tort Law in the Aristotelian Tradition, in Philosophical Foundations of Tort Law, supra note 104, at 131. See generally 5 Aristotle, The Nicomachean Ethics 142–49 (J.E.C. Welldon trans., 1987).

\textsuperscript{120} Stephen R. Perry, Tort Law, in A Companion to Philosophy of Law and Legal Theory 57, 72 (Dennis Patterson ed., 1996) [hereinafter Perry, Tort Law].

\textsuperscript{121} See generally Tony Honoré, Responsibility and Luck, 104 L. Q. Rev. 530 (1988). For a revealing discussion of Honoré’s theory, see Perry, Moral Foundations, supra note 94, at 489–96.

\textsuperscript{122} Perry, Tort Law, supra note 120, at 75; see also Stephen R. Perry, The Distribution Turn: Mischief, Misfortune and Tort Law, 16 Q.L.R. 315, 326–27 (1996) [hereinafter Perry, The Distribution Turn] (arguing that fault without outcome-responsibility does not justify forced compensation, but instead might be satisfied by punishment alone).
is established simply when one person’s act causes harm to another.\textsuperscript{123} His reasoning is as follows. Each of us has a choice whether or not to act in the world,\textsuperscript{124} or at least regarding which of any number of acts we will undertake.\textsuperscript{125} This ability to choose is commonly described as one’s capacity for “agency.”\textsuperscript{126} Every act necessarily entails some risk.\textsuperscript{127} In freely choosing one act over another, an actor makes an implicit gamble that the payoff of that act will outweigh its potential cost.\textsuperscript{128} The outcome of this gamble is thus an expression of the actor’s agency,\textsuperscript{129} and the gamble’s outcome is credited (or discredited) to the agent in a kind of “social ledger.”\textsuperscript{130} This social ledger serves as the moral basis for legal liability.\textsuperscript{131}

As many have pointed out, however, what has come to be known as the libertarian account of outcome-responsibility is wanting in two important respects. First, it fails as a normative matter to distinguish between plaintiff and defendant, each of whose actions necessarily led

\textsuperscript{123} See Perry, \textit{Moral Foundations}, supra note 94, at 499 (“[T]he essential characteristic of outcome-responsibility is the fact of having voluntarily performed an action or actions that causally contributed to the outcome in question.”).

\textsuperscript{124} Perry, \textit{Tort Law}, supra note 120, at 76.

\textsuperscript{125} One might argue that in fact we do not have a choice but to act in the world. Living is acting; even if one chooses to sit unmoving as a Jainist monk, one has chosen that act over other alternatives. See Holmes, supra note 29, at 77 (“A man need not, it is true, do this or that act,—the term act implies a choice,—but he must act somehow.”).

\textsuperscript{126} Perry, \textit{Moral Foundations}, supra note 94, at 489–90.

\textsuperscript{127} Others have also made this point. See, e.g., Prosser & Keeton, supra note 27, § 31, at 170 (“Nearly all human acts, of course, carry some recognizable but remote possibility of harm to another.”); Ernest J. Weinrib, \textit{Right and Advantage in Private Law}, 10 CARDOZO L. REV. 1283, 1305 (1989) (stating that “risk is an unavoidable concomitant of all action”).

\textsuperscript{128} Honoré, \textit{supra} note 121, at 539.

\textsuperscript{129} Ripstein, \textit{supra} note 113, at 98.

\textsuperscript{130} Perry, \textit{Moral Foundations}, supra note 94, at 489; see also Ripstein, \textit{supra} note 113, at 97–98 (“Honoré argues that in a world of risk, every action involves an implicit gamble. How those various gambles turn out is constitutive of a person’s agency. . . . When things go well, the credit redounds to us. When they turn out badly, we are responsible for the bad results.”); Honoré, \textit{supra} note 121, at 539–40 (noting that in deciding to do X instead of Y, “we are choosing to put our money on X and its outcome rather than Y and its outcome” and that although we receive credit if the bet turns out well, “if we botch it, . . . that is chalked up against us”); Perry, \textit{Tort Law}, supra note 120, at 76 (“I have a choice about whether to become active in the world, and if I choose activity over passivity then all subsequent consequences, both good and bad, are appropriately chalked up to my moral ledger and no one else’s.”) (emphasis omitted).

\textsuperscript{131} This conclusion rests, in part, on the normative, libertarian claim that forced redistribution is illegitimate and that it is unfair to hold a person responsible for costs imposed by another. Perry, \textit{Tort Law}, supra note 120, at 76.
to the relevant injury. For example, where A drives into B’s parked car, both A’s act in driving and B’s act in parking the car caused the collision—according to the libertarian view, both A and B are outcome-responsible for the accident. This approach leaves courts without guidance on how to allocate liability between causal elements. Second, even if courts were to find principled grounds by which to allocate liability, the libertarian view ultimately leads to a system of general strict liability, and thus fails to justify our largely fault-based system. For these two reasons, the concepts of action and causation have proven to be necessary, but not sufficient, preconditions for outcome-responsibility.

Stephen Perry has led the way in developing an account of outcome-responsibility that justifies both strict and fault-based liability. According to Perry, outcome-responsibility depends on the notion of control. Only an agent who is in control of his or her actions and, to a certain degree, of the consequences of those actions, may be said to be outcome-responsible. According to Perry, the necessary degree of control over the outcome is defined by whether the outcome was avoidable, and avoidability exists only in the presence of a “general capacity on the part of an agent to foresee an outcome and to take steps to avoid its occurrence.” Foreseeability, or at least a “general capacity to foresee,” thus enters the outcome-responsibility calculus. Perry explains:

132 Perry, Moral Foundations, supra note 94, at 463; see also Ronald H. Coase, The Problem of Social Cost, 3 J.L. & Econ. 1, 1–44 (1960) (arguing that tort injury is not caused by one party alone, but rather is caused by the interaction of the acts of both parties to an action).

133 Id.; see also Epstein, supra note 117, at 157–60 (justifying his proposal for general strict liability on libertarian grounds).

134 An interesting consequence of this reasoning is that tort liability based upon violation of an affirmative duty would seem not to be an instance of outcome-responsibility. See Perry, Responsibility for Outcomes, supra note 104, at 125 n.34 (“[T]he breach of an affirmative duty typically gives rise to non-causal liability and should not, therefore, be regarded as an instance of outcome-responsibility.”). I discuss the import of this realization in Part III.D below.

135 See id. at 82 (“Outcome-responsibility . . . , like most other responsibility concepts, . . . involves a notion of control. . . . Outcome-responsibility assumes that the agent had control of his action of the kind posited by action-responsibility, but it also assumes that he had control over the outcome itself.”); see also Ripstein, supra note 113, at 53 (“The idea of responsibility thus carries with it an idea of responsible agency. In order to be a responsible agent, one must be able both to pursue one’s own ends and to moderate one’s claims in light of the legitimate claims of others.”).

136 See id. at 104, at 73.

137 See id.
Suppose A engages in an activity that results in foreseeable harm to B. Because the harm was foreseeable, A had it within his power to avoid causing it; even if there were no precautions he could have taken to reduce the risk, he could have forgone the activity altogether. He thus had a certain measure of control over the situation, and . . . it seems reasonable to ascribe to him a special responsibility for the outcome that, in general, other persons do not have.\textsuperscript{139}

Outcome-responsibility, according to Perry, thus consists of (1) an act (2) that caused injury (3) committed by a person who had the general capacity to foresee and avoid causing such injury.\textsuperscript{140} Outcome-responsibility alone, however, is an insufficient justification for forcing compensation for an injury, but it serves, rather, as a basis for such an obligation.\textsuperscript{141} Even if a defendant \textit{could} have avoided causing an injury, it does not necessarily follow that the defendant \textit{should} have done so.\textsuperscript{142} For the final link in the normative connection between an agent’s act and the harm it caused, Perry looks to the concept of fault.\textsuperscript{143} Only when an avoidable risk should have been avoided—that is, when the agent’s conduct was faulty—does outcome-responsibility ripen into an obligation to compensate.\textsuperscript{144} Perry suggests that foreseeability plays a role in gauging fault as well, although perhaps foreseeability of a different sort and serving a different purpose than that ascribed to outcome-responsibility. Foreseeability informs outcome-

\textsuperscript{139} Perry, \textit{Tort Law}, supra note 120, at 76–77; see also Ripstein, \textit{supra} note 113, at 95 (noting that according to Perry, “foresight is relevant to liability because an agent is only morally responsible for things that he or she can foresee”).

\textsuperscript{140} See Perry, \textit{Responsibility for Outcomes}, \textit{supra} note 104, at 92 (“The normative power of this conception of outcome-responsibility resides in the idea that the exercise of a person’s positive agency, under circumstances in which a harmful outcome could have been foreseen and avoided, leads us to regard her as the author of the outcome. . . . The agent acted and caused harm under circumstances in which she had a sufficient degree of control to avoid its occurrence, and for that reason she has a special responsibility for the outcome that other persons do not have.”).

\textsuperscript{141} Id. at 73.

\textsuperscript{142} Id. at 91; Perry, \textit{Tort Law}, \textit{supra} note 120, at 77.

\textsuperscript{143} See Perry, \textit{The Distribution Turn}, \textit{supra} note 122, at 334 (“[A] moral obligation to compensate, of a kind that was enforceable in tort law, rests on two main foundations: first, a pre-political moral responsibility for those outcomes of our actions that we have the capacity to foresee and avoid; and second, an objective fault standard, shaped by liberal conceptions of fairness and autonomy, that determines which risks may and which may not be imposed on others.”).

\textsuperscript{144} See Perry, \textit{Tort Law}, \textit{supra} note 120, at 77 (“If . . . a risk was not only foreseeable but should have been avoided, it seems appropriate to conclude that outcome-responsibility takes the form of an obligation to compensate.”).
responsibility by indicating avoidability—if an actor was generally capable of foreseeing an outcome, the actor *could* have avoided it. Foreseeability informs fault by indicating reasonableness—if the epistemic probability of an outcome was high enough, perhaps the actor *should* have avoided it.\textsuperscript{145} I will refer to his latter incarnation of foreseeability as “normative foreseeability.”

Perry is clear that at least as a positive matter, outcome-responsibility is the stuff of duty in negligence, and fault is the essence of breach.\textsuperscript{146} This approach makes some intuitive sense. It seems right to say that if one is responsible for an outcome, one therefore owes a duty to have acted reasonably in bringing it about. Stated in terms of foreseeability, when one is capable of foreseeing and avoiding an outcome, one owes a duty to have acted reasonably in foreseeing and avoiding it. Assuming that Perry’s overall justification for negligence liability is correct, however, whether it *commands* a role for foreseeability in duty is another matter. Might outcome-responsibility still accurately describe and justify negligence liability were courts to purge duty of considerations of foreseeability?

An answer to this question begins with the observation—to which Perry concedes—that as a positive account of duty, outcome-responsibility is incomplete. It does not explain courts’ refusal to impose a duty due to any number of consequentialist policy considerations\textsuperscript{147}—for

\textsuperscript{145} See Perry, *Responsibility for Outcomes*, supra note 104, at 105 (“[I]t is only by referring to . . . a cost-benefit analysis that we can say whether or not a given type of harm is reasonably foreseeable, i.e., whether or not it *should* be foreseen. Thus the notion of reasonable foreseeability inevitably involves more than the general capacity to foresee that I have been describing. It also involves reference to action-guiding norms of the kind that figure in the Learned Hand Test.”); see also Ripstein, *supra* note 113, at 105 (explaining that in the context of judging fault, “consequences are foreseeable if a person showing appropriate regard for the interests of others would have taken them into account”).

\textsuperscript{146} See Perry, *The Distribution Turn*, supra note 122, at 326 (“On its face, however, a fault standard tells us nothing about when, or even whether, misfortune should be shifted from one person to another. What it tells us, rather, is how we should behave towards one another; it tells us to conduct our activities with reasonable care, or to take precautions when B < PL, or whatever.”) (emphasis omitted); Perry, *Responsibility for Outcomes*, supra note 104, at 95 (explaining that outcome-responsibility, and the foreseeability component of that, are questions of duty, for the court to decide); id. at 116 (“[I]n the absence of outcome-responsibility, all we have in fault is the violation of a norm. . . . [B]ut if this is not systematically related to and constrained by a requirement of outcome-responsibility (as embodied in the defendant’s duty of care), then there is nothing to distinguish this particular norm of conduct from any other.”).

\textsuperscript{147} Perry, *Responsibility for Outcomes*, supra note 104, at 95 n.34.
example, a concern for sweeping liability\textsuperscript{148} or the desire to protect accepted social institutions.\textsuperscript{149} Nor does outcome-responsibility justify or explain either the imposition of a duty in cases of nonfeasance\textsuperscript{150} or courts’ limitation of one’s duty to avoid causing emotional or economic injury\textsuperscript{151}—decisions that might turn on concerns other than foreseeability.\textsuperscript{152} The fact that duty turns on non-foreseeability-related factors in some cases suggests the possibility that duty, at its core, turns on non-foreseeability-related factors in all cases.\textsuperscript{153} Such an explanation of duty need not rob outcome-responsibility of its justificatory or even its explanatory power, however. Negligence, in relevant part, might be described as follows: if one owed a duty of care (determined solely, for example, by reference to community notions of obligation, policy considerations, a concern for the rule of law, and administrative convenience),\textsuperscript{154} then one will have breached that duty if (1) one was generally capable of foreseeing the injury and avoiding it, and (2) one should have foreseen and avoided it. Just as it seems right to say that if one is capable of foreseeing and avoiding an outcome, one owes a duty to have acted reasonably in foreseeing and avoiding it (Perry’s approach), it also seems intuitive to say that one who is incapable of foresight generally cannot have acted unreasonably in failing to foresee and avoid causing injury. Neither explanation, at this basic level, is normatively superior. A closer look at foreseeability’s part in outcome-responsibility supports this initial conclusion.

\textsuperscript{148} See, e.g., Strauss v. Belle Realty Co., 482 N.E.2d 34, 38 (N.Y. 1985) (limiting the liability of a public utility for gross negligence in causing a power outage because sweeping liability might lead to the utility’s insolvency).

\textsuperscript{149} See, e.g., Kelly v. Gwinnell, 476 A.2d 1219, 1224 (N.J. 1984) (recognizing that many courts do not impose a duty in cases of social host liability because of the concern that such duties would “interfere with accepted standards of social behavior; [and] . . . intrude on and somewhat diminish the enjoyment, relaxation, and camaraderie that accompany social gatherings at which alcohol is served”); Thompson v. McNeill, 559 N.E.2d 705, 707 (Ohio 1990) (declining to impose a duty where a golfer had been struck in the head by an errant ball because liability “might well stifle the rewards of athletic competition”).

\textsuperscript{150} Perry, Responsibility for Outcomes, supra note 104, at 95 n.34.

\textsuperscript{151} See id.

\textsuperscript{152} That affirmative duty cases turn on factors other than foreseeability is a matter discussed in Part III.D below. See infra notes 295–305 and accompanying text. Recitation of some of the policy considerations that underlie economic harm cases may be found supra note 318, and emotional harm cases may be found infra note 61.

\textsuperscript{153} Most existing theories of duty purport to be descriptive and so include foreseeability in some manner. Although it is not the mandate of this Article to propose a foreseeability-free substantive duty theory, Part IV sketches the beginnings of such an approach.

\textsuperscript{154} See supra notes 58–61 and accompanying text.
Outcome-responsibility is dependent upon an actor’s “general capacity to foresee.” By this, Perry refers at least in part to the subjective characteristics of the actor—for example, whether the actor has an IQ sufficient to facilitate a general minimum capacity to foresee injury, or whether the actor’s age and mental health are similarly sufficient. Were this Perry’s only benchmark for judging capacity, this narrow query perhaps would fit best within the concept of duty.

As a matter of policy and commonly-held notions of obligation, a court might well wish to set a minimum physical and mental capacity for foresight below which an actor cannot be said to owe a duty of reasonable care at all. The prescription of such bright-line rules seems quintessentially the mandate of duty.

Thus understood, however, judgments regarding an actor’s general capacity to foresee would not constitute decisions of “foreseeability” at all, at least not as that term is commonly used in the law of negligence. Inquiries into foreseeability presume an acceptable capacity to foresee and ask whether the relevant risk, injury, or plaintiff was fore-

155 See Perry, Responsibility for Outcomes, supra note 104, at 73.

156 See id. at 103 (“[T]he capacities to foresee and avoid harmful outcomes are appropriately understood as general abilities that the individual who caused a given harm ordinarily succeeds in exercising in other, similar situations. . . . It is in this sense that the avoidability-based conception of outcome-responsibility treats the capacities to foresee and avoid harm as subjective.”). Perry further notes that “a plausible non-consequentialist theory of tort law must suppose that mental disorders serious enough to undermine the capacity to foresee and avoid harm should excuse the defendant from liability.” Id. at 106. Perry also cites, for this proposition, H.L.A. Hart’s capacity/opportunity principle: “[W]hat is crucial is that those whom we punish should have had, when they acted, the normal capacities, physical and mental, for doing what the law requires and abstaining from what it forbids, and a fair opportunity to exercise those capacities.” Id. at 88 (quoting H.L.A. Hart, Punishment and Responsibility: Essays in the Philosophy of Law 152 (1968)).

157 See Perry, The Distribution Turn, supra note 122, at 320 (“If I were not an agent, I would not be the sort of entity that could even owe a duty.”).

158 Of course, a court might also hold that a person of a certain age or mental ability cannot be deemed to have breached a duty of care.

159 In fact, courts have generally declined to alter the duty owed by those of impaired mental capacity. Kelley, supra note 85, at 231; see Restatement (Second) of Torts § 283B (1965) (“Unless the actor is a child, his insanity or other mental deficiency does not relieve the actor from liability for conduct which does not conform to the standard of a reasonable man under like circumstances.”). By contrast, courts often delimit the age at which a person gains meaningful agency; below that age, an actor is held either to owe no duty at all or to owe a duty of care keyed to the reasonable child of the actor’s age. See, e.g., Maksaliunas v. Chi. & W.I.R. Co., 149 N.E. 23, 26 (Ill. 1925) (holding that children under the age of seven are presumed to be incapable of negligence); Cox v. Hugo, 329 P.2d 467, 469 (Wash. 1958) (holding same for children under the age of six); Kelley, supra note 85, at 239–40 (explaining courts’ general unwillingness to impose a negligence duty on children under a certain age).
seeable by a hypothetical reasonable person in the actor’s shoes. The aspect of general capacity described above, by contrast, is divorced from the facts of the case. It does not ask whether the relevant risk, injury, or plaintiff was foreseeable under the circumstances. Nor does it even examine whether a certain category of risk, injury, or plaintiff was foreseeable to a category of actor. Such an analysis instead focuses solely on the physical and mental traits of the actor, on whether the actor possesses some minimum mental ability to foresee the consequences of his or her actions. In this sense, this aspect of the general capacity inquiry is akin to a competency hearing in a criminal case. Thus, even if this aspect of the general capacity to foresee were to fit more comfortably in duty than in breach, such a conclusion would do little to advance analysis of the proper doctrinal home for foreseeability.

Perry’s conception of the “general capacity to foresee” cannot possibly stop at an actor’s subjective characteristics, however. It would be illogical to refer to an actor as “outcome-responsible”—even when the actor enjoys some minimum general capacity for foresight—if not even a person of ordinary capacity could have foreseen the relevant outcome. Perry’s conception of capacity thus includes an objective as well as a subjective component, and it considers not one’s capacity to foresee in the abstract, but one’s capacity to foresee a category of events that fits the facts of the case. Specifically, Perry’s capacity inquiry proceeds as follows. First, the decisionmaker must divine the actor’s subjective capacity to foresee outcomes of the same general sort as the outcome at hand. (The proper description of this subject category, as explained in the introduction to this section, is determined according to the “common knowledge of the ordinary person.”) Second, the decisionmaker must compare the actor’s subjective capacity to foresee outcomes of the same general sort as the outcome at hand. (The proper description of this subject category, as explained in the introduction to this section, is determined according to the “common knowledge of the ordinary person.”)

---

160 Whether outcome-responsibility depends on the foresight of the actor or a hypothetical reasonable person is a matter discussed below. See infra notes 161–64 and accompanying text.

161 See Perry, Responsibility for Outcomes, supra note 104, at 103 (“[C]apacities to foresee and avoid harmful outcomes are appropriately understood as general abilities that the individual who caused a given harm ordinarily succeeds in exercising in other, similar situations.”).

162 See supra notes 109–11 and accompanying text.

163 Perry, Responsibility for Outcomes, supra note 104, at 103–04. A broader quotation might be helpful:
foresee such outcomes meets (or exceeds) this objective minimum standard, the actor’s general capacity to foresee is sufficient to support a conclusion that the actor is outcome-responsible.\textsuperscript{164}

This explanation of the general capacity to foresee brings the concept more squarely in line with courts’ common understanding of foreseeability. Still, Perry’s account does not require that its resolution occur in the context of duty. The closest Perry comes to such an argument is his emphasis on the distinction between foreseeability’s role in outcome-responsibility and its role in reasonableness. There is some aesthetic appeal to the proposition that as a prerequisite to normative foreseeability and breach, one’s general capacity to foresee should be resolved separately, as a part of duty. Aside from aesthetics, however, there is no practical reason that the constituents of outcome-responsibility need remain a unit or that they must remain segregated from other phases of the liability inquiry. The distinction between the general capacity to foresee and normative foreseeability would remain meaningful even if they were to exist as sequential steps in an analysis of breach. In such a case, the breach inquiry would ask (in relevant part) first whether the actor \textit{could} have foreseen and avoided the outcome, and second whether the actor \textit{should} have done so.\textsuperscript{165}

Some people can foresee possible future outcomes more often than others, and some are sufficiently out of touch with reality that they possess only the most minimal capacity to predict the consequences of their actions. A line has to be drawn to determine what degree of the capacity a person must possess before he or she is capable of being outcome-responsible at all. . . . That, in turn, seems inevitably to require looking . . . to the degree of capacity regularly exercised by ordinary or average people who strike us as possessing meaningful agency.

\textit{Id.}

\textsuperscript{164} See id. at 104 (“The capacity consists in being able to foresee and avoid outcomes in various sorts of circumstances on a sufficiently regular basis, where what counts as a sufficient degree of regular success is determined by the idealized conception of the ordinary person.”). If outcome-responsible, however, an actor is then—for the purpose of determining the reasonableness of her conduct—deemed to have the capacity to foresee enjoyed by the ordinary person. See id. (“[A]ll those who meet the minimum standard are treated as having an equal capacity to foresee and avoid outcomes, and hence are subject in an equal degree to being held responsible for the consequences of their actions.”). Thus, although the standard for whether an actor \textit{could} foresee injury is at least partly subjective, the standard for whether an actor \textit{should} foresee injury is completely objective.

\textsuperscript{165} Of course, as a practical matter, the question of whether an outcome could have been foreseen might well be subsumed in the decision-making process by the question of whether it should have been foreseen. Such a result would not pose a problem, however—the latter finding is outcome-determinative in any case.
By analogy, consideration of the means available to an actor to avoid an outcome is, in fact, traditionally considered to be a constituent of reasonableness. Suppose that a plaintiff, struck by the defendant’s car, were to argue that the collision would have been avoided had the defendant performed a reverse 180° slide (a professional stunt-driving maneuver). The reasonableness of the defendant’s failure to perform such a maneuver would depend on (1) whether it is within the capacity of the reasonable person to have done so generally and, if so, (2) whether a reasonable person should have done so under the circumstances. Both considerations are clearly matters within the scope of the jury’s resolution of breach. Perry’s analysis of an actor’s general capacity to foresee is analogous to (1) and therefore similarly might fit within a jury’s reasonableness inquiry.\(^1\) Put in terms amenable to Judge Learned Hand (where negligence exists if \(B < P \times L\)),\(^2\) if an outcome was in no meaningful way foreseeable because the actor was incapable of foresight, then not only would the epistemic probability of the outcome (\(P\)) be zero, but the actor’s burden in preventing it (\(B\)) would be almost immeasurably great. To hold one responsible for failing to avoid an unforeseeable outcome would be to place one in charge of serendipity. In sum, negligence works properly whether the general capacity to foresee is a part of duty or a part of breach.

Embracing Perry’s concept of the general capacity to foresee as a constituent of breach rather than duty would require no sacrifice of outcome-responsibility or of reasonableness. Thus, the more compelling reason to keep the two inquiries doctrinally separate would be a conclusion that the court should decide the former and a jury the latter. At least by implication, Perry sheds some light on this matter as well, although to see this one must understand Perry’s answer to the following question: at what level of epistemic probability is an ordinary person deemed minimally capable of foreseeing a type of outcome? It is a tautology to say that if the epistemic probability of an event occurring is greater than zero, the ordinary person is capable of foreseeing it. Yet if this were the test for judging one’s capacity to foresee, the in-

\(^{1}\) One might point out that the capacity to avoid the crash is judged pursuant to a completely objective standard, whereas Perry’s capacity to foresee inquiry is in part subjective. This difference is unproblematic, however. There is no reason that a fact-specific subjective inquiry cannot take place within the context of reasonableness. Indeed, a fact-specific inquiry into an actor’s subjective characteristics is squarely within the jury’s mandate to decide breach.

quiry would lose much of its probative value because almost any outcome is, by some stretch of the imagination, foreseeable.\textsuperscript{168} (This is especially true considering that judgments of foreseeability are made ex post the subject event.)\textsuperscript{169} Some normative judgment is therefore required in determining how much epistemic probability is sufficient to render an actor outcome-responsible. Perry hints that the requisite degree of epistemic probability is not particularly high\textsuperscript{170} and that it is potentially less than the epistemic probability necessary to support a conclusion that the relevant type of outcome was normatively foreseeable—that is, that the actor should have foreseen it.\textsuperscript{171} Perry also urges that the judgment should be made “with an eye to what constitutes meaningful agency in the world.”\textsuperscript{172} Despite such guidance, however, Perry concedes that decisions regarding an ordinary person’s “minimum capacity to foresee” remain indeterminate, an indeterminacy curbed only by the extent to which decisionmakers rely on the “common knowledge of the ordinary person.”\textsuperscript{173}

Two points may be drawn from Perry’s account. First, Perry explains that each constituent judgment of foreseeability—the judgment regarding an actor’s general capacity to foresee and the judgment regarding normative foreseeability—exists merely as a point drawn by the decisionmaker on the spectrum of epistemic probability.\textsuperscript{174} Each

\textsuperscript{168} See Ripstein, supra note 113, at 105 (“[E]xcept for the most bizarre of coincidences, everything is in principle foreseeable, and everything that has happened as a result of natural forces is in fact foreseeable.”).

\textsuperscript{169} The results of social science research suggest that knowledge that an event has occurred influences one’s judgment, in hindsight, of its foreseeability. Subjects who are told that an event in fact happened are more likely (than those who are not so told) to report that the event was foreseeable. This effect is known as “hindsight bias.” See, e.g., Chris Guthrie et al., Inside the Judicial Mind, 86 CORNELL L. REV. 777, 816 (2001) (finding that judges are empirically just as likely as other members of the public to fall prey to hindsight bias).

\textsuperscript{170} See Perry, Responsibility for Outcomes, supra note 104, at 94 (describing Perry’s conception as follows: “In general, the epistemic probability that will support a judgment of reasonable foreseeability need not be particularly high. In a case of unintentional harm, it is typically much more likely that the harmful outcome will not materialize than that it will.”).

\textsuperscript{171} See id. at 104–05 (describing generally the difference in foreseeability requisite to prove outcome-responsibility and the reasonableness of the actor’s conduct).

\textsuperscript{172} Id. at 104. This suggestion, although perhaps helpful at some level, is circular: one must determine outcome-responsibility (and hence agency) by looking to one’s understanding of “meaningful agency in the world.”

\textsuperscript{173} See id. at 103; see also supra notes 109–14 and accompanying text (describing the necessary reliance of such normative judgments on the common knowledge of the average person).

\textsuperscript{174} See supra notes 136–45 and accompanying text.
judgment is also drawn by reference to what is presumably the same
event-type and the same reference class of events.\textsuperscript{175} Furthermore, 
with regard to each judgment, both the characterization of the event-
type and reference class of events and the ultimate line-drawing re-
garding epistemic probability are properly made by reference to the 
common experience of the ordinary person.\textsuperscript{176} In light of these com-
monalities, one might argue that for consistency’s sake (and for 
efficiency’s sake) both the general capacity to foresee and normative 
foreseeability ought to be determined by the same decisionmaker. 
Second, because both foreseeability-related inquiries turn upon the 
common knowledge of the ordinary person, what better body to dis-
still and apply such knowledge than the jury, a group of just such peo-
ple?\textsuperscript{177} Perhaps for these reasons, Perry suggests, at least in passing, 
that it is the factfinder that properly determines an actor’s general 
capacity to foresee.\textsuperscript{178}

The foregoing discussion addresses specifically the account of 
foreseeability advanced by Stephen Perry, but it might similarly apply 
to other corrective justice theories of negligence.\textsuperscript{179} Although this Ar-
ticle cannot accommodate a comprehensive examination of such, Ar-
thur Ripstein’s rich description of risk and responsibility, and of fore-
seeability’s role in those concepts, adds an interesting dimension to 
the discussion. Ripstein’s explication of negligence liability differs 
from Perry’s in two important respects. First, rather than grounding 
his account in outcome-responsibility, Ripstein argues that negligence 
represents courts’ attempts to enforce fair terms of social interaction, 
a process guided by the need to balance shared interests in liberty and 
security.\textsuperscript{180} Ripstein asserts that the balance courts have generally 
achieved is best explained by the idea of “risk ownership”—only risks 
wrongfully imposed are “owned” by the actor in a sense that requires

\textsuperscript{175} See supra notes 109–11 and accompanying text.
\textsuperscript{176} See supra notes 109–14 and accompanying text.
\textsuperscript{177} See Cardi, supra note 7, at 794–804 (explaining the historical and institutional rea-
sons mitigating in favor of this proposition); Weinrib, supra note 94, at 519 (suggesting that 
the standard for permissible risk imposition “is not susceptible of precise measurement 
and is applied by the trier of fact on a case by case basis”).
\textsuperscript{178} See Perry, Responsibility for Outcomes, supra note 104, at 100 (indicating that it is the 
“trier of fact” that will decide the capacity of the ordinary person to foresee harm).
\textsuperscript{179} For example, Jules Coleman offers a “mixed” conception of corrective justice. See 
\textsuperscript{180} See Ripstein, supra note 113, at 49 (“My interpretation of fault liability . . . [em-
braces] the Kantian idea that the boundaries of individual rights are given by fair terms of 
social interaction. . . . by a concern for equal liberty and security for all.”).
compensation. Second, Ripstein objects to Perry’s conception of outcome-responsibility to the extent that it requires subjective knowledge (or even subjective capacity for knowledge) of a risk. Ripstein urges that instead, negligence embodies a completely objective standard in this regard. Thus, according to Ripstein, “foresight is not required because it is a general condition of agency. . . . [Rather,] it is implicit in the idea of fair terms of interaction.”

Importantly, although Ripstein seems to define duty and breach such that each includes some consideration of foreseeability, he contends that foreseeability generally is an independent doctrinal requirement for liability, separate from and prior to both duty and breach. In Ripstein’s words,

[Foresight is] a preliminary test for liability . . . . On this understanding, if an injury was unforeseeable, there is no further question of liability to ask, because the defendant could not have taken account of the risk that it would happen. Once the test of foreseeability has been passed, the further inquiry is fixed by questions of duty and risk.

---

181 See id. (“[T]he boundaries between persons are given by a concern for equal liberty and security for all. . . . [F]or purposes of liability, the distinction between what someone does and what merely happens is a normative distinction. Provided I exercise appropriate care, the consequences of my actions are treated as though they merely happened. If I fail to exercise appropriate care, though, the risks I create are mine to bear, and if they ripen into injuries, I must bear the costs of those injuries.”).

182 See id. at 100 “[Perry’s account] cannot . . . be combined with the idea that the duty of care is itself objective, in the sense that someone can be responsible for something that he did not himself foresee.”). Perry has responded that Ripstein’s conception of control is too narrow. Perry, The Distribution Turn, supra note 122, at 332–33; see also Perry, Responsibility for Outcomes, supra note 104, at 89–90 (suggesting that due to this argument, it is unclear whether Ripstein is a corrective justice theorist at all; and that Ripstein’s argument may ultimately devolve into a distributive account of negligence liability). Perry also promises to respond more extensively to Ripstein’s critique in some future work. See Perry, Responsibility for Outcomes, supra note 104, at 130 n.80.

183 See RIPSTEIN, supra note 113, at 56 (“To protect all equally requires weighing liberty against security, but any weighing is done within the representative reasonable person, rather than across persons. The point of weighing interests within a representative person is to avoid allowing the particularities of one person’s situation to set the limits of another’s liberty or security.”).

184 Id. at 105.

185 Id. at 104. In this way, Ripstein’s approach is analogous to the provocative suggestion of Ben Zipursky that foreseeability is in fact a constituent of substantive standing. See generally Zipursky, supra note 97, at 27–40 (proposing his substantive standing theory in light of several aspects of negligence).
Unfortunately, Ripstein does little to expound on this suggestion. The implication, however, is that duty and foreseeability are not co-dependent concepts. Ripstein thus appears to believe that the essence of duty is something other than foreseeability. Perhaps duty is simply the step in liability analysis at which courts create categorical guidelines with regard to society’s competing interests in liberty and security.186 By this reasoning, duty determinations might not involve forays into foreseeability at all.

B. Foreseeability as Instrument

A second theoretical approach explains and justifies negligence liability, and the doctrinal elements on which liability turns, solely according to the goals to be achieved by such a system. From this perspective, negligence is merely an instrument; even if negligence doctrine exists in some coherent form, it is only meaningful and valid to the extent that it serves the law’s ultimate goals.187 Possible goals of negligence include providing compensation for injured persons, loss spreading, wealth redistribution, deterrence, punishment, social and personal retribution, efficiency, the maximization of wealth or utility, and perhaps others.188 A particular instrumentalist approach might focus on one goal or serve some combination of goals. Although the reasoning in this section would apply to any instrumentalist approach, I will focus on the most comprehensive instrumentalist approach existing: the economic theory of negligence.

The most prominent economic account of tort law is that of William Landes and Richard Posner.189 For Landes and Posner, the goal of tort law is to promote the efficient allocation of resources or as Posner has separately proposed, to maximize societal wealth.190 According to this view, negligence law should create incentives for actors to take into account the potential costs of their actions. Where the cost of avoiding a risk of injury would be less than the cost of the risk

186 See Ripstein, supra note 113, at 92 (“In cases of misfeasance, the existence of duty of care does not depend on the ease with which it can be discharged in the particular instance. Instead, it depends on the significance of the relevant interests in liberty and security.”).
188 Id. at 487.
190 Id. at 1.
itself, the threat of negligence liability should encourage one to avoid taking the risk. Conversely, where the cost of prevention is greater than the cost of the risk, liability should not ensue. The imposition of liability in such case would have no appreciable deterrent effect and would punish the defendant for having acted rationally. In this sense, according to Landes and Posner, negligence should seek to create an efficient level of deterrence, but no more.

The economic theory of negligence relies on foreseeability in two ways. First, foreseeability plays a role in the determination of the reasonableness (in economic terms) of a defendant’s conduct. In Landes and Posner’s view, foreseeability aids in calibrating the proper level of deterrence. Only an injury that is foreseeable is capable of being deterred. Moreover, in determining the reasonableness of the defendant’s conduct, the foreseeable probability of injury, combined with its foreseeable range of costs, is to be weighed against the cost of avoiding them. Where the cost of avoiding injury is less than the cost of the risk, the defendant is deemed to have acted unreasonably in failing to avoid causing injury. The higher the risk, the more care-

193 Because the cost of avoidance would exceed the benefit of doing so, the defendant (and similarly situated future actors) will not take preventive measures despite the threat of legal liability. This reasoning is, of course, subject to an assumption that the level of risk of the activity is appropriate—if it is not, then a court might impose strict liability rather than negligence.
194 See Weinrib, supra note 94, at 504 (explaining that “spending more money to prevent an injury than the injury itself costs would be wasteful”).
195 See Landes & Posner, supra note 189, at 247 (“[T]he [one-bite] rule is that the owner is liable only if he has reason to suspect the dog’s vicious disposition; and ordinarily there is no reason to suspect it until the dog has bitten someone. Even if the probability of the dog’s biting someone is very high, the owner will not be liable unless he has reason to know it is high. Otherwise, as we have said, liability will have no allocative effect.”).
196 Professor Mark Grady has made two important clarifications of Landes and Posner’s theory. First, there is some question whether Landes and Posner condition breach upon reasonable foreseeability or upon simple objective probability. Professor Grady has sufficiently demonstrated that it must be the former. See Mark F. Grady, Proximate Cause and the Law of Negligence, 69 Iowa L. Rev. 363, 364 (1984) (“The theory to be explained here demonstrates that the breach-of-duty question depends on the amount of information concerning the risk that it was reasonable for the injurer to have had.”); id. at 385–391 (demonstrating the same). Second, Landes and Posner apparently believed that for practical purposes the breach determination should be made only by reference to the risk of the particular injury suffered by the plaintiff. Id. at 382–83. Professor Grady has demonstrated that this is not and cannot be the case and that instead, breach must be determined with regard to the entire range of foreseeable risks created by a defendant’s conduct. Id. at 383–85.
197 The cost of foreseeing injury—for example, the cost of investigating the various societal effects of a pesticide that one plans to use on one’s garden—is included as part of the defendant’s burden under the Learned Hand formula.
ful the defendant is required to have been. This, of course, is the “Learned Hand test” first proposed by Judge Learned Hand in the case of United States v. Carroll Towing Co.

Even if $B < P \times L$, however, a defendant should be liable only for harms with regard to which liability will have some allocative effect. The second way in which economic theory relies on foreseeability is in deciding which of the various risks created by negligent conduct an actor should internalize. At one level, it may seem harmless to hold a defendant liable for all of the consequences of his or her negligent actions. As Landes and Posner explain, “[i]t may indeed be harmless, because if the accident is unforeseeable then so is the imposition of liability for its consequences. Hence there is no danger . . . of inducing too much care.” Assuming perfect enforcement, however, unforeseeable injuries are unnecessary and ineffective in producing the efficient level of deterrence. From the perspective of the individual actor, the threat of liability for unforeseeable risks will not deter the relevant conduct because the cost of foreseeing the risk exceeds the benefit of avoiding that particular risk. Thus, according to Landes & Posner, courts should not force a defendant to internalize a risk that is not reasonably foreseeable because such liability would confer no economic benefit. To impose liability nonetheless would be to “merely require a costly transfer payment.”

Landes and Posner offer, by example, the case of Watson v. Kentucky & Indiana Bridge & Railroad Co., in which the defendant negligently caused a railroad tank car to be derailed and leak fuel, which

198 See, e.g., Lollar v. Poe, 622 So. 2d 902, 905 (Ala. 1993) (“The degree of care required of an animal owner should be commensurate with the propensities of the particular animal and with the place where the animals are kept, including its proximity to high-speed highways.”); Indust. Chem. & Fiberglass Corp. v. Chandler, 547 So. 2d 812, 831 (Ala. 1988) (“[T]he who deal with dangerous instrumentalities, such as explosives or chemicals, must exercise a great amount of care because the risk is great.”).

199 159 F.2d at 173–74 (enshrining these factors in the mathematical formula in which liability lies where $B$ (burden of precautions) $< P$ (probability of loss) $\times L$ (magnitude of loss)); see also Markowitz v. Ariz. Parks Bd., 706 P.2d 364, 369 (Ariz. 1985) (recognizing that foreseeability of risk and the burden of precautions are “factors which determine the reasonableness of the defendant’s conduct”); Dobbs, supra note 2, §§ 143–146, at 334–48 (explaining in detail the interplay of foreseeability and reasonableness here summarized).

200 Zipursky, supra note 97, at 46.

201 LANDES & POSNER, supra note 189, at 247.

202 Id. at 246.

203 Id. at 247.

204 Id.

205 126 S.W. 146 (Ky. 1910).
subsequently exploded when the plaintiff deliberately set fire to it.\textsuperscript{206} Although the defendant’s conduct in \textit{Watson} was unreasonable in economic terms (because the cost of avoiding the derailment was less than the resulting risk of injury),\textsuperscript{207} Landes and Posner concur with the court’s finding for the defendant on a foreseeability analysis:

[T]he possibility of arson was so slight (it was an act of pure malice, with no possibility of pecuniary or any other benefit—except the delights of pyromania) that . . . the defendant would not have taken account of it in deciding how much care to use; so imposing liability would have had no allocative effect.\textsuperscript{208}

According to Landes and Posner, foreseeability’s part in reasonableness serves the negligence element of breach, whereas the foreseeability judgment described above falls under the doctrinal umbrella of proximate cause.\textsuperscript{209} Indeed, in their discussion of \textit{Watson}, Landes and Posner implicitly justify the “risk rule,” a theory of proximate cause discussed at length in the following section.\textsuperscript{210} In theory, however, so long as negligence liability is conditioned on foreseeability in the right ways, an economic approach to negligence is indifferent to whether such is accomplished by means of duty or breach and proximate cause. By definition, an instrumentalist approach cares only about the outcomes generated by the system and little about the system’s intrinsic ordering or conceptual coherence.\textsuperscript{211} Economic theory is thus indeterminate as to foreseeability’s proper doctrinal fit.\textsuperscript{212}

\textsuperscript{206} Id. at 147.

\textsuperscript{207} See \textsc{Landes} & \textsc{Posner}, \textit{supra} note 189, at 247 (explaining that in \textit{Watson}, “the defendant was negligent with respect to the accident probability that was known”).

\textsuperscript{208} Id. Interestingly, the Kentucky Supreme Court later overruled \textit{Watson}, rejecting the reasoning cited by Landes and Posner. \textit{See Britton v. Wooten}, 817 S.W.2d 443, 449–52 (Ky. 1991).

\textsuperscript{209} See \textsc{Landes} & \textsc{Posner}, \textit{supra} note 189, at 243–51 (discussing the foregoing matters within the chapter entitled “Causation”).

\textsuperscript{210} See \textit{infra} notes 265–69 and accompanying text.

\textsuperscript{211} See \textsc{Weinrib}, \textit{supra} note 94, at 487–88.

\textsuperscript{212} This conclusion must be hedged in the following way. In a phone conversation regarding this Article, Professor Ken Kress pointed out that foreseeability’s doctrinal placement might result in a shift in its substance due to the “gravitational effect” of surrounding doctrine. For example, as the duty question mulls through the brain of the judge, the judge’s consideration of other duty norms might “attract” foreseeability, causing it to become more akin to those norms, or it might repel foreseeability, having the opposite effect. The jury’s consideration of foreseeability in the context of breach or proximate cause would take place in the presence of a different set of norms, with correspondingly differ-
This is not to say, however, that the field of law and economics has nothing to add to the ultimate debate. To the contrary, economic theory might lend considerable insight into whether court or jury is the better arbiter of foreseeability. One might argue, for example, that court-rendered, categorical foreseeability determinations induce greater efficiency by providing clearer rules for market actors. In addition, early dismissal of a class of cases almost certainly saves in the administrative costs of trials.213 On the other hand, jury decisions are more precisely calibrated (because they are fact-specific) and may therefore result in a more accurate efficiency determination in any particular case and perhaps even over time.214

Finally, it is perhaps significant that the distinction between foreseeability’s place in determining economic reasonableness and its role in deciding which risks a negligent actor should internalize is not so much a distinction in kind as it is in focus. Consider how the latter judgment of “reasonable foreseeability” is made. The marginal cost to the defendant of foreseeing and preventing the particular class of injuries suffered by the plaintiff is weighed against the risk of that class of injuries occurring. If the marginal cost of foresight is less than the risk of such injury, then the plaintiff’s harm was reasonably foreseeable, and liability ensues. If, on the other hand, the marginal cost of foresight is greater than the relevant risk, liability would have no allocative effect—the defendant would not have been deterred even in the face of certain liability. This foreseeability calculus is similar to the economic determination of reasonableness, except that it focuses on a particular set of costs and injuries rather than on the entire universe of the same. In light of this similarity, one might argue that efficiency


214 Cf. id. at 256–57 (“[C]ategory-by-category allocation makes the introduction of individualized moral judgments into the allocation decision harder. This explains why category allocation cannot be used in the fault system.”). Calabresi also suggests that because court determinations of foreseeability are less flexible than fact-specific allocations, they may therefore suffer lags in recognizing changes in who is the cheapest cost-avoider. Id. at 257. Finally, he points out that although categorical, duty-based decisions concentrate on “general or recurring causes of accidents,” they might overlook “some very cheaply avoided particular causes of very high accident costs.” Id.
is best served if the same decisionmaker—the jury—is entrusted with both decisions.  

C. Foreseeability as a Constituent of Relation and Limitation

Negligence liability is, at least in some sense, relational. Ours is not a New Zealand-style social compensation system pursuant to which everyone pays into a central fund and from which tort victims collect. Rather, American courts coerce compensation directly from tortfeasor to victim. Negligence doctrine must tie the defendant’s wrongful acts to the victim’s injuries in a way that justifies coerced repayment. Although the link between the defendant’s wrong and the plaintiff’s right to compensation might serve various instrumental goals, it also has moral underpinnings rooted in our desire to effect recourse and correct moral imbalance. The tails-side of this relational concept is that negligence liability must typically be limited to those plaintiffs or injuries for which coerced compensation is morally justified. Coerced transfer that is not so justified offends our general commitment to freedom and to property rights, values inexorably part of the American landscape.

A debate roils in the courts and within the academy regarding the proper doctrinal means for effectuating the concepts of relation and limitation. According to one view—which I will call the “non-relational” approach—the concept of duty or obligation is largely actor-centered. Pursuant to this view, one owes to society in general an

---

215 I leave more extensive economic evaluation of the court-jury debate to others. Again, however, I urge that this is the important question to be answered.


217 See Weinrib, supra note 94, at 520 (“The concepts of proximate cause and duty normatively link the wrongfulness of the defendant’s unreasonable risk-creation with the wrongfulness of the plaintiff’s suffering.”).

218 See, e.g., supra notes 200–04 and accompanying text (describing relevant economic goals).


220 There may be a number of bases for limiting negligence liability other than for lack of a moral justification for coerced transfer of resources. This Section, however, only focuses on this particular concept of limitation.

221 Indeed, these are the pillars of classic liberalism. See Daniel A. Farber & Suzanna Sherry, A History of the American Constitution 5, 9 (1990); Philip A. Hamburger, Natural Rights, Natural Law, and American Constitutions, 102 Yale L.J. 907, 918–19 (1993).
obligation to exercise care in one’s actions. Courts thus impose a duty of care solely on the basis of public policy and community standards of obligation, without regard to the question of to whom such duty is owed. From this non-relational perspective, both the relational aspect of negligence liability and the requisite limitations on liability are to be resolved primarily via the elements of factual and proximate causation.

According to the “relational” conception, by contrast, the doctrinal means for deciding relation and limitation are splintered. Although proximate cause is seen to be the element best suited for imposing limits on the scope of liability, duty is the vessel through which courts establish the proper relation between defendant and plaintiff. To a relationalist, the concept of duty or obligation is incomplete without some connection between the actor and a person or class of persons. The relational approach thus requires a two-step duty query: (1) did the defendant owe a duty to the plaintiff or to a class of people of which plaintiff is a member, and (2) was it this duty

---

222 See, e.g., PROSSER & KEETON, supra note 27, § 53 (“Certainly [in the early common law] there is little trace of any notion of a relation between the parties, or an obligation to any one individual, as essential to the tort. The defendant’s obligation to behave properly apparently was owed to all the world . . . .”); Oliver W. Holmes, Jr., The Theory of Torts, 7 AM. L. REV. 652, 661 (1873) (describing the universal tort duty as “a duty imposed on all the world, in favor of all”).

223 Goldberg & Zipursky, supra note 18, at 1817–18. Of course, even from a generally non-relational view of duty, many affirmative duties depend upon the relationship between the plaintiff and the defendant—for example, the affirmative duties owed by doctor to patient or innkeeper to guest.

224 See Weinrib, supra note 94, at 521 (“Duty focuses on the class of persons affected by the defendant’s negligence, proximate cause on the kind of accident or injury generated by that negligence. The two concepts cover, respectively, the questions of risk to whom and risk of what.”).

225 John C. P. Goldberg & Benjamin C. Zipursky, The Restatement (Third) and the Place of Duty in Negligence Law, 54 VAND. L. REV. 657, 707 (2001) (“[D]uty in its primary sense is an analytically relational concept: It concerns obligations of care that are owed by certain persons to certain other persons.”) (emphasis omitted). To illustrate, assume that I am the last living human on earth. In such case, intuition suggests that I cannot possibly owe any duty of care. According to the relationalist, this intuition stems from the fact that the concept of duty is incomprehensible without considering to whom duty is owed.
that the defendant allegedly breached.226 The latter aspect of this inquiry is sometimes called the “duty-breach nexus.”227

At the center of the debate between the relational and non-relational views of duty lies the concept of foreseeability. According to the non-relational account, foreseeability of the plaintiff and foreseeability of the type of the plaintiff’s injury operate to limit the consequences of a defendant’s negligent act. Although the defendant may have wrongfully created some risk of injury to some category of people, the consequences of that wrong must be limited to people and injuries that were reasonably foreseeable. These limitations might be justified on instrumental grounds—for example, liability for harm to an unforeseeable plaintiff or injury cannot serve as a deterrent228—or on moral grounds—for example, one cannot be responsible for an injury or plaintiff that could not have been foreseen and thus could not have influenced one’s choice in acting.229 Because proximate cause is generally seen to be the seat of such “limitation” decisions, the non-relationalist believes that foreseeability of plaintiff and of the type of manner of injury fit seamlessly in that element.

From a relational view, by contrast, although foreseeability of the type or manner of injury serves to limit liability in a non-relational way and therefore fits within the element of proximate cause, plaintiff-foreseeability is a necessary constituent of relation.230 Only if some harm to the plaintiff was foreseeable might a defendant’s relationship

226 See Goldberg, supra note 43, at 1336–39 (explaining this concept in relation to the idea that the defendant’s unreasonable conduct subjected the plaintiff to a “wronging”); Weinrib, supra note 94, at 520 (“[These concepts] exclude[] liability when the plaintiff is injured by the defendant’s wrongful act without being wrongfully injured.”); Zipursky, supra note 97, at 7–15 (explaining the concept generally).
227 See Goldberg & Zipursky, supra note 225, at 709–12 (explaining the duty-breach nexus requirement). Separately, Professor Zipursky makes the provocative suggestion that this duty-breach nexus foreseeability is so important and omni-present that it constitutes a substantive standing requirement for all negligence claims. See Zipursky, supra note 97, at 27–40 (proposing his substantive standing theory in light of several aspects of negligence).
228 See supra notes 200–04 and accompanying text.
229 See supra notes 136–40 and accompanying text.
230 See Goldberg, supra note 43, at 1337–39 (distinguishing a hypothetical that turns on the foreseeability of the type of injury—to be handled by proximate cause—from Palsgraf, which turned on the foreseeability of the plaintiff and is a matter of duty); Perry, Responsibility for Outcomes, supra note 104, at 95 (noting that although foreseeability of the plaintiff is a matter of duty, foreseeability of the type of harm is “the basic principle of proximate cause”).
with the plaintiff give rise to an obligation of care. Because relation-based obligation is the essence of duty, plaintiff-foreseeability must, to a relationalist, constitute a part of the duty calculus.

This divide over the proper doctrinal scheme for the relational and limiting uses of foreseeability was brought into the spotlight by the contrasting opinions of Judge Cardozo and Judge Andrews in *Palsgraf v. Long Island Railroad Co.* *Palsgraf* involved the claim of a railroad passenger who was injured by a falling set of scales while standing on the station platform. The injury resulted when railroad employees, standing at the opposite end of the platform from the plaintiff, negligently dislodged a package from the hands of a passenger while helping him board a departing train. Unbeknownst to the employees, the package contained fireworks, which exploded upon impact. The force of the explosion toppled the scales, which in turn harmed the plaintiff, Mrs. Palsgraf. Writing for the dissent, Judge Andrews held that the railroad owed a “duty of refraining from those acts that may unreasonably threaten the safety of others.” According to Judge Andrews, this duty was not relational but a duty to the world. In his words, “[d]ue care is a duty imposed on each one of us to protect society, . . . not to protect A, B, or C alone.” With the elements of duty, breach, and factual causation satisfied in Mrs. Palsgraf’s case, Judge Andrews opined that any limitation on the railroad’s liability must therefore be effected by proximate causation, in which foreseeability plays an important role.

Writing for the majority, Judge Cardozo approached the case from a relational perspective. Although Judge Cardozo held that the railroad indeed owed a duty to its passengers and that it had breached that duty with respect to the man carrying the package (and perhaps to others nearby), the railroad had nevertheless breached no duty

---

231 See, e.g., Ripstein, *supra* note 113, at 52 (“Parties engaging in potentially risky activities must show reasonable care for those who might be injured by those activities, not simply for the persons who turn out to be so injured.”).


233 Id.

234 Id.

235 Id.

236 Id.


238 Id. at 102.

239 Id. at 103–04. Although Judge Andrews wrote that foreseeability plays a central role in proximate cause determinations, he felt that the decision was largely a matter of “practical politics.” Id. at 103.
owed to the plaintiff.\textsuperscript{240} In reaching this conclusion, Judge Cardozo reasoned that “[t]he risk reasonably to be perceived defines the duty to be obeyed, and risk imports relation.”\textsuperscript{241} In other words, because it was not reasonably foreseeable that the defendant’s actions would injure Mrs. Palsgraf, the requisite relation between the defendant’s wrong and the plaintiff’s injury was missing. Without relation there can be no duty, or at least the duty alleged to have been breached was not a duty owed to the plaintiff.\textsuperscript{242}

The central difference between the majority and the dissent in \textit{Palsgraf} is whether the requisite relational link between defendant’s wrong and plaintiff’s injury should be decided in the context of duty or proximate cause. As mentioned above, the argument most frequently made in favor of Judge Cardozo’s approach is that the “duty-breaching nexus” is a question of relation, not a policy-driven limitation on liability.\textsuperscript{243} It therefore fits best in, and indeed is necessary to, the element of duty. Foreseeability of the type of injury, on the other hand, limits liability in a non-relational way and therefore fits within the element of proximate cause. Professor John Goldberg illustrates this distinction with the following pair of hypotheticals:

\begin{itemize}
\itemSee id. at 100 ("What the plaintiff must show is ‘a wrong’ to herself; i.e., a violation of her own right, and not merely a wrong to some one else, nor conduct ‘wrongful’ because unsocial, but not ‘a wrong’ to any one."); see also Goldberg, \textit{supra} note 43, at 1336 ("[T]he conduct in question—although perhaps antisocial, as well as wrongful to persons standing near the conductors—could not have amounted to a wronging of [Mrs. Palsgraf]; it did not constitute a harm flowing from a breach of any of the duties that were owed to her."); Zipursky, \textit{supra} note 97, at 7–15 (explaining the duty-breaching nexus as the core of Judge Cardozo’s opinion).
\itemPalsgraf, 162 N.E. at 100.
\itemProfessor Perry offers another interesting hypothetical scenario involving the duty-risk nexus:
\begin{itemize}
\itemThe defendant kidnaps someone, and in the course of his felonious activity quite unforeseeably and non-negligently injures a third party. Does the injured third party have a morally justified claim in tort against the defendant, just because her injury would not have occurred but for the defendant’s violation of the moral and legal norm against kidnapping?
\end{itemize}
\itemPerry, \textit{Responsibility for Outcomes}, \textit{supra} note 104, at 116. As Perry explains, “the element of fault [in this hypo] is not related in the right way to the antecedent element of outcome-responsibility.” \textit{Id.} at 118.
\itemSee Goldberg & Zipursky, \textit{supra} note 18, at 1820 (“[F]oreseeability is not intended as a policy-driven or fairness-based limitation on the harms for which a wrongdoer may be held liable. To read the opinion this way is to convert what Cardozo regarded as a duty question concerning conduct and obligation into a proximate cause question concerning the extent of liability. For Cardozo, the foreseeability of harm to a class of persons goes to the question of whether certain conduct is owed to those persons, not to whether certain liabilities are appropriately borne by defendant[].”.
\end{itemize}
First, imagine a person who, while driving his car on a street in a moderately busy part of town, carelessly throws a half-filled paper coffee cup out of the driver’s side front window. The coffee proceeds to splatter on the windshield of a car coming in the opposite direction, temporarily blocking the vision of the driver of that car, who swerves, strikes, and seriously injures a pedestrian standing on the far sidewalk.

Now imagine the same careless act—the throwing of the half-filled coffee cup by the driver—except that, instead of hitting another car’s windshield, the cup lands harmlessly in the road. However, a pedestrian walking along the far curb sees the tossing of the coffee, stops in his tracks, and raises his arms in indignation over the driver’s act of littering. Only because the pedestrian happens to stop at that precise point and gesticulate, he makes contact with a tree limb, which in turn disturbs a hidden nest of bees, many of which sting him, causing him to suffer disfiguring welts and considerable pain.\(^\text{244}\)

According to Professor Goldberg (and presumably to other relationalist), the difference in likely outcomes of the first and second hypotheticals turns on proximate cause rather than on duty.\(^\text{245}\) Unlike Mrs. Palsgraf, the pedestrian in Professor Goldberg’s hypothetical was a foreseeable plaintiff—thus, the relation between driver and pedestrian was sufficient to establish the existence of a duty.\(^\text{246}\) Rather, it was the pedestrian’s type of injury that was unforeseeable. Presumably stemming from the view that this brand of foreseeability has little to do with the relation between plaintiff and defendant, Professor Goldberg maintains that although duty is the proper forum for cases such as \textit{Palsgraf}, “the notion of proximate cause better explains cases such as the bee-sting hypothetical.”\(^\text{247}\)

The relational account of plaintiff-foreseeability is, in large part, accurate as a positive description of courts’ behavior. Despite what

\(^{244}\) Goldberg, supra note 43, at 1337.  
\(^{245}\) \textit{Id.} at 1338.  
\(^{246}\) \textit{Id.}  
\(^{247}\) \textit{Id.} at 1338 n.64. I say “presumably” because Professor Goldberg does not explain the reasons underlying this statement—perhaps in light of his more extensive discussion of this topic in works that he has co-authored with Professor Zipursky, which are cited throughout this Article.
some scholars have posited, many courts seem to view plaintiff-foreseeability as a duty question. Such an approach, however, is in my view conceptually flawed, and one need not adopt a non-relational view of duty to expose these flaws. One might simultaneously maintain the views that duty generally (or at least often) is relational and that foreseeability’s function as a constituent of relation and limitation is best instantiated by the element of proximate cause. Three false dichotomies weaken the relational approach: (1) the dichotomy between plaintiff-foreseeability and foreseeability of the type of injury; (2) the dichotomy between relation and limitation; and (3) the dichotomy between the relational nature of duty and the non-relational nature of proximate cause. I discuss each in turn.

1. The Dichotomy Between Plaintiff-Foreseeability and Foreseeability of the Type of Injury

A conclusion that plaintiff-foreseeability must exist as a constituent of duty but that injury-foreseeability is a matter of proximate cause rests in part on the presupposition that there exists some meaningful distinction between the two concepts. It is this distinction that Professor Goldberg seeks to illustrate by comparing Palsgraf with the bee-sting hypothetical. The most that Professor Goldberg’s comparison reveals, however, is that fact patterns which, from a relational view, present an issue of plaintiff-foreseeability are merely a subset of cases that call for a determination of the foreseeability of a particular

---

248 As Professor Zipursky points out, the predominant approach in torts casebooks is to discuss Palsgraf in the section detailing proximate cause. See Zipursky, supra note 97, at 3 & n.4 (citing the relevant casebooks).

249 A recent canvass of state court decisions revealed only two jurisdictions that clearly favor an Andrews-like approach to plaintiff-foreseeability. See Wintersteen v. Nat’l Cooperage & Woodenware Co., 197 N.E. 578, 582 (Ill. 1935) (“It is axiomatic that every person owes a duty to all persons to exercise ordinary care to guard against any injury which may naturally flow as a reasonably probable and foreseeable consequence of his act . . . . This duty . . . . extends to remote and unknown persons.”); Alvarado v. Sersch, 662 N.W.2d 350, 353 (Wis. 2003) (“Wisconsin has long followed the minority view of duty set forth in the dissent of Palsgraf v. Long Island Railroad. . . . ‘[E]veryone owes to the world at large the duty of refraining from those acts which may unreasonably threaten the safety of others.’”) (citation omitted) (quoting Palsgraf, 162 N.E. at 103 (Andrews, J., dissenting)). But see Restatement (Third) of Torts: Liab. for Physical Harm (Basic Principles) § 29, Reporter’s Note to cmt. f, at 45–49 (Tentative Draft No. 3, 2003) (citing cases that purportedly consider plaintiff-foreseeability to be a matter of proximate cause). Interestingly, a growing number of courts also sweep foreseeability of the type and manner of injury into the penumbra of duty. See Cardi, supra note 7, at 755–60 (describing this phenomenon and citing examples).

250 See Goldberg, supra note 43, at 1337.
risk or a particular type of injury. One might, for example, describe the issue in *Palsgraf* to be not whether Mrs. Palsgraf was foreseeably harmed by the act of dislodging a package out of a distant passenger’s hands, but whether it was foreseeable that such act could result in harm due to an explosion.\(^{251}\) Of course, a relationalist might suggest that even so, the question remains whether Mrs. Palsgraf was foreseeably injured by such an explosion. This too, however, is unproblematically characterized in terms of injury rather than plaintiff: if an explosion was foreseeable, was it also foreseeable that injury would result both of the type Mrs. Palsgraf received and in the manner in which she received it? Similar reasoning might apply to many cases that courts currently deem to involve plaintiff-foreseeability.\(^{252}\)

In light of this alternative perspective on plaintiff-foreseeability, the distinction between *Palsgraf* and Professor Goldberg’s bee-sting hypothetical is rendered at least potentially meaningless. That is, the relational distinction between plaintiff-foreseeability and injury-foreseeability becomes unnecessary absent some practical or metaconceptual reason to keep it.\(^{253}\) A desire that courts rather than juries decide such questions might supply such a reason. The more common justification, however, is that duty, not proximate cause, is relational, and that a relational understanding of duty requires the dis-

\(^{251}\) Judge Andrews approximated this reasoning in applying his “hindsight” test, a peculiar variant of foreseeability. See *Palsgraf*, 162 N.E. at 105 (Andrews, J., dissenting) (“[G]iven such an explosion as here, it needed no great foresight to predict that the natural result would be to injure one on the platform at no greater distance from its scene than was the plaintiff.”). Professor Grady also proposes a risk-oriented interpretation of *Palsgraf*. See Grady, supra note 196, at 414 (explaining *Palsgraf* as having applied the “reasonable-foresight doctrine” where “the same set of precautions reduces two different expected harms or risks,” one foreseeable and one not).

\(^{252}\) For example, Professor Perry’s kidnapping hypothetical, for instance, described supra note 242, is equally susceptible to such an analysis. Rather than presenting the question of whether kidnapping posed a foreseeable risk to the injured driver, the issue might instead be described as whether injury by car crash is a foreseeable risk of kidnapping. See also Fawley v. Martin’s Supermarkets, Inc., 618 N.E.2d 10, 12–14 (Ind. Ct. App. 1993) (analyzing plaintiff-foreseeability when plaintiffs sued a supermarket for negligent failure to protect them from a drunk driver outside the defendant’s business); Valentine v. On Target, Inc., 727 A.2d 947, 949–50 (Md. 1999) (same when a plaintiff sued the defendant gun dealer for the murder of the plaintiff’s decedent, committed by a third party with a gun stolen from the defendant’s store).

\(^{253}\) See Restatement (Third) of Torts: Liab. for Physical Harm (Basic Principles) § 29, Reporter’s Note to cmt. m (Tentative Draft No. 3, 2003) (suggesting that *Palsgraf* and similar cases do not necessarily turn on plaintiff-foreseeability, but might just as easily be characterized as foreseeability of harm cases); William Powers, Jr., *Reputology*, 12 Cardozo L. Rev. 1941, 1949 (1991) (same).
tinction and separation of plaintiff-foreseeability from injury-foreseeability. It is to this assertion that I now turn.

2. The Dichotomy Between Relation and Limitation

As described above, the primary justification for assigning plaintiff-foreseeability to duty and foreseeability of type of injury to proximate cause is the distinction between relation and limitation: the essence of duty is a relation-based obligation of which plaintiff-foreseeability is a necessary constituent, and the essence of proximate cause is to serve as a tool for limiting liability with regard to which injury-foreseeability is relevant. The basic dichotomy on which this justification rests, however, is flawed. Relation and limitation are not distinct concepts, at least not in the context of foreseeability determinations. Rather, they are merely two possible conclusions to the following question: is the link between defendant’s wrong and plaintiff’s injury of a kind for which community notions of obligation and policy suggest that a defendant be held liable? A finding that the plaintiff and the plaintiff’s injury were foreseeable provides a basis for finding such a link and therefore justifies coerced compensation; a finding that the plaintiff or injury was unforeseeable means that such a link is absent and therefore serves to “limit” the reach of coerced compensation even in the face of wrongful conduct.254

This is not to say that the relation and limitation inquiries completely overlap. Indeed, as relationalists point out, plaintiff-foreseeability is merely a necessary, but not a sufficient condition for a finding of the requisite relation.255 Similarly, limitations on the extent of liability also might stem from other sources—256—for example, due to the lapse of time between the negligent act and injury.257

---

254 Although Professor Perry considers, at least as a normative matter, plaintiff foreseeability to be a matter of duty and injury-foreseeability to be a matter of proximate cause, Perry seems to recognize that relation and limitation are inexorably tied: “Some [facets of proximate cause] . . . seem to be built into the concept of outcome-responsibility itself. . . . If a foreseeable type of harm occurs in too freakish or improbably a manner, there is insufficient control over the outcome to support a judgment of outcome-responsibility . . . .” See Perry, Responsibility for Outcomes, supra note 104, at 95.


256 See Perry, Responsibility for Outcomes, supra note 104, at 95 (“There are many facets to the law of proximate cause in addition to the basic foreseeability principle.”).

257 See, e.g., Martinez v. California, 444 U.S. 277, 285 (1980) (holding that the defendant state could not be held liable for the murder of a girl by a parolee five months following the parolee’s release from prison, in part because of the significant lapse in time between the parole release and the murder).
seeability aspect of relation and limitation, however, simply combines two opposing forces of the same inquiry—the desire for relation and the preservation of residual freedoms. Division of the concept between separate elements of negligence is therefore misleading and redundant.

Such an argument is, of course, indeterminate as to whether duty or proximate cause is the better home for the combined inquiry. In fact, this is precisely the point. If the distinction between relation and limitation is indeterminate as to the proper doctrinal home for foreseeability, perhaps the only determinative factor is whether court or jury is the preferable primary decisionmaker.

3. The Dichotomy Between the Relational Nature of Duty and the Non-Relational Nature of Proximate Cause

Even if one disagrees with the assertion that relation and limitation are opposing aspects of the same inquiry, the common charge that relation is exclusive to duty and that limitation is exclusive to proximate cause is, in my view, not accurate. Rather, duty and proximate cause are both relational in some respect, and both also limit liability. The charge is therefore indeterminate as grounds for conditioning duty on plaintiff-foreseeability.

Despite Judge Andrews’s protests that each of us owes to the world a duty to avoid causing harm, it is difficult to deny that duty is at least sometimes relational. Many of our daily obligations, both legal and pre-legal, exist only in relation to a limited group of people. For example, the duty to clothe and feed others extends only to our dependent children, the duty to warn or rescue others extends only to those with whom we have some “special relationship,” and the duty

---

258 The view that duty often serves to limit negligence liability is relatively uncontroversial. See, e.g., Friedman, supra note 78, at 470, 474 (explaining that judges have used duty to keep negligence decisions out of the hands of the jury in order to limit the liability of burgeoning enterprise); Peter F. Lake, Common Law Duty in Negligence Law: The Recent Consolidation of a Consensus on the Expansion of the Analysis of Duty and the New Conservative Liability Limiting Use of Policy Considerations, 34 San Diego L. Rev. 1503, 1525 (1997) (explaining the expanded reach of duty as a means of effecting a “conservative” vision of the proper extent of liability). I therefore focus on the claim that proximate cause, as well as duty, is relational.

259 See Palsgraf, 162 N.E. at 103 (Andrews, J., dissenting).

260 See, e.g., Harper v. Herman, 499 N.W.2d 472, 474 (Minn. 1993) (holding that a boat owner did not have a duty to warn a guest against diving into shallow water); Methola v. County of Eddy, 629 P.2d 350, 353–54 (N.M. Ct. App. 1981) (holding that jailors have a duty to protect and rescue inmates from other abusive inmates); Restatement (Second) of Torts § 314A (1965) (explaining that common carriers, innkeepers, land possessors,
to protect those on our property depends upon the status of our relationship with them.\textsuperscript{261} Other obligations we owe to all others. As a general matter, for instance, we must not intentionally punch anyone in the nose or drive a car so unreasonably that we crash into another. These “universal duties” might also be described as relational in the sense that they too are owed to a defined class of persons, albeit an inclusive one.\textsuperscript{262}

Even conceding that duty is relational, however, one need not conclude that the relational aspect of negligence liability rests exclusively with duty. Nor is it self-evident that duty is the most effective tool for linking the defendant’s wrong with the plaintiff’s injury. Proximate cause also plays a role in establishing relation—a role that neatly overlaps the work performed by duty and by plaintiff-foreseeability in \textit{Palsgraf}.\textsuperscript{263} A comparison of \textit{Palsgraf}’s relational duty analysis with the “risk rule” for proximate cause illustrates the point.

Recall that a relational duty inquiry consists of two steps: (1) whether the defendant owed a duty \textit{to the plaintiff} or to a class of people of which plaintiff is a member, and (2) whether it was \textit{this duty} that the defendant allegedly breached.\textsuperscript{264} Proximate cause, as defined by the risk rule, asks whether the plaintiff’s injury was “different from the harms whose risks made the actor’s conduct tortious.”\textsuperscript{265} At least with regard to plaintiff-foreseeability, these duty and proximate cause inquiries pose the very same question: was the defendant’s act wrong-

\begin{itemize}
\item \textsuperscript{261} See supra note 56 and infra note 308 (citing sources that explain the distinction in many jurisdictions between duties owed to licensees, invitees, and trespassers).
\item \textsuperscript{262} Goldberg & Zipursky, supra note 225, at 707.
\item \textsuperscript{263} Factual causation also requires a relational link (in the form of causation) between the defendant’s wrong and the plaintiff’s injury.
\item \textsuperscript{264} See supra note 226 and accompanying text. Actually, the duty inquiry requires a third step as well—the question of if a duty was owed, what is the general scope of that duty? Was it a general duty to use reasonable care, or some more specific duty?
\item \textsuperscript{265} \textsc{Restatement (Third) of Torts: Liab. for Physical Harm (Basic Principles)} § 29 (Tentative Draft No. 3, 2003); see also Robert E. Keeton, Legal Cause in the Law of Tort 10 (1963) (“A negligent actor is legally responsible for the harm, and only the harm, that . . . is a result within the scope of the risks by reason of which the actor is found to be negligent.”); Ripstein, supra note 113, at 65 (“[L]iability is limited to the risks that make the conduct wrongful.”).
\end{itemize}
ful in relation to the plaintiff’s injury? Consider the following explanation of the reasoning involved in each. In acting, a defendant created a range of potential risks—picture this range as a circle, within which any given point represents a potential risk. In deciding that the defendant’s act was wrongful, the jury draws a smaller circle within this range. The smaller circle represents the jury’s normative judgment regarding which of the potential risks were reasonably foreseeable and avoidable. Every point within the smaller circle represents a risk that the defendant ought to have taken into account when acting. Every point outside of the smaller circle represents a risk that the reasonable person would not have taken action to avoid. In the context of Palsgraf, for example, the railroad employees ought to have considered that in helping the man onto the train they might bruise him or cause him to fall. This risk of injury is therefore within the circle of risk that made the employees’ conduct wrongful. Similarly, the employees ought to have considered that they might cause his package to drop and its contents to break. Perhaps they should even have considered the risk of jostling nearby passengers. The risk of explo-

266 At least one relationalist appears to admit as much. In a provocative article in which he ascribes to Palsgraf’s duty-risk rule the weight of a substantive standing requirement, Professor Zipursky states that:

[I]t appears that the risk rule can be understood as a conflation of the substantive standing rule with some form of causation requirement. The risk rule presupposes that the conduct must have involved the wrongful taking of a risk of certain injuries to the plaintiff—in other words, negligence in relation to the plaintiff. To the contrary, such is precisely part of the relation that the risk rule seeks to establish. See Ripstein, supra note 113, at 65, 67–68 (explaining the relationality of negligence liability, and more specifically the “ownership” of risk, in terms of the risk rule: “[I]f we understand the boundaries between persons as existing in moral rather than geometric space, we see that the fault system can coherently hold someone liable only for the losses that are within the risk implicit in the violated standard of care.”); Perry, Moral Foundations, supra note 94, at 484 (“Risk is the relational concept that connects the active and passive aspect of injurious conduct, so that what the defendant did and what the plaintiff suffered are not regarded as two discrete happenings.”) (quoting Ernest J. Weinrib, Right and Advantage in Private Law, 10 Cardozo L. Rev. 1283, 1304–05 (1989)); Weinrib, supra note 94, at 521 (“The concepts of proximate cause and duty connect wrongful doing to wrongful suffering by requiring the plaintiff’s injury to be the fruition of the unreasonable risk that renders the defendant’s action wrongful.”).

267 In terms of foreseeability, there are perhaps two smaller circles within the range of risks created by the railroad employees’ action. There is a smaller circle representing the foreseeable risks and an even smaller circle still representing the reasonably foreseeable risks. See supra notes 105–08 and accompanying text.
sion, on the other hand, or the risk that their act might topple distant scales, or the risk that they might hurt a passenger as far away as Mrs. Palsgraf all are risks that would not have been considered by the ordinary person. In the language of the risk rule, such risks were not what made the employees’ act wrongful. In the language of Judge Cardozo, the defendant breached no duty with respect to such risks. In either case, such risks were not reasonably foreseeable.

Each of the false dichotomies described above indicates that proximate cause standing alone might fulfill the component of relationality often currently decided in the context of duty by plaintiff-foreseeability. This is not, of course, to say that the elements of duty and proximate cause are redundant. The question central to duty remains whether the defendant owed some obligation of care. The essence of proximate cause remains whether the plaintiff’s injury exceeds the scope of the risk that the defendant created in breaching a duty. Furthermore, although both elements involve some aspect of the relation between the defendant’s wrong and the plaintiff’s injury, there is a fundamental difference in the level of generality from which each element approaches the question. The existence of a duty is a threshold question. It is the gatekeeper for negligence liability, including or excluding potential liability with regard to broad groupings of fact patterns. Duty questions and the relational inquiries relevant to them are therefore, to the extent possible, acontextual. Proximate cause, on the other hand, serves as the final and finest sieve. Whereas duty examines the relation between broad categories of defendants and plaintiffs, wrongs and injuries, proximate cause considers relation in the full dress of context.

268 Palsgraf, 162 N.E. at 100.

269 I offer this discussion of the similarity between the duty-risk nexus and the risk rule for the purpose of showing that proximate cause is equally as capable of establishing this aspect of relation as is duty. I would contend that the risk rule is virtually identical in effect as a generalized foreseeability inquiry properly formulated, although a full discussion of the point exceeds the scope of this Article. For a rigorous, if ultimately flawed, critique of the risk rule, see Heidi M. Hurd & Michael S. Moore, Negligence in the Air, 3 THEORETICAL INQUIRIES IN LAW 333, 365–411 (2002) (criticizing the risk rule on grounds that (1) because there is no “correct” description of a risk, the risk rule is arbitrary; (2) because every risk is considered in determining fault, the risk rule leads to liability for all resulting harms; and (3) the risk rule is incompatible with other rules such as superseding cause, the thin-skull rule, and the denial of recovery for harms that are remote in spatio-temporal terms). For a compelling response to these arguments, see Israel Gilead, Harm Screening Under Negligence Law 14–17 (Sept. 26, 2005) (unpublished manuscript, on file with the author).
The foregoing account admittedly leaves two pieces of the foreseeability puzzle unanswered. First, if the only meaningful distinction between the relational inquiries of duty and proximate cause is the level of generality at which they are made, then why not a scheme in which foreseeability plays a role in both elements, differing only in the generality of its approach? It is to this question that I turn in Part III.D, immediately below. Second, even if proximate cause can accommodate the relational determinations often currently rendered in the context of duty, can a relational view of duty exist without the looking glass of foreseeability? Although a complete examination of this question must be left to future work, my intuition is that the answer is yes. In an every-day sense, the essence of an affirmative caretaking relationship between two people is not that one of them might become injured. Rather, such a relationship is the result of a voluntary commitment from one to the other, an inclusion of the other within one’s sphere of influence—the mother’s exercise of her parental rights, the innkeeper’s invitation to stay the night, the doctor’s promise to treat, the police officer’s promise to guard, the driver’s non-negligent creation of a driving hazard, or the rescuer’s attempt to rescue. In each of these situations, the actor’s conduct is a self-initiated curtailment of her own liberty, a commitment—whether actual or implied—to a relationship with another. It is this commitment, and not foreseeability, that provides the foundation for a court’s examination of relation and obligation.

D. Categorical Foreseeability

What about a theory of duty and proximate cause according to which foreseeability plays a role in both elements, differing only in the generality of its approach? Such an understanding is not uncommon. In duty, so the theory goes, courts determine whether the alleged class of plaintiff, risk, or type of injury is generally foreseeable to the class of people of which the defendant is a member. Such a theory is often limited to the context of plaintiff-foreseeability, leaving injury-foreseeability to proximate cause. David W. Robertson et al., Cases and Materials on Torts 199–200 n.9 (3d ed. 2004); see Goldberg & Zipursky, supra note 18, at 1818–20, 1828 (explaining that the duty inquiry focuses on whether the relevant class of defendants owes a duty to the relevant class of plaintiffs to act in accordance with a certain standard of care). Courts frequently, however, decline to impose a duty after concluding that a class of risk or injury is not generally foreseeable to a particular class of defendants. See, e.g., Ballard v. Uribe, 715 P.2d 624, 628 n.6 (Cal. 1986) (“The foreseeability of a particular kind of
the court must dismiss the case for lack of duty. If so—and if other duty-relevant factors mitigate in favor—then the defendant will be deemed to have owed the alleged duty. The only question remaining for proximate cause, under such an approach, is whether the particular plaintiff, risk, or injury was sufficiently unforeseeable within the factual context of the case as to preclude liability. In other words, foreseeability in the context of duty is categorical, whereas foreseeability in proximate cause is individualized and context-dependent. Although such an approach addresses the conceptual requirements of negligence in some workable fashion, I urge in this section that it should be abandoned nonetheless.

In considering the validity of this or any other conception of duty in which categorical foreseeability plays a starring role, three aspects of categorical foreseeability require separate attention: (1) the categorical incapacity to foresee, (2) categorical unforeseeability as grounds not to impose a duty, and (3) the imposition of an affirmative obligation due to categorical higher-than-average foreseeability. I discuss each in turn.

1. Categorical Incapacity to Foresee

In certain limited circumstances, courts have held that a class of defendants’ mental or physical capacity to foresee is so undeveloped or diminished that a defendant within such a class cannot be said to owe a duty of care at all. Thus, most courts hold that defendants below a certain age presumptively are not liable in negligence. Similarly, a driver who unexpectedly is rendered unconscious is not held liable for the resulting damages. One might characterize such harm plays a very significant role in this calculus, but a court’s task—in determining ‘duty’—is not to decide whether a particular plaintiff’s injury was reasonably foreseeable in light of a particular defendant’s conduct, but rather to evaluate more generally whether the category of negligent conduct at issue is sufficiently likely to result in the kind of harm experienced that liability may appropriately be imposed on the negligent party.” (citation omitted).

Determining the relevant class presents the same potential for indeterminacy as does the characterization of the relevant reference class of events for purposes of deciding foreseeability generally. See supra notes 109–11 and accompanying text (explaining this indeterminacy).

See supra note 159 and accompanying text.

See supra note 109 and accompanying text.

See Bashi v. Wodarz, 53 Cal. Rptr. 2d 635, 638 (Ct. App. 1996) (citing a line of cases holding as much). In addition, some forms of incapacity lead courts to impose a lesser standard of care, rather than to deny duty altogether. For example, a blind defendant is generally held to an obligation to use only the care of a reasonable blind person. E.g. Roberts v. Louisiana, 396 So. 2d 566, 568 (La. Ct. App. 1981).
bright-line rules as judgments regarding the categorical unforeseeability of injury or plaintiff to such classes of defendants. Because any risk, injury, or potential plaintiff is categorically unforeseeable to a four year-old, for example, the child cannot owe a duty of care.

As explained in Part III.A above, however, such rules are not best described as foreseeability determinations at all, even categorical ones. The rules limiting the negligence liability of children or unconscious drivers are not conditioned upon any category of risk, injury, or plaintiff, but rather depend solely on the physical and mental characteristics of the class of defendant. Unlike even categorical foreseeability determinations, incapacity decisions are completely divorced from any other factual context.

Furthermore, an explanation of incapacity rules that turns on foreseeability faces a serious descriptive anomaly. Courts generally impose an ordinary duty of care upon mentally incapacitated but not upon physically incapacitated defendants, although the potential consequences of their acts may be equally unforeseeable to members of either class. In light of this inconsistency, perhaps the justification behind incapacity rules is not categorical unforeseeability, but rather the determination that unlike the mentally incapacitated, neither the physically incapacitated nor the very young can be deemed to have acted at all.

275 See supra note 160 and the text of the surrounding paragraph.

276 One might, of course, characterize such determinations as relational in a way that is similar to a relational description of the general duty of care to avoid causing harm to another. Such decisions embody the judgment that a class of risk, injury, or plaintiffs described as “any” is unforeseeable to a class of defendants. Even so, however, the distinction between a class defined as “all the world” and some more contextually limited class is an important one in sorting out the proper doctrinal home for foreseeability.

277 This distinction is not only descriptively accurate, but it also supports the argument that categorical incapacity, but not categorical unforeseeability, is properly decided as a matter of duty. The more context-independent a decision, the better the case for leaving it in the hands of the court; the more context-dependent, the more reason to send it, in the first instance, to the jury. See Cardi, supra note 7, at 801–05 (explaining why juries are the better entity for deciding fact-dependent inquiries).

278 Kelley, supra note 85, at 237. One possible explanation for this seeming discrepancy is society’s general apathy toward the mentally ill. See id. at 205 (“The generic law review article often goes on to suggest that the different treatment of the mentally ill reflects outdated and unjustified prejudice against them.”). This fails to explain why negligence liability for the mentally ill has become stricter since the 1800s, see id. at 183–203 (describing the law’s historical development), while society’s understanding and tolerance of mental illness has presumably increased.

279 See id. at 238 (“When the defendant driver fainted without forewarning and his unguided car drifts into the left lane the driver has not acted in violation of the safety convention because at the time his car crossed into the left lane he was not acting at all.”).
In short, although it is practical and conceptually proper that courts set specialized duty rules based on categorical judgments of incapacity, such judgments are not dependent upon foreseeability. Hence, the existence of categorical incapacity rules does not afford foreseeability a foothold in duty.

2. Categorical Unforeseeability as Grounds Not to Impose a Duty

In most jurisdictions, the host of a non-commercial social event may serve alcohol without fear of liability for injuries caused by a drunken guest. Typically, courts have barred social host liability on the grounds that a host owes no duty to an intoxicated guest or to third parties injured thereby, either because social hosts as a class cannot foresee how much alcohol a guest will consume and cannot know when a guest has become intoxicated, or because a host cannot foresee or control an intoxicated guest’s conduct. Such reasoning seems disingenuous, however, in light of the fact that a social host often know that a guest has imbibed excessively and is able to foresee, for example, that the guest might harm herself or others while driving home. This seeming contradiction might be generalized as follows. The denial of a duty due to categorical unforeseeability is tantamount to holding that even if the defendant was not vigilant, the defendant’s lack of vigilance would not be blameworthy because the class of risk, injury, or plaintiff is generally unforeseeable to the defendant-class. If the risk, injury, or plaintiff was in fact foreseeable to the particular defendant, however, the stated reason for refusing to impose a duty is gone. The categorical no-duty rule thus stands on quicksand, and some unstated justification must therefore underlie these rulings.

Similarly, as anyone with experience with young children can attest, it is sometimes difficult to say whether any particular movement of limb is deliberate or the result of some random firing of maturing neurons. At any rate, even if many of a child’s actions do seem deliberate, a court might yet wish to steer clear of the thicket of determining which are which.

280 See, e.g., McGuiggan v. New Eng. Tel. & Tel. Co., 496 N.E.2d 141, 142 (Mass. 1986) (“Under traditional common law tort analysis, our inquiry is whether a social host violated a duty to an injured third person by serving an alcoholic beverage to a guest . . . . [which] require[s] . . . consider[ation of] whether the social host unreasonably created a risk of injury . . . .”); Gilger v. Hernandez, 997 P.2d 305, 310–12 (Utah 2000) (considering whether the social host-guest relationship gives rise to a special relationship that imposes on the social host an affirmative duty either to control or protect her guests).

One possibility is that categorical unforeseeability serves as a vessel for hidden decisions of policy; in the social host context, for example, courts may simply wish to protect the institution of social gatherings at which alcohol is served.\textsuperscript{282} Another possibility is that in deciding such cases based on duty, courts unwittingly engage in what is commonly understood to constitute breach or proximate cause analysis without the requisite deference to the jury.\textsuperscript{283} I have discussed the hazards of each of these possibilities elsewhere.\textsuperscript{284}

Two further potential justifications exist, however. First, categorical unforeseeability rulings might serve as a tool to control transaction costs. A court might reason, for example, that because the incidence of social host cases in which a risk, injury, or plaintiff is foreseeable is extraordinarily low, identifying such cases is not worth the requisite expenditure of judicial resources. This reasoning is, of course, not limited to social host cases; any case in which the court declines to impose a duty on grounds of categorical unforeseeability might thus be justified. By dismissing such cases based on categorical unforeseeability, courts arguably avoid a likely unfruitful context-specific foreseeability inquiry and discourage future analogous suits. A court is, of course, empowered to make such a decision, although it is unclear just how much cost is saved thereby.\textsuperscript{285} More important, however, is the gravity of a decision to sacrifice the interests of a plaintiff (and of future plaintiffs) whom another's wrong has foreseeably injured simply because it would be too costly over time to sort the wheat from the chaff. At the very least, courts should be clear that such is the basis for their rulings—that they are denying liability in a class of cases not because the defendant was not blameworthy, but be-

\textsuperscript{282} See Kelly, 476 A.2d at 1227–29 (discussing the opposing policy interests in social host cases); Cardi, supra note 7, at 763–65, 787–90 (same); Linda E. Fisher, Alcohol, Tobacco, and Firearms: Autonomy, the Common Good, and the Courts, 18 Yale L. & Pol'y Rev. 351, 370–73 (2000) (same).

\textsuperscript{283} Cardi, supra note 7, at 774–81.

\textsuperscript{284} See id. at 790–804 (arguing that the former is improper to the extent that foreseeability obscures the underlying policy decision and that the latter is improper because it robs the jury of its rightful role in deciding negligence and proximate cause).

\textsuperscript{285} For example, a court must still determine whether a particular case fits within an established no-duty category—an inquiry that may yet require fact-finding by the jury. Furthermore, it is uncertain to what extent no-duty decisions discourage future lawsuits; attorneys are trained to distinguish their cases from potentially controlling precedent, a task made especially inviting by the particularly fact-specific nature of negligence cases.
cause they have decided not to consider the question for administrative convenience. As a final potential justification for categorical unforeseeability, Professors Goldberg and Zipursky suggest that by excluding liability categorically, courts prioritize certain obligations over others and provide clearer rules for actors. Although Professors Goldberg and Zipursky do not explicitly state as much, I understand their point to be that categorical foreseeability is useful insofar as it encourages actors to prioritize their efforts to take care where doing so will have some reasonable chance of avoiding injury. Only classes of scenarios in which the risk, injury, and plaintiff are “foreseeable enough” are worthy of a “prioritized” duty of care. Professor Goldberg and Zipursky’s account represents a plausible but exceedingly strong view of the judge’s role in negligence cases—a view of the judge as frequent rule-maker, as micromanager. This was the view of Justice Holmes as he imposed specific rules for railroad-crossings in *Baltimore & Ohio Railroad Co. v. Goodman*, an approach resoundingly (if respectfully) rejected seven years later by Justice Cardozo in *Pokora v. Wabash Railway Co.* I concede that in the decades since *Pokora* (and especially since the early 1980s), courts seem to be sliding back to the idea that they must keep a tighter

---

286 See Calabresi, supra note 213, at 250–51, 255–56 (arguing that the largest administrative cost of the tort system is the case-by-case determination of accidents by juries).

287 See Goldberg & Zipursky, supra note 18, at 1831–32 (articulating the value of the prioritizing effect of categorical duties). As they explain:

If one has a duty of care to another, that other person figures (or should figure) in one’s deliberation in a certain way. Because the possibility of duty serving a prioritizing role is compromised by casting the duty net too wide, the questions arises as to which types of persons are obligated to be vigilant to avoid causing certain types of harm to certain others.

288 275 U.S. 66, 70 (1927) (holding categorically that “if a driver cannot be sure otherwise whether a train is dangerously near he must stop and get out of his vehicle”). In so holding, Justice Holmes acted upon his longstanding view that judges should more actively apply their cumulative wisdom to prescribe specific duties in categories of cases. See id. (“It is true . . . that the question of due care very generally is left to the jury. But we are dealing with a standard of conduct, and when the standard is clear it should be laid down once for all by the Courts.”) (citation omitted); Glen O. Robinson & Kenneth S. Abraham, *Collective Justice in Tort Law*, 78 Va. L. Rev. 1481, 1485 (1992) (noting that Justice Holmes’s *Goodman* opinion was an attempt to implement the theories he had propounded in *The Common Law*).

289 292 U.S. 98, 106 (1934) (holding that it was for the jury to decide whether the defendant’s conduct was unreasonable, where the layout of a grade crossing was such that to exit one’s car to check for oncoming trains—as was one’s duty pursuant to *Goodman*—would be to increase, rather than to reduce, the risk of catastrophe).
rein on duty and hence on the jury’s power to decide negligence. In
doing so, however, courts sacrifice justice in the individual case in
which risk, injury, or plaintiff in fact was foreseeable.290

Furthermore, it has not been demonstrated that the particular-
ized guidance of court-made rules is necessary or effective in prioritiz-
ing people’s efforts at care. Nor is there any evidence that particular-
ized duties are normatively superior in this respect to the generalized
duties to avoid unreasonably causing harm to another or to take rea-
sonable precautions to protect another with whom one has a special
relationship.291 Scant sociological research exists to support such a
hypothesis,292 and the comparative merit of rules to standards gener-
ally is a matter of perpetual jurisprudential debate.293 Without evi-
dence that particularized rule-guidance is preferable, the long-
deliberated policy judgments embodied by the general duties of care
should stand.

Finally, one might question the institutional ability of courts to
render categorical unforeseeability determinations. The categorical
unforeseeability inquiry consists of three steps: (1) a description of
the relevant categories of defendant and risk, injury, or plaintiff, (2) a
finding that the incidence of foreseeability is lower for the applicable
pair of categories than it is for society generally, and (3) a normative
judgment that the incidence is sufficiently low to justify the denial of
liability. Although the last of these determinations is arguably within
the court’s bailiwick, the second entails a factual investigation perhaps
best accomplished by a legislature or an administrative agency, and a

290 Cardi, supra note 7, at 798–804.

291 Cf. Ripstein, supra note 113, at 105 ("[A]gents need to be able to know which ac-
tions can be performed without fear of legal sanction, and, more to the point, the interests
of others of which they must take account. The limitation of liability to foreseeable injuries
is an expression of the idea that people must be in a position to know their rights and the
rights of others. Reasonable foreseeability is required . . . because it is public in the right
way, that is, accessible in principle both to those who might injure others and those who
might be injured by them.") (emphasis omitted).

292 This conclusion is based upon conversations with two sociology professors at the
University of Kentucky. As of publication, no specific sources for this conclusion could be
identified.

in the Making and Application of Law 138–41 (1994) (discussing generally the relative
value of rules versus standards). See generally Ronald Dworkin, Taking Rights Seriously
(1977) (challenging the primacy of rules over principles and proposing a structure for
their co-existence); Frederick Schauer, Playing by the Rules (1992) (advocating the
law’s reliance on specific rules rather than on generalized standards).
description of the relevant categories is rendered less indeterminate, if at all, only by reliance on the jury.  

In sum, there are no dispositive conceptual justifications for courts’ use of categorical unforeseeability as a reason to deny the existence of a duty. The justifications offered by courts and scholars turn out to be merely instrumental—that categorical unforeseeability serves administrative convenience or better guides actors’ behavior. Furthermore, the objects of such instrumentalism appear to be normatively irresolute. Without good reasons to embrace categorical unforeseeability, it therefore should be abandoned.

3. Imposition of an Affirmative Duty Due to Categorical Higher-Than-Average Foreseeability

As many courts look to categorical unforeseeability as a reason to deny duty, courts also treat the particular ability of a class of defendants to foresee a class of risk, injury, or plaintiff as a reason to impose an affirmative duty. The quintessential example of this affirmative use of categorical foreseeability is the California Supreme Court’s decision in *Tarasoff v. Regents of the University of California*. In that case, the court imposed on the defendant psychologist a duty to take reasonable care on behalf of the decedent Tatiana Tarasoff, a non-patient, with regard to the threats on her life made by one of the psychologist’s patients. The court’s duty analysis did not turn on the particular context of the case. Rather, the court imposed a duty in part due to the special relationship between psychologist and patient, but also because psychologists as a class are often privy to their patients’ potentially injurious intentions toward third parties.

---

294 See supra notes 109–111 and accompanying text. It is interesting that although Professors Goldberg and Zipursky explicitly endorse categorical foreseeability decisions, they decry intrusions into the realm of the legislature when discussing the duty-as-policy approach of Justice Holmes and Dean Prosser. See Goldberg & Zipursky, supra note 18, at 1739–44 (assessing the multi-factored “policy-driven” approach to duty as “arbitrary, indeterminate, and doctrinally unstable” and concluding that “[o]ur understanding of the relative strengths and weaknesses of political institutions often leads to the conclusion that the legislative and executive branches are more capable, or at least more appropriate, institutions for making such decisions”). I would contend that categorical foreseeability decisions are similarly suspect.


296 Id. at 340.

297 Id. at 343–45. Although the court stated that it was not necessary to “decide whether foreseeability alone is sufficient to create a duty to exercise reasonable care to protect a potential victim of another’s conduct,” id. at 343, the court’s analysis of the special relationship between psychologist and patient turned in part on the fact that psy-
In other words, just as a low incidence of class-foreseeability might lead a court to deny the existence of a duty, a higher-than-ordinary incidence of class-based foreseeability might be reason to impose a duty of care where otherwise none would exist.\(^{298}\) This conclusion implicitly stems from the following reasoning. In the ordinary case, courts refuse to impose an affirmative duty to warn or protect another from danger posed by a third source\(^ {299}\)—a rule generally understood to arise from courts’ concern for individual liberty.\(^ {300}\) Where a particular class of defendant has a higher-than-ordinary likelihood of foreseeing a class of risk, injury, or plaintiff, the proportionately greater opportunity for accident-prevention afforded by imposing an affirmative duty may outweigh the countervailing liberty interests of the defendant class.\(^ {301}\)

Furthermore, unlike no-duty holdings based on categorical un-foreseeability, the affirmative use of categorical foreseeability does not forgo a context-specific foreseeability inquiry. Consider, for example, if the \textit{Tarasoff} court had held that a psychologist owes no duty because psychologists are uniquely privy to knowledge of dangers posed by their clients. Benjamin C. Zipursky, \textit{The Many Faces of Foreseeability}, 10 Kan. J.L. & Pub. Pol’y 156, 157–58 (2000).

\(^{298}\) See Goldberg & Zipursky, \textit{supra} note 18, at 1838–39 (“The ease or difficulty for persons in the defendant’s category to anticipate those harms is relevant to whether it makes sense for such persons to be said to have a duty to be vigilant against causing them. Hence, . . . the decision that certain defendants are particularly well-situated to foresee the sort of harm that befell the plaintiff is not only relevant to whether there was a breach. It is also relevant to whether a category of defendant may properly be declared to owe a duty of due care to a category of plaintiff.”).

\(^{299}\) \textit{Restatement (Second) of Torts} § 314 (1965); Dobbs, \textit{supra} note 2, § 314, at 853.

\(^{300}\) E.g., Epstein, \textit{supra} note 117, at 197–98; \textit{see also} Ripstein, \textit{supra} note 113, at 91 (“The fact that you are in peril, and I know of your peril, does not make that risk mine. . . . To shift risk in this way would be unduly burdensome to liberty . . . .”); Arthur Ripstein, \textit{The Division of Responsibility and the Law of Tort}, 72 Fordham L. Rev. 1811, 1840 (2004) (explaining that defendants’ nonfeasance does not typically generate a duty because “[a]s private parties, they cannot be under any corresponding obligation to confer any benefit on me, no matter how significant the benefit, and no matter how easy it is to confer. Such an obligation would undermine the sense in which what they have is their own”). \textit{See generally} Steven J. Heyman, \textit{Foundations of the Duty to Rescue}, 47 Vand. L. Rev. 673 (1994) (offering a range of possible justifications).

\(^{301}\) In economic terms, suppose that courts would prevent 100,000 injuries a year, valued at $100,000 each, by imposing on society generally a duty to protect, warn, or rescue. By the absence of such a general affirmative duty, we might presume the conclusion that society’s combined interest in liberty is worth more than the combined value of 100,000 yearly injuries: \((300,000,000 \text{ people} \times \$100 \text{ yearly liberty interest per person} = \$30,000,000,000) > (100,000 \text{ yearly injuries} \times \$100,000 \text{ per injury} = \$10,000,000,000)\). Suppose now that imposing an affirmative duty on 10,000 psychologists would prevent 100 injuries per year. The value of these 100 yearly injuries may well outweigh the combined liberty interest of the psychologist class: \((10,000 \text{ psychologists} \times \$100 \text{ yearly liberty interest per person} = \$1,000,000) < (100 \text{ yearly injuries} \times \$100,000 \text{ per injury} = \$10,000,000)\).
generally it is not foreseeable that a psychologist’s patient may pose a risk to a third party. In such a case, psychologists would escape liability even where a jury might have found that the risk was indeed foreseeable (or, as in Tarasoff, actually foreseen). As noted above, such a result would be nonsensical. A court’s affirmative use of class-based foreseeability does not elicit this problem, however. Although the Tarasoff court imposed a duty based in part on the relatively high incidence of psychologists’ foresight, the jury remained free to deny liability in the context of breach or proximate cause because the risk, injury, or plaintiff was not foreseeable under the circumstances.

Despite the promise of categorical foreseeability, however, it proves conceptually unstable. Specifically, the affirmative use of categorical foreseeability is at odds with the canonical distinction between misfeasance and nonfeasance. In deciding that a psychologist’s special ability to foresee is enough to override the general no-duty-to-rescue rule (and its underlying principles), it is difficult to argue that the ability of others to foresee injury should not also be sufficient. For example, suppose that a small stretch of public street outside a haberdashery becomes icy due to a combination of inclement weather and a peculiar dip in the road. Suppose also that the haberdasher witnesses several accidents throughout the day and does nothing to prevent them. The haberdasher is surely more likely than the ordinary person to foresee the risk that passing cars might slide on the ice—and yet the haberdasher will not owe a legal duty to warn those in danger. The difference between the psychologist and the haberdasher is not the incidence with which each might foresee the relevant class of harms. Psychologists are surely no more privy to impending doom than is a class defined as “people who work in proximity to an observed potential road hazard.” With no greater potential for accident-prevention, the categorical foreseeability of psychologists provides no greater ballast against a countervailing liberty interest than does the categorical foreseeability of the haberdasher. Indeed, along this line, a few scholars have urged that the logical extension of Tarasoff leads to abrogation of the traditional distinction between misfeasance and nonfeasance in favor of an approach to duty based almost entirely on foreseeability.

---

302 John M. Adler, Relying upon the Reasonableness of Strangers: Some Observations About the Current State of Common Law Affirmative Duties to Aid or Protect Others, 1991 Wis. L. Rev. 867, 911–14 (1991) (proposing the abandonment of the misfeasance/nonfeasance distinction and suggesting that such a change is necessitated by the same forces that led to California duty cases such as Tarasoff); Lake, supra note 4, at 121–23 (characterizing Tarasoff as having
One might attempt to distinguish Tarasoff from the case of the apathetic haberdasher according to the ease with which a court can identify the psychologist class in advance. Such an assertion is not necessarily accurate, however. Although it might be a relatively straightforward inquiry for a court to evaluate psychologists’ ability as a class to discern a patient’s potential for violence, the foreseeability-related attributes of “business owners in proximity to potential road hazards” or even more generally, “people who are witness to an obvious risk” seem at least as readily apparent. In addition, although administrative ineffectiveness often righteously serves as a reason not to impose a duty, the ease with which a court might identify a class is not so clear a justification on its own to impose a duty. A court cannot nullify the liberty interests of an entire class of defendants simply because it is easier to identify those defendants than others—there must be some underlying substantive reason to do so. Administrative ease in such a case only has substantive content as compared to administrative difficulty.

All of this is not, of course, to say that the Tarasoff court was wrong to impose an affirmative duty on psychologists to warn third parties of risks posed by patients. The court’s decision might, for example, be justified by a perceived need to stem a growing tide of injuries caused by psychological patients. Or, as the Tarasoff court reasoned, the therapist-patient relationship might be sufficiently custodial that the therapist must protect the patient not only from harm but also from harming others. The imposition of an affirmative duty cannot, however, be justified by categorical foreseeability.

The conclusion to be drawn from the foregoing analysis of the three subspecies of categorical foreseeability—the capacity to foresee, strengthened the prominence of foreseeability in duty decisions); Murphy, supra note 4, at 175–76 (arguing that the duty analysis has gradually embraced a duty rule turning primarily on an analysis of foreseeability, an evolution culminating in Tarasoff). Indeed, there can be no doubt that Tarasoff has extended its influence beyond the realm of affirmative duties. See, e.g., Randi W. v. Muroc Joint Unified Sch. Dist., 929 P.2d 582, 588–91 (Cal. 1997) (applying Tarasoff factors, chief among which is foreseeability, to an action for negligent misrepresentation—a misfeasance claim—in which a school recommended an ex-teacher for hire despite its knowledge of past complaints of sexual misconduct).

303 See supra note 61 and accompanying text.

304 Such would be functionally equivalent to making theft a crime only for those with a low IQ because such people are easier than others to catch.

305 See 551 P.2d at 344 (“[B]y entering into a doctor-patient relationship the therapist becomes sufficiently involved to assume some responsibility for the safety, not only of the patient himself, but also of any third person whom the doctor knows to be threatened by the patient.”).
categorical unforeseeability, and categorical increased foreseeability—is obviously unique among the subsections of Part III of this article. Rather than an argument that categorical foreseeability might fit within breach or proximate cause instead of duty, I have urged that categorical foreseeability is (1) normatively deficient, and/or (2) a façade for some other duty-relevant determination. I now turn to a more robust discussion of the latter as a distinct conceptual use for foreseeability.

E. Foreseeability as Proxy

In a variety of types of cases, courts use foreseeability as a convenient tool by which to limit liability in order to reflect community notions of obligation or for reasons of policy. Properly understood, this use of foreseeability is unproblematic. To the extent that a court imposes atypical boundaries on a jury’s determination of foreseeability in order to effect a policy-based limitation on liability, such a determination lies squarely within the province of the court to delineate the standard of care or to define the legal standard for proximate cause. It would be a mistake, however, to characterize such holdings as “foreseeability decisions.” Courts in such cases do not decide foreseeability at all, but leave (or at least properly leave) the actual foreseeability determination to the jury. Instances of this conceptual use of foreseeability may be found in cases alleging emotional harm, economic harm,\textsuperscript{306} economic harm,\textsuperscript{307}

\textsuperscript{306} See, e.g., Portee v. Jaffee, 417 A.2d 521, 526 (N.J. 1980) (limiting claims for bystander emotional distress, in relevant part, to cases in which the “plaintiff was located near the scene of the accident” as least partly because “as physical proximity between plaintiff and the scene of the accident becomes closer, the foreseeable likelihood that plaintiff will suffer emotional distress from apprehending the physical harm of another increases”). At least one court has gotten it right, however. See Gates v. Richardson, 719 P.2d 193, 196 (Wyo. 1986) (deciding to impose a duty to avoid negligent infliction of emotional distress and recognizing that foreseeability plays no role in such determination while noting that “[i]nstead of perpetuating the illusion, we prefer to set forth the legal duty and outline the policy principles which persuade us to recognize the legal duty and its limitations”).

\textsuperscript{307} See, e.g., Citizens State Bank v. Timm, Schmidt & Co., 335 N.W.2d 361, 366 (Wis. 1983) (explaining that, in a suit for economic harm, “[i]t is enough [to permit liability] that the maker of the representation intends it to reach and influence either a particular person or persons, known to him, or a group or class of persons, distinct from the much larger class who might reasonably be expected sooner or later to have access to the information and foreseeably to take some action in reliance upon it”) (quoting Restatement (Second) of Torts § 552, cmt. h (1965)).
landowner-liability, and other scenarios with regard to which a court wishes to limit liability in some at least ostensibly principled way. Consider the following representative case.

In Posecai v. Wal-Mart Stores, Inc., the plaintiff sued Wal-Mart after having been robbed in a Sam’s Club parking lot. Mrs. Posecai alleged that Wal-Mart was negligent for having failed to post a security guard outside of its store. Characterizing Mrs. Posecai’s case as alleging violation of an affirmative duty, the court summarized the various approaches of state courts in analyzing the existence of a duty on the part of a business owner to protect patrons from the crimes of third parties. As the court explained, many courts have imposed such a duty, limited in a variety of ways by foreseeability. In some jurisdictions, for example, a business owner owes a duty to protect patrons only if “he is aware of specific, imminent harm about to befall them.” In others, a duty exists only in light of “evidence of previous crimes on or near the premises.” In still others, the owner owes a duty to protect customers against any harm that is foreseeable under a “totality of the circumstances.” Each test represents a balance between the security interest of customers and the liberty interest of owners. Foreseeability is the means by which a court manifests the balance chosen. If a court decides that the balance should favor the

---

308 See, e.g., Carter v. Kinney, 896 S.W.2d 926, 928 (Mo. 1995) (describing the common-law duty standard owed by landowners to invitees to protect against foreseeable conditions and the duty to protect licensees only from conditions actually foreseen).

309 See, e.g., Langle v. Kurkul, 510 A.2d 1301, 1306 (Vt. 1986) (holding that a social host has a duty of care only in situations in which the host “furnishes alcoholic beverages to one who is visibly intoxicated and it is foreseeable to the host that the guest will thereafter drive an automobile”).

310 752 So. 2d 762, 764 (La. 1999).

311 Id. at 765.

312 Id. at 766. Such a conclusion is not foregone, however. By choosing to run a business in a potentially criminal neighborhood, perhaps Wal-Mart in fact created the risk of harm and therefore owed a general duty to run its business reasonably safely—a standard that might necessitate the posting of security guards.

313 Id. at 766–68.

314 Id. at 766.

315 Posecai, 752 So. 2d at 767.

316 Id.

317 Id.

318 Cf. Ripstein, supra note 113, at 50 (explaining the limited duty not to cause purely economic harm—a duty which often turns on foreseeability—as follows: “Only some forms of attachment to particular goods are protected; protecting all economic interests would place too great a burden on the liberty of others. If I could not act unless I was sure that your financial position would not be adversely affected, I could not act at all.”) (footnote omitted).
business owner, it might impose liability only where crime is actually foreseen. If a court is more sympathetic to the interests of customers, the court may define foreseeability according to a totality of the circumstances. The test restricting foreseeability to evidence of prior similar crimes represents a choice somewhere in between.\(^{319}\)

Defining the parameters of the foreseeability inquiry does not, however, constitute the inquiry itself. Rather, it is a means of delineating the scope of a business owner’s duty. Pursuant to the typical negligence standard of care—the duty to use reasonable care under the circumstances—a judgment regarding a business owner’s reasonableness would depend (in part) upon the degree to which crime on the premises of the business was foreseeable. The more foreseeable the risk of crime, the more likely the conclusion that the defendant’s failure to account for it was wrongful. In cases like Posecai, courts have altered this standard of care, conditioning it on a particularized notion of foreseeability. In such cases, the defendant’s reasonableness does not turn on foreseeability unmodified but instead on whether the owner was actually aware of the crime or whether crime was foreseeable in the cramped light of previous similar crimes. Delineating the standard of care thus is undoubtedly a task for the court within the context of duty.\(^{320}\) The decision whether a defendant’s conduct met that standard, however—including the judgment regarding to what degree risk was foreseeable, however foreseeability is defined—is the essence of the jury’s determination of breach.\(^{321}\) Hence, the actual foreseeability determination in cases such as Posecai properly is left to the jury.

Unfortunately, courts in such cases often adopt a different view of foreseeability’s role. Even if such courts agree that their various calibrations of foreseeability represent an attempt to strike an appropriate balance between the relative interests of the parties, they often view this exercise as part of their inquiry into duty’s existence, not as part of a delineation of the standard of care.\(^{322}\) The import of this perhaps subtle distinction is clear. If narrowing the definition of foreseeability is a matter of setting the appropriate standard of care, then the

---

319 One might also conceptualize these duty decisions in terms of a court’s appropriate recognition of community-based notions of obligation between business owners and their patrons. The variety of tests would embody courts’ sundry impressions of such notions.

320 Restatement (Second) of Torts § 328B (1965).

321 See id. § 328C.

322 See, e.g., Posecai, 752, So. 2d. at 765 (“The sole issue presented for our review is whether Sam’s owed a duty to protect Mrs. Posecai from the criminal acts of third parties under the facts and circumstances of this case.”).
actual foreseeability determination is left to the jury. If, on the other hand, the court conditions the very existence of a duty on foreseeability, then the court has co-opted the foreseeability determination from the jury. The court in Posecai took the latter approach, although there is not uniformity among the courts in this regard.

In sum, Posecai and cases like it represent courts' attempt to limit the liability of a class of defendants for a class of risk or injury not based on categorical foreseeability (as described in the previous subsection), but due to a non-foreseeability-related policy reason. Although a court might—as did the court in Posecai—thus choose to limit liability by conditioning the existence of a duty on a limited form of foreseeability, a court might also do so by redefining the standard of care or the legal standard for proximate cause. The only practical difference between the two approaches lies in whether the court or the jury renders the ultimate foreseeability determination. The effectiveness of foreseeability's use as a proxy thus need not depend on a court's determination of foreseeability within the confines of duty.

IV. Duty Without Foreseeability: A Conclusion and a Beginning

In this Article, I have highlighted the shortcomings of an account of duty that hinges upon determinations of foreseeability. I have also attempted to show that each of foreseeability's conceptual functions might instead operate wholly and seamlessly within the element of breach or proximate cause or both (or, in the case of categorical foreseeability determined by the jury).

323 See id. at 769 (concluding in light of all the evidence that "Sam's did not possess the requisite degree of foreseeability for the imposition of a duty to provide security patrols in its parking lot").

324 See, e.g., Nycal Corp. v. KPMG Peat Marwick LLP, 688 N.E.2d 1368, 1370 (Mass. 1998). In Nycal Corp. v. KPMG Peat Marwick LLP, the Massachusetts Supreme Judicial Court considered the viability of a third-party claim for pure economic loss resulting from alleged auditor malpractice. Id. at 1369. The court viewed the issue not as whether a duty existed, but rather as what was "the scope of liability of an accountant to persons with whom the accountant is not in privity." Id. at 1370. The court reviewed the various approaches taken by other courts, which ranged from allowing such claims pursuant to the typical reasonableness standard to limiting claims, at least in part, by narrowly defining the conditions under which injury to another might be found foreseeable. Id. at 1370–72. After picking a test, the court upheld summary judgment for the defendants with apparent deference to the jury's right, in the first instance, to determine foreseeability. Id. at 1373. Although Nycal represents the intersection of a number of knotty negligence issues—including a Palsgraf-like duty/proximate cause issue and the always puzzling intersection of tort and contract law—it is safe to say that the various approaches were necessitated by the fact that courts are more hesitant to extend liability for purely economic loss.
seeability, that negligence law would benefit from its disavowal altogether). It is my hope that the foregoing discussion will rekindle debate regarding the more important question—whether the court or the jury is the better institutional arbiter of foreseeability. This debate might rage on many fronts—whether judge or jury is more accurate, more just, more lawful, or more consistent, and perhaps at the center of it all, whether the benefits of clear, ex ante rules outweigh the concern for individualized justice.

Still, one question yet nags at my conscience—after all the dust of my reconstruction of foreseeability settles, what is left of duty? One might argue that even a flawed conception of duty is preferable to one gutted of foreseeability, an empty shell of the element of duty as many now understand it.

The basis for a foreseeability-free conception of duty is far from revolutionary, however. It stems from the proposition that at its core, duty consists of the following inquiry:

Assuming that the defendant’s actions were unreasonable and assuming that the plaintiff’s injury was within the scope of the risk that made the actor’s conduct unreasonable, should the court give legal force to an obligation on the part of the defendant to have acted reasonably?

With regard to foreseeability specifically, this initial duty inquiry might be translated, in relevant part, as follows:

Assuming that some risk of injury was reasonably foreseeable and assuming that the plaintiff’s injury in particular was reasonably foreseeable, should the court give legal force to an obligation on the part of the defendant reasonably to have foreseen it?

Understood pursuant to the reasoning set forth in Part III above, this general duty provision serves as a lens, filtering out questions that the law has decided to leave to the jury, unless incontestable, and focusing the duty inquiry instead on two prime considerations: (1) whether an obligation exists within the broadly generalized class of circumstances at issue and (2) whether courts should give legal cognizance to that obligation.

Of course, this duty standard leaves the ultimate question unanswered—it offers little guidance to courts regarding how to recognize an obligation or whether to gird an obligation with the force of law. It is not within the purview of this Article to set forth a comprehensive theory of how this ought to be done. My instinct is that duty is a com-
plex organism, which necessarily consists of a mix of the considerations listed in Part I.C above: community notions of obligation, a broad sense of social policy, concern for the rule of law, and administrative capability and convenience.\footnote{See supra notes 58–61 and accompanying text.} Furthermore, I resist the view that duty must exist either as a system of pure rule and principle or as an ad hoc balancing of instrumental policy goals.\footnote{See, e.g., Green, supra note 19, at 76–77 (proposing that duty decisions are the result of courts’ balancing of the following “factors”: administrative, ethical or moral, economic, prophylactic or preventative, and justice); Goldberg & Zipursky, supra note 18, at 1739–43, 1819 (characterizing the view of Justice Holmes and Prosser to be that duty amounts to little more than a multi-factored “policy-driven check on liability” and criticizing such an approach because it (1) is unprincipled, (2) usurps the power of the legislature, (3) obscures the rationale for judicial decisions, and (4) fails to provide clear rules for society); James A. Henderson, Jr., Expanding the Negligence Concept: Retreat from the Rule of Law, 51 Ind. L.J. 467, 468 (1976) (urging that “under all the circumstances” tests for liability are potentially devoid of meaning and incapable of supplying guidance). But see George C. Christie, The Uneasy Place of Principle in Tort Law, in PHILOSOPHICAL FOUNDATIONS OF TORT LAW, supra note 104, at 129 (suggesting, in the context of a discussion of the role of foreseeability in economic harm cases, that “we have asked too much of principles. For the appeal of principles to be useful, we must not only have more modest expectations of their role in the decision of tort cases but also perhaps a more expansive notion of what comes within the ambit of a principle”). See generally Gary T. Schwartz, Mixed Theories of Tort Law: Affirming both Deterrence and Corrective Justice, 75 Tex. L. Rev. 1801 (1997) (offering a reconciliation of the leading instrumentalist and non-instrumentalist understandings of tort law).} In my view, it is both.
OUTSOURCING POWER: HOW PRIVATIZING MILITARY EFFORTS CHALLENGES ACCOUNTABILITY, PROFESSIONALISM, AND DEMOCRACY

Martha Minow*

Abstract: Private contractors have played key roles in recent high-profile scandals. These scandals hint at the degree to which the U.S. military has increased the scope and scale of its reliance on private security companies in recent decades. This trend offers many advantages, including nimbleness in the deployment of expertise and geographic flexibility. But it also departs from conventional methods of accountability through both public oversight and private market discipline. The lack of transparency in the use of private contractors compounds the problem of assessing the impact of their increasing role. Failures of basic governmental oversight to ensure contract enforcement by the Department of Defense are well-documented. Departures from conventional government contracting procedures exacerbate these failures and obscure whether inherently governmental functions are in effect privatized. The large sums of money involved contribute to risks of corruption and a scale of private lobbying that can distort the legislative process. These developments jeopardize the effectiveness of military activities, the professionalism of the military, the integrity of the legislative process and foreign policy decision making, public confidence in the government, national self-interest, and the stability of the world order.

Introduction

In a time of scandals, three recent ones share a disturbing, though perhaps not initially obvious, element.

* Jeremiah Smith, Jr. Professor, Harvard Law School. This paper derives from the Cecil A. Wright Lecture, University of Toronto Faculty of Law, January 20, 2004. For insightful comments and questions, the Author thanks the editors of the Boston College Law Review; Dean Ron Daniels and members of the audience at the Cecil A. Wright Lecture at the University of Toronto; David Luban, Nancy Sherman, and students in their Joint Seminar on War at Georgetown University; and participants in the March 24, 2004, Harvard Law School Conference, Governance by Design. Special thanks to Brishon Rogers, Mira Edmonds, and Amanda K. Edwards for research assistance, and Jody Freeman, Bob Monks, and Joe Singer for conversations about this work.
Congressional and media sources charge the Halliburton Company, a Houston-based oil services firm previously headed by Vice President Richard Cheney, with overcharging $61 million worth of gasoline and for charging $186 million for meals not actually served as part of its $10 billion worth of contracts with the Department of Defense (the “DOD”) to support the U.S. military effort in Iraq. Halliburton holds the two largest contracts for reconstruction in Iraq with $2.5 billion to restore the oil infrastructure and $6.5 billion to provide the troops with housing, food, laundry, and other services. The contracts guaranteed the company a profit and allowed it to pass on all of its expenses to the government. One account indicates that the unaccounted-for charges amount to 43% of the amount the company billed, though that estimate understates the full

---


5 David Teather, *Halliburton Accused of Not Justifying £1bn Army Bills, Guardian* (London), Aug. 12, 2004, at 17; see also Dorgan Hearing, supra note 1, at 20 (testimony of Danielle Brian, Exec. Director of Project on Government Oversight). On transport subcontracting, a subcontract administrator discovered that the price was inflated by 500%, but no effort was made to recover these costs. Dorgan Hearing, supra note 1, at 25–26 (testimony of Maria deYoung, former Halliburton subsidiary Kellogg Brown & Root, Inc. (“KBR”) employee).
$5.6 billion of contracts awarded since the start of the Iraq war.⁶

• The Defense Advanced Research Projects Agency (“DARPA”), a research and development division within the DOD, launched in early 2002 an undertaking it initially called the Total Information Awareness project, but for political reasons later renamed as the Terrorist Information Awareness project (“TIA”).⁷ The project developed advanced information technology tools to use domestic and foreign databases in both governmental and commercial hands to search for “patterns that are related to predicted terrorist activities.”⁸ TIA used mathematical algorithms and other features of governmental software to “mine” data about any person, including “religious and political contributions; driving records; high school transcripts; book purchases; medical records; passport applications; car rentals; and phone, e-mail and internet search logs.”⁹ Overseen by John Poindexter, who was indicted for his involvement in efforts to provide secret and illicit support to a military force in Nicaragua in the Iran-Contra scandal,¹⁰ TIA formally

---

⁶ O’Harrow, supra note 4, at E1. Full information about the billing has not been available to the Congress or to the public. The Pentagon blocked Congressional efforts to secure full information about Halliburton’s billing under a $2.5 billion contract for oil site repairs and fuel imports. Erik Eckholm, Lawmakers, Including Republicans, Criticize Pentagon on Disputed Billing by Halliburton, N.Y. Times, June 22, 2005, at A10.


ended when Congress shut down its funding in the face of media critiques and opposition from political leaders on both the right and the left wings.\textsuperscript{11} Yet Congress still permits counterterrorism intelligence that targets foreign nations; 199 data mining projects receive or will receive federal support, and many TIA projects may proceed under classified programs outside of public visibility.\textsuperscript{12}

- Shocking digital photographs of cruel and abusive practices used by interrogators and guards at the U.S.-run Iraq prison at Abu Ghraib sparked massive investigations.\textsuperscript{13} Confirming the worst allegations of imperialism and inhumanity leveled by terrorist propaganda, these revelations brought shame and worldwide criticism to America.\textsuperscript{14}

Although not the major element in each story, one common thread among these scandals is the role played by private contractors working for the U.S. government in its response to the 9/11 terrorist attacks. The Halliburton scandal stands out as the most obvious. The largest private oil and military services company in the country, Halliburton received, without competitive bidding, the contract to manage the logistical planning for the Iraq war—and promptly reported a


\textsuperscript{11} \textit{Balance Information With Privacy}, ATLANTA J.-CONST., Dec. 10, 2003, at A22; William Safire, \textit{Privacy Invasion Curtailed}, N.Y. TIMES, Feb. 13, 2003, at A41 (noting range of opponents); \textit{Virtual Borders vs. Civil Liberties}, DENVER POST, May 31, 2004, at C7. DARPA had identified a range of technologies contributing to TIA, and there is no indication that termination of TIA involved terminating development or use of these other technologies. See DARPA, supra note 8, at app. B.


62% jump in revenues.\textsuperscript{15} Government officials defended the selection of Halliburton on grounds that it was the only firm large enough to manage the job, but Peter Singer notes that Halliburton often outsources further, hiring subcontractors with nationals from countries including Bangladesh and the Philippines.\textsuperscript{16}

With private contractors working for the military comes the practice of subcontracting, which raises the question of whether such an arrangement should allow the corporation to distance itself from responsibility any more than a contracting relationship should allow the military to separate itself from the acts of its contractor’s employees. In addition, the use of private contractors by the military foreseeable creates extra problems of supervision and control. Legal and political responses will determine whether those problems are excused or instead monitored sufficiently to create incentives for adequate oversight.

Investigative reports into the abuses at the Abu Ghraib prison identify participation of private contractors as one example of the poor supervision, confused lines of authority, and improper procedures at that site.\textsuperscript{17} One report, produced by a panel of U.S. generals, concludes that “[c]ontracting-related issues contributed to the problems at Abu Ghraib prison.”\textsuperscript{18} An earlier internal Army report traces the involvement of civilian interpreters and interrogators working for CACI International, Inc. (“CACI”) and Titan Corp. (“TTN”) in conjunction with military officers, and concludes that they were either directly or indirectly responsible for the abuses at Abu Ghraib.\textsuperscript{19} CACI denied involvement of its employees in the abuse, but also asserted

\begin{flushleft}
\end{flushleft}

\begin{flushleft}
\textsuperscript{16} Id.
\end{flushleft}

\begin{flushleft}
\end{flushleft}

\begin{flushleft}
\textsuperscript{18} See id. at 47. One of the companies denies any involvement, although an internal Army report had identified one of its employees as a key player. Ellen McCarthy, \textit{Changes Behind the Barbed Wire: New Standards Are in Place for the Oversight of Contract Workers at Abu Ghraib Prison}, \textsc{Wash. Post}, Dec. 13, 2004, at E1. Steven Schooner concludes that “CACI International provided more than half of the interrogators employed at the facility, while Titan supplied linguistics personnel,” and that “more than a third of the improper incidents involved contractor personnel.” Steven L. Schooner, \textit{Contractor Atrocities at Abu Ghraib: Compromised Accountability in a Streamlined, Outsourced Government}, 16 \textsc{Stan. L. & Pol’y Rev.} 549, 555 (2005).
\end{flushleft}

\begin{flushleft}
\textsuperscript{19} McCarthy, \textit{supra} note 18, at E1 (quoting a leaked internal Army report by Major General Antonio Taguba).
\end{flushleft}
that the individuals in question were no longer employees.\textsuperscript{20} Titan quickly responded that the individual associated with its operations actually worked for a subcontractor.\textsuperscript{21} Minimal or missing oversight seems to characterize recent military use of contractors.\textsuperscript{22} Dan Guttman at the Center for Public Integrity reports that only a thin layer of official oversight existed for the contractors running the TIA data mining project.\textsuperscript{23} Admiral Poindexter came to the DOD agency DARPA after working with Syntek Technologies to develop the research tools that became the basis of TIA.\textsuperscript{24} TIA worked by awarding contracts to private companies.\textsuperscript{25} Through a lawsuit seeking enforcement of the Freedom of Information Act (the “FOIA”),\textsuperscript{26} a nonprofit organization obtained information about the DOD contracts with private entities for TIA activities, including large corporations such as Booz Allen Hamilton,\textsuperscript{27} Lockheed Martin, and several universities.\textsuperscript{28} The government has long contracted out research and development work as well as information technology.\textsuperscript{29} Yet outsourcing the tasks of monitoring Americans as part of government investigations seems to represent a new degree of privatization.

The roles played by private contractors in these incidents offer a glimpse of not only the variety and scale of outsourcing by the U.S. military in recent years, but also the departures from conventional methods of accountability accomplished by extensive use of private

\textsuperscript{20} Id.


\textsuperscript{22} See Dorgan Hearing, supra note 1, at 1–2.

\textsuperscript{23} See generally Guttman, supra note 21.


\textsuperscript{25} See generally Mayle & Knott, supra note 24.


\textsuperscript{28} Id.

\textsuperscript{29} Nancy Ferris, \textit{Give and Take}, GovExec.com, July 1, 2003, http://www.govexec.com/features/0603/ots03s4.htm (“Even before the current push to outsource, contractors were doing at least three-quarters of the federal government’s IT work . . . .”).
contractors. Typically proceeding without much publicity or disclosure, private contractors working for the military theoretically could be subject to two systems of accountability: public oversight and private market discipline. Yet in practice, military contractors often evade the oversight intended to determine contract performance and also often bypass private market competition through sole-source bids and other waivers of marketplace practices. Private contractors may also enjoy exclusions from other legal constraints that would attach to government actors engaged in the very same activities.

The prospect of unaccountable private military contractors is disturbing, but also inconsistent with growing demands for compliance with human rights globally. Federal courts in the United States have opened their doors to claims by foreigners of human rights violations by U.S. citizens and U.S.-based corporations acting outside U.S. boundaries.

---


against U.S. corporations consulting with foreign governments—even though judicial remedies would be unavailable for abuses arising from the same contractors’ work for the U.S. military, given the court-made government contractor defense.33

The current scale of the military’s use of private contractors makes the question of accountability particularly pressing.34 Private military companies—not merely individuals offering their services as mercenaries35—have a long lineage,36 but never have been more central to the U.S. military strategy than in the deployment in Iraq.37 Peter Singer, one of the key observers of this trend, notes that private military firms in Iraq employ between 20,000 and 30,000 people—and that this, taken together, represents the second-largest force in Iraq after the U.S. military.38 Private military companies guard U.S. gener-

In Sosa v. Alvarez-Machain, the Supreme Court declined to immunize corporations from application of the Alien Tort Claims Act. 124 S.Ct. at 2766; see also Lisa Girion, Court Oks Foreign-Abuse Suits, L.A. TIMES, June 30, 2004, at C1. This development produced a settlement in the suit against Unocal, in which noncitizens alleged that the corporation was complicit with the Myanmar army that allegedly engaged in forced labor, murder, and rape while working jointly with Unocal on a pipeline. See Doe I v. Unocal Corp., 395 F.3d 992, 999–100 (9th Cir. 2002), rev’d en banc 403 F.3d 708 (9th Cir. 2005); Duncan Campbell, Energy Giant Agrees Settlement with Burmese Villagers, GUARDIAN (London), Dec. 15, 2004, at 17.


37 See Yanagisawa, supra note 34, at 2.

38 Id.
als, essential military sites, and U.S. government compounds in Iraq.\(^{39}\)

The private military companies offer their services to other nations, of course, generating about $100 billion in annual revenues around the globe.\(^{40}\) For instance, one U.S. company trained the Croatian leadership for eight months, which then defeated Serb forces in 1995.\(^{41}\)

Familiar in the medieval era, private companies creating and deploying troops diminished with the rise of nation-states, which were capable of raising and supporting their own military and which developed the conception that sovereign nation-states alone may legitimately use military force.\(^{42}\) The rise of the nation-state marked a shift from multinational armies, composed of soldiers recruited from many places, to the citizen-army.\(^{43}\) Emerging nation-states asserted their monopoly over not only the legitimate use of force, but also their own citizens’ use of force in and outside the nation’s territory.\(^{44}\) By the twentieth century, only those places with legacies of imperialism or temporary shortages of manpower deployed soldiers-for-hire, and the list of such exceptions (for example, Saudi Arabia, Libya, Angola, and Nigeria) matches a list of global trouble spots.\(^{45}\)

Yet the downsizing of major military efforts at the end of the Cold War and the end of apartheid in South Africa created a supply of individuals with military training in this country and elsewhere who could market their services.\(^{46}\) The widespread hope that privatization would improve efficiency and save costs for nation-states in turn generated enormous governmental demand for private contractors.\(^{47}\) Privatization in the abstract could mean ending government involvement alto-


\(^{44}\) *Id.* at 30–31, 34–39.

\(^{45}\) See *id.* at 27.


\(^{47}\) See Singer, *supra* note 36, at 66–70.
gether either by selling off public assets or terminating government funding and involvement in a particular activity. But with much federal governmental action—and most DOD initiatives—privatization simply becomes reliance on nongovernmental actors who are paid under publicly-funded contracts or vouchers. Increasingly, ordinary day-to-day government operations proceed this way. Private companies work under contracts to manage welfare programs, enforce child support obligations, and build and operate prisons. Intended to save money and bring efficiency, the private sector introduces competitive bidding and techniques of business management to government operations.

In my own work, I have explored the potential benefits to society from privatization of schools and human services—invoking for-profit, as well as nonprofit and religious organizations. These benefits include innovation, efficiency, and replenishment of pluralism that itself can support individual freedom. There are risks, too: risks of exclusionary practices, fraud, and religious coercion. Therefore, I have argued, privatization should be accompanied by an insistence on public values following private dollars. The content of those values, in turn, should stem from the Constitution and from public debate.

Does privatization make sense when the operations themselves involve the use of force—as they do with war, prisons, and police? In Canada, considerable privatization of social services has been followed by large-scale privatization of prisons, police, and the military. A decision to privatize the Canadian Forces’ Department of Supply and Warehousing generated critical television ads and the ironic

---

49 See Martha Minow, Partners, Not Rivals: Privatization and the Public Good 3 (2002).
50 See id. at 20; Minow, supra note 48, at 1242–46.
51 See Minow, supra note 48, at 1243.
52 See Minow, supra note 49, at 7–22.
53 See Minow, supra note 48, at 1242–46.
54 See id. at 1246–55.
55 See Minow, supra note 49, at 142–44.
56 See id. at 144–50.
58 See Alan Cairns, Jailhouse Blues: Ontario Towns Feel Sense of Betrayal as Gov’t Leans to Private-Run Jails, Toronto Sun, Apr. 2, 2000, at 32.
phrase, “welcome to War-Mart.”59 Private military companies recruit employees from the United States, Canada, and around the world.60 This investigation of privatization within the U.S. military requires further discussion about the nature and scope of its use of private military companies. It also invites consideration of three kinds of normative analysis that I pursue here. First, do the military contractors, once engaged, deliver what they promise? Second, does the use of private contractors overall advance or detract from purposes articulated by the military itself? Third, how well does the use of private contractors serve (or disserve) the purposes of a military in a constitutional democracy?

Before proceeding any further, I must acknowledge one great limitation of this effort: the lack of transparency and disclosure makes it difficult for the public—and for me—to know what is going on with the military’s use of private contractors. The private firms disclose some of their activities in promoting their services, but they can resist media and Congressional inquiries, claiming that they need to do so to protect proprietary information.61 Private companies are free from the disclosure obligations placed on the government by the FOIA, the federal law intended to make democracy work by ensuring access to all of the government’s information compatible with security.62 There is some authority that private companies enjoy the ability to enjoin the government from disclosing information they have shared with the government in the course of doing business together.63 As the

---

59 Ian McDougall, Bargain Warehouse: Union Irate as Army Privatizes Supply Department, Toronto Sun, Mar. 13, 2002, at 33.
60 Cf. Danilo Burzan, Arrests Reignite Kosovo Cauldron, Adelaide Advertiser (Australia), Aug. 5, 2000, at 56 (reporting that two Canadians, working for a private contractor, were arrested for participation in training Yugoslav Republic forces).
61 Silverstein, supra note 35, at 145.
United States fears for its security and engages in war, the executive branch is especially secretive, while Congress and the courts are more than usually deferential. The general public seems mainly unaware of and uninterested in examining the expanding role of private corporations in managing and operating force—weapons, prisons, and policing—in the name of the government. It remains surprising to people to learn that private contractors are engaging in interrogations in Guantanamo and Iraq, that publicly traded companies run prisons in this country, or that Halliburton built the military prison at Guantanamo Bay, Cuba. The lack of transparency about the scope and effects of private company contracts with the U.S. military is not merely a caveat admitting the limitations of my analysis, but an important demerit as I try to assess how well the contractors are performing, how well they are achieving goals of military purposes, and how well they are achieving goals of a constitutional democracy.

Apparently, at least on some level, there is growing public awareness of the presence of private military companies. A recent cartoon in the New Yorker magazine shows a group of kids dressed up with combat gear and toy guns. One says to another, “Tommy and Ben are like Green Berets, Dan and Jerry are Navy SEALS, and me and Scott are like private contractors.” Another cartoon earlier last year shows one worker referring to two soldiers standing behind his desk as his “private army.” Tamed to be part of children’s games or office cubicle humor in an elite magazine, the outsourcing of force disturbs at least our humorists. Let’s see if we, too, should be disturbed. I will pursue this inquiry both in terms of contract compliance and potential jeopardy to larger missions for the military, the nation, and democracy.

I. Is There Basic Contractual Compliance?

Whatever debates may arise over the norms that ought to apply in evaluating the performance of private military contractors, the basic
elements of contract compliance should not give rise to dispute. Do contractors do what they are asked to do—and not do what they are not asked to do? Massive outsourcing has drawn civilian military contractors into military work without meaningful methods for ensuring contractual compliance.  

The combination of the scale of the outsourcing, poor communication and coordination of contractors with military personnel, and defective managerial systems within both the DOD and military services produce this unavoidable conclusion.  

Presidents Bill Clinton and George W. Bush both embraced outsourcing as a way to help downsize the military. Congress adopted bipartisan caps on the number of civil servants employed by the government—but did not limit the number of persons who could be employed through contracts with private companies. For nearly a decade this has allowed policymakers to hide from public view the true size of the government. Many of the functions the government could outsource involve not only commercial activities such as transportation, laundry, and food services, but also planning, policy development, managing weapon systems, and managing the military workforce. President Clinton used private contract employees to administer the nuclear nonproliferation agreement with Russia.  

That supply and demand resulted in the growth of private corporations that offered military training and assistance, including logistical services and planning, weapons management and servicing, as well as weapons development and management of other private corpora-

---

68 See Guttman, supra note 30, at 323–24.  
69 See Schooner, supra note 18, at 556–60.  
72 Congress required agencies to inventory what work should be understood as commercial and what instead is inherently governmental. Office of Mgmt. & Budget, OMB Cir. No. A-76 (Revised), Performance of Commercial Activity (2003), available at http://www.whitehouse.gov/omb/circulars/a076/a76_incl_tech_correction.pdf [hereinafter OMB Cir. No. A-76 (Revised)] (regarding the Federal Activities Inventory Reform Act (FAIR) of 1998); see infra note 169 (discussing departures from Circular A-76).  
74 Schooner, supra note 18, at 554. See generally Guttman, supra note 21.  
75 Guttman, supra note 30, at 335.
tions doing business with the military.\textsuperscript{76} The government increasingly turned to private individuals to perform core military tasks and to private companies to orchestrate plans and implementation.\textsuperscript{77} The private firms have strong ties to U.S. government: \textsuperscript{78}

The Pentagon and CIA have long used private contractors for a variety of tasks, from building base infrastructure to assisting with covert operations. The current situation differs in both scope and size from past practice, most famously revealed in the Iran/contra scandal. Today, the firms most heavily involved are not CIA cut-outs but multimillion-dollar corporations with diverse interests. Their work is implemented not by foreign locals trained by the CIA but by high-ranking U.S. military officers fresh out of the armed forces.\textsuperscript{79}

On September 10, 2001, Secretary of Defense Donald Rumsfeld announced what he called an “all-out campaign to shift [the] Pentagon’s resources from bureaucracy to the battlefield, from tail to the tooth.”\textsuperscript{80} He declared that the Pentagon would be challenged to eliminate or shift to private suppliers any but the core activities of defense.\textsuperscript{81} He announced a commitment to ensure that the Pentagon would learn from, and take advantage of, the private sector’s expertise in management, technology, and business practices.\textsuperscript{82}

\textsuperscript{76} Guttman, \textit{supra} note 21.
\textsuperscript{77} See Lisa L. Turner & Lynn G. Norton, \textit{Civilians at the Tip of the Spear}, 51 A.F. L. REV. 1, 8 (2001). In 1999, noting that judge advocates increasingly see issues arising from the growing numbers of contractor personnel—and nongovernmental private organizations working alongside or near the military—Colonel Steven J. Zamparelli commented, “[n]ever has there been such a reliance on nonmilitary members to accomplish tasks directly affecting the tactical success of an engagement.” \textit{Id.} at 3 (quoting Steven J. Zamparelli, \textit{Competitive Sourcing and Privatization: Contractors on the Battlefield, What Have We Signed Up For?}, 23 A.F. J. LOG. 11, 11 (1999)).
\textsuperscript{78} Silverstein, \textit{supra} note 35, at xv (“[The private firms] are licensed by the State Department and are staffed by former military officers—effectively serving as an extension of foreign policy.”)
\textsuperscript{79} \textit{Id.} at 143.
\textsuperscript{81} \textit{Id.}
\textsuperscript{82} Secretary of Defense Donald H. Rumsfeld explained:

Already we have made some progress. We’ve eliminated some 31 of the 72 acquisition-related advisory boards. We now budget based on realistic estimates. We’re improving the acquisition process. We’re investing $400 million in public-private partnerships for military housing. Many utility services to military installations will be privatized.
delivered $300 billion worth of contracts to private military industries between 1992 and 2002. It has even outsourced security at Army bases within the United States.

When the work of national defense dramatically increased after 9/11, Pentagon reliance on private military companies escalated. Private contractors played key roles in the U.S. war in Afghanistan. They served in paramilitary units with the Central Intelligence Agency (the “CIA”) that hit the ground before other combat troops; they maintained combat equipment, provided logistical support, and worked with surveillance and targeting plans. Currently, private contracts are part of the military operation trying to locate Osama bin Laden. Major Gary Tallman, an Army spokesman, acknowledged the unprecedented level of outsourcing since 9/11 and commented, “[t]he Army is much smaller than in the past. When you run out of soldiers and they don’t have an expertise, one way to get that capability on the battlefield is to contract it.”

Before the war in Iraq started, the Army announced that it would permit contractors to compete for 154,910 civilian jobs—more than half of its civilian workforce—as well as 58,727 military positions. Economist and columnist Paul Krugman observes that the Bush administration has privatized everything in sight in Iraq, including guards for U.S. installations, interrogators, and other seemingly central military functions. Privatization at the DOD and the armed services has come to mean buying the time of people who work alongside Pentagon officials and troops on the ground. These people do, as Dan Guttman puts it, “what citizens consider the stuff of government: planning, policy writing, budgeting, intelligence gathering, nation

We’re tightening the requirements for other government agencies to reimburse us for detailees, and we’re reviewing to see whether we should suspend assignments where detailees are not fully reimbursed.

We have committed $100 million for financial modernization, and we’re establishing a Defense Business Board to tap outside expertise as we move to improve the department’s business practices.

Id.

84 Id.
85 Singer, supra note 15.
86 Id.
87 Id.
89 Guttman, supra note 21.
90 Schooner, supra note 18, at 553 (quoting Paul Krugman).
building,” but under the employment relationship of a temporary worker. After research and development and aircraft, the third largest category for military expenditures on contracted work recently has been for “professional, administrative and management support services.” In a letter to the Congressional Armed Services Committee on May 4, 2004, Secretary of Defense Donald Rumsfeld estimated there were 20,000 private security workers in the DOD’s employment in Iraq, which would make these private workers easily the largest group—larger than the British deployment—working alongside the U.S. military.

Immediate benefits are clear. A private company can handpick the team for a given project, and reassemble or disassemble the team when the job is done or changes. A private company can hire and send twenty former colonels, while the U.S. Army would have to “strip more than an entire combat division to muster that many,” observed Colonel Bruce Grant of the Institute for National Strategic Studies.

This kind of nimbleness is especially difficult for the DOD in its deployment of its civilian workers, even though they increasingly perform key roles in defense policy, intelligence, acquisitions, and weapon system maintenance. By depending on private companies, the military can obtain the newest technology and the staffs trained to maintain it—and even avoid the costs of retraining simply by shifting to a new team. But the heavy reliance on private contractors for both the enormous scale of work underway and for management, planning, and even supervision of other contracts raises questions at the

---

91 Guttman, supra note 21.
92 Id. These expenditures more than doubled over the past decade. Id.
93 Ante, supra note 21, at 76; see Singer, supra note 15 (estimating 15,000–20,000 contract workers in Iraq); see also Robert Collier, Global Security Firms Fill in As Private Armies, S.F. Chron., Mar. 28, 2004, at A1 (estimating 15,000 private security agents were employed by 25 private security firms in Iraq in 2004).
94 Silverstein, supra note 35, at 167.
95 Describing these functions performed by the civilian employees of DOD, a 2003 study by the GAO (then called the General Accounting Office, renamed the Government Accountability Office in 2004) observed that such deployment was impossible for the Department of Defense when it came to its civilian workforce because even it did not include “data on the skills and competencies needed to successfully accomplish future missions; therefore, DOD and the components risk not being able to put the right people, in the right place, and at the right time,” nor did it include how to integrate the civilian employees with their military counterparts or sourcing initiatives. U.S. Gen. Accounting Office, GAO-03-475, DOD Personnel: DOD Actions Needed to Strengthen Civilian Human Capital Strategic Planning and Integration with Military Personnel and Sourcing Decisions Exec. Sum. (2003), available at http://www.gao.gov/new.items/d03475.pdf [hereinafter GAO-03-475].
most basic level about whether anyone knows what the private contractors are doing. Are the contractors doing what they promised to do? Ensuring that the contracted work is performed, and performed without fraud, overcharging, or mismanagement, turns out to require clear and consistent oversight. Yet this seems precisely what has been missing certainly during the Iraq war period, but with clear signs of serious problems long before.

Management and oversight problems of two sorts appear to be growing to staggering proportions for the U.S. military: 1) failures of basic oversight and management of specific existing contracts both reflecting and resulting in confused lines of authority and accountability; and 2) widespread departures from established contracting processes due not only to exigencies of emergency but also to convenience and inattention, with disregard for established distinctions between inherently governmental and nongovernmental functions.

A. Failures of Basic Oversight to Ensure Contract Enforcement

Managing contracts may seem a nicety in the middle of a war, but when the contracts concern deployment of weaponry, security forces, and interrogators, that excuse seems less convincing. At the Abu Ghraib prison, military personnel did not receive guidance about how to use contracted personnel and did not know the terms of the contracts nor their procedures. Whatever else may be said about Abu Ghraib, this lack of clarity about contract workers and military is not anomalous. The Government Accountability Office (the “GAO”) found the same problem in the Balkans in 2000; military officers were confused about whether they could control the actions of contractors, and this lack of understanding was a major factor in the government’s inability to control adequately contract costs.

---

96 See Guttman, supra note 21.
98 See GAO-04-615, supra note 97, at 3; Schooner, supra note 18, at 572; infra notes 101–02.
99 See infra notes 101–49 and accompanying text.
100 See infra notes 150–78 and accompanying text.
101 Schooner, supra note 18, at 563–64 (discussing the Fay Report).
It is difficult to administer a contract when the contracting officer has no representative on site, yet this is the case with many DOD contracts.\textsuperscript{103} Even if the lines of authority clearly locate the civilian contractor employees under military command, these civilians do not face the same rewards and sanctions as do the members of the military, and other kinds of sanctions for misbehavior are limited and remote.\textsuperscript{104} The Fay Report, following the abuses in Abu Ghraib, concluded that a properly trained contracting officer’s representative must be on site to prevent a recurrence of that situation.\textsuperscript{105} Congress adopted the Military Extraterritorial Jurisdiction Act to allow prosecutions of contractors for offenses occurring while deployed if those offenses would be felonies in the United States, but this is a very limited tool for holding the line against misconduct far from home.\textsuperscript{106}

The contract enforcement failures at the DOD are pervasive and basic.\textsuperscript{107} Its financial management and related business operations, when reviewed by the GAO, show persistent fundamental failures that produce waste and inadequate accountability and threaten its mission.\textsuperscript{108} Specifically asking whether the DOD has adequate capacity to oversee contracts for logistic support across the military services, the GAO found that the contractors proved responsive when reviewed, but generally received inadequate reviews in part because insufficient numbers of appropriately trained military staff were on hand to provide effective oversight and also because many military units lacked understanding of their role in directing and monitoring the work of contractors.\textsuperscript{109} In another study in 2004, the GAO concluded that the fundamental flaws in the DOD’s business systems affect the effective-

\begin{itemize}
  \item \textsuperscript{103} See Schooner, supra note 18, at 557–58.
  \item \textsuperscript{104} See Singer, supra note 15.
  \item \textsuperscript{105} Fay Report, supra note 17, at 47; Dorgan Hearing, supra note 1, at 17 (testimony of Steve Schooner, Co-Director, Government Procurement Law Program, George Washington University Law School).
  \item \textsuperscript{106} Military Extraterritorial Jurisdiction Act, 18 U.S.C. §§ 3261–67 (2000); Fay Report, supra note 17, at 50.
  \item \textsuperscript{107} See Guttman, supra note 30, at 323–24 (discussing high-level official admissions of agency-wide deficiencies and GAO reports).
\end{itemize}
ness of the agency and contribute to fraud, waste, and abuse.\textsuperscript{110} Yet the government rarely suspends contractors, even for misconduct, overcharging, and other violations.\textsuperscript{111}

In effect confirming Secretary Rumsfeld’s 2001 call to revolutionize the DOD bureaucracy, the GAO has documented several problems: year-long delays in processing security clearances for contractor employees (with failures by individual military services to respect the security clearances granted by another service),\textsuperscript{112} inadequate information and no coordinated plan to assess whether various efforts to promote small businesses in the procurement activities could work,\textsuperscript{113} and millions of dollars in payment adjustments for mistaken billing.\textsuperscript{114} The DOD does not follow the best commercial practices in managing service contracts by analyzing spending patterns to enable savings.\textsuperscript{115} The DOD also fails to use value engineering to control costs in weapons until it is too late to make substantial savings.\textsuperscript{116} Logistical support contracts used by several of the armed forces with good results in earlier periods have been marred during the recent Iraq efforts by inadequate oversight, poor definition of terms, and failed cost containment.\textsuperscript{117}

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{110} GAO-04-615, \textit{supra} note 97, at 3.
\item \textsuperscript{111} Dorgan Hearing, \textit{supra} note 1, at 20 (testimony of Danielle Brian).
\item \textsuperscript{117} GAO-04-854, \textit{supra} note 109, at Highlights. The report notes:

In its contingency operations since the early 1990s, the Department of Defense (DOD) has relied extensively on logistics support contractors to provide many of the supplies and services needed by deployed U.S. forces. As requested, GAO assessed DOD’s planning in its use of logistics support con-
\end{itemize}
\end{footnotesize}
The form of some government contracts itself renders cost management difficult. Share-in-savings contracts allow the contractor who develops a system for the government to share in the savings the system is supposed to generate. Yet it is difficult to measure such savings. The lack of caps in these contracts exposes the government to potentially limitless demands for payment by the contractors, and questionable accounting practices used with these contracts exacerbate these problems.

Some problems with contract oversight are not new with the war in Iraq and substantially pre-date the Bush Administration. The GAO reported confusion by Army officers over whether they could control the actions of contractors in the Balkans in 2000, given the performance-based nature of the contract. Years of reports document problems with Los Alamos, the weapons facility run under contract by the University of California. Repeatedly losing classified and sensitive material and technology, more than three-quarters of the security personnel there also failed tests of required skills and were unable to demonstrate abilities to arrest intruders or shoot with

tracts in contingency operations; determined whether DOD has had contract oversight processes that are adequate to ensure that quality services were provided in an economical and efficient manner; and assessed the extent to which DOD provided trained personnel qualified to oversee its contractors. GAO focused its efforts on four logistics support contracts chosen because of their size and chosen to represent more than one military service—the Army’s Logistics Civil Augmentation Program (LOGCAP) and Balkans Support Contract, the Navy’s Construction Capabilities Augmentation Program, and the Air Force’s Contract Augmentation Program.

DOD did not have sufficient numbers of trained personnel in place to provide effective oversight of its logistics support contractors. The Army has deployed units responsible for supporting the LOGCAP contract, but some of the personnel have little knowledge of the contract. The Air Force did not consistently train evaluators to monitor its logistics support contractor’s performance. Military units across the services receiving contractor support have lacked a comprehensive understanding of their roles and responsibilities, which include establishing the work to be done by contractors and monitoring contractors’ performance.

Id.

Id.

Id.

See infra notes 122–25 and accompanying text.

GAO/NSIAD-00-225, supra note 102, at 21.

accuracy.\footnote{124}\footnote{See id.} But the deeper problem is the failure of the U.S. Department of Energy—after years of repeated complaints, congressional investigations, and reports—to demand accountability of the contracting parties, whether nonprofit institutions like universities or for-profit companies.\footnote{125}\footnote{See id.} To demand accountability is either to secure compliance or else find the contractors in breach, unworthy of payment, or renewal. To hold contractors accountable requires knowledge of what they have done, and if they have failed to perform adequately, then to replace them with another contractor or else to take the tasks back in-house.

Problems with contract monitoring may have been most acute during the Iraq war, and they did not much improve despite infusions of resources and the passage of time.\footnote{126}\footnote{Dorgan Hearing, supra note 1, at 18 (testimony of Steve Schooner). Congress called upon the DOD to devise a plan to supervise the private security companies and monitor human rights violations, overcharging, and quality. DOD has missed the deadline. See Bergner, supra note 39, at 54.} Steven Schooner attributes this to a failure to recruit aggressively the staff necessary to manage the contracting responsibilities.\footnote{127}\footnote{See generally Dorgan Hearing, supra note 1.} The DOD repeatedly has failed to hold private contractors accountable for their performance of multi-million dollar contracts in Iraq.\footnote{128}\footnote{A contracting officer who told Congress that Halliburton was abusing the contracting protocol faced a demotion; some viewed this as retaliation. See T. Christian Miller, Democrats Demand Probe of Demotion, L.A. TIMES, Aug. 30, 2005, at A8 (describing treatment of Bunnatine Greenhouse); Mark A. Stein, Indictments and Statistics All Overwhelmed by Tragedy Down South, N.Y. TIMES, Sept. 3, 2005, at C3 (same).} The Office of the Inspector General issued a 2004 report on the Coalition contracts in Iraq and found not only missing and incomplete records, but an inef-
fective system for contract review, tracking, and monitoring.\textsuperscript{130} Again, the lack of transparent information about all of this makes assessment necessarily incomplete, but the failure to obtain sufficient information apparently runs all the way to the DOD, which lacks basic information about how many private contract employees are on the ground in Iraq, the specific tasks each are to perform, and if they are in fact performing those tasks.\textsuperscript{131} This means that the DOD simply is defaulting on its contractual role as the paying, bargaining partner.

The DOD also relies heavily on contracts with private companies to monitor other private contracts.\textsuperscript{132} For example, Aegis was hired as the coordinating hub for more than 50 other private security companies in Iraq; it provides its own force of heavily armed protection teams to oversee reconstruction efforts, but also oversees the work of the other companies under a contract worth $293 million (£163m).\textsuperscript{133} The award to Aegis became a lightning rod. Former British army officer, Lieutenant Colonel Tim Spicer, heads Aegis.\textsuperscript{134} He commanded and defended two soldiers charged and convicted with a murder in Northern Belfast.\textsuperscript{135} He also previously directed a company that sold arms to Sierra Leone in violation of a United Nations embargo.\textsuperscript{136} The particular worries about him or his company should not obscure the larger problem: how well can one company monitor another in advancing governmental purposes? Such an arrangement is doomed if there are no clear guidelines and no sustained monitoring by the government of the oversight process itself.\textsuperscript{137}

In the Iraq war, the DOD has relied heavily on contractors who in turn subcontracted.\textsuperscript{138} At a Congressional hearing about Hallibur-

\begin{flushright}

\textsuperscript{131} See generally Singer, supra note 15.


\textsuperscript{133} Id.

\textsuperscript{134} See id.

\textsuperscript{135} Family Lobbies over U.S. Contract, Irish Times, June 18, 2004, at 8.

\textsuperscript{136} Id. The GAO rejected a challenge to the award of the contract brought by rival DynCorp, which pointed out Spicer’s involvement in violating an arms embargo. Jimmy Burns & Thomas Catan, DynCorp Seeks to Overturn Iraq Security Contract: Dispute Over Award to Company Headed by Controversial Former British Army Officer, Fin. Times (London), July 22, 2004, at 8.

\textsuperscript{137} The downsizing of the Department of Defense’s workforce dedicated to acquisitions coincided with the expanded use of outsourcing. See David A. Whiteford, Negotiated Procurements: Squandering the Benefit of the Bargain, 32 Pub. Cont. L.J. 509, 555 (2003).

\textsuperscript{138} Dorgan Hearing, supra note 1, at 12.
\end{flushright}
ton cost-overruns and inefficiencies, one of the witnesses, Marie deYoung, was an employee of Halliburton and a former army captain.\(^{139}\) Marie deYoung testified that Halliburton subcontracted to companies that in turn subcontracted, producing two or three layers of subcontracts.\(^{140}\) She concluded, “[w]e, essentially, lost control of the project and paid between four to nine times what we needed to fund that project.”\(^{141}\) An element of the scandal around Halliburton was its own failure to act promptly in paying its subcontractors who in turn faced bankruptcy and even threatened to stop performance—putting the security and effectiveness of the troops in jeopardy.\(^{142}\) The heavy reliance on subcontractors and on contractors to monitor other contractors may result from reductions in the government’s own acquisition workforce—a legacy of the downsizing movement.\(^{143}\) Steven Schooner concludes, “[t]he government has no choice at this point but to enter into larger, more complicated contracts, because they don’t have enough people to manage the contracts. So we’re being penny wise and pound foolish by not staffing up our acquisition workforce.”\(^{144}\)

The stream of studies by the GAO—as requested by members of Congress—offers some spot checks.\(^{145}\) Those checks, however, raise alarms about Pentagon failures to monitor contract performance. Reviews from the outside, by the GAO, Inspectors General of other departments, or the media, show failures to ensure cost-effectiveness and achievement of DOD purposes—and also failures to gather and review information sufficiently to permit ongoing contract oversight.\(^{146}\) The promise of cheaper and more efficient services at the heart of privatization in any field bears little relationship to reality in the military context where competition is difficult or impossible to sustain and where the private industry rather than the contracting

---

139 Id. at 10.
140 Id. at 12.
141 Id. (testimony of Marie deYoung).
142 Id. at 13 (“All Halliburton had to do was to negotiate reasonable prices and then pay its subcontracts for services rendered. But many of these vendors were not paid for months for a year at a time.”).
143 See Dorgan Hearing, supra note 1, at 33.
144 Id. (testimony of Steve Schooner).
146 See GAO-03-661, supra note 115, at 3, 6; GAO-03-574T, supra note 145, at 1. On failures to even secure the data that would allow control and management see Gutman, supra note 21.
government actors calls the shots.\textsuperscript{147} When the government is the sole purchaser, and a handful of contractors dominate the field, it is difficult to bar or suspend a major contractor and often difficult even to get the contractors to perform and document their costs.\textsuperscript{148} Ironically, as the Iraq war unfolded, the GAO changed its name from the General Accounting Office to the Government Accountability Office, just as the proliferating privatization of the military made government accountability newly elusive.\textsuperscript{149} It almost has an Orwellian quality; if we declare there is an office of accountability, there is accountability—yet the GAO reports repeatedly expose the failures of accountability within the DOD.

B. Failures to Comply with Contracting Processes

Several critics of private contracting by the DOD focus not on the DOD’s failures to manage and enforce particular contracts, but on its departure from the norms and standards of established contracting processes.\textsuperscript{150} Given the urgency of the situation in Iraq, Congress permitted waivers and irregularities in established procurement procedures,\textsuperscript{151} just as Congress authorized the Department of Homeland Security to pilot the use of acquisition agreements that departed from the usual government contract procedures.\textsuperscript{152} With Halliburton, the irregularities were justified because of the exigent circumstances—and also by claims that it was the only company with the relevant experience.\textsuperscript{153} Yet subsequent evidence indicated that another company had equal, if not superior, experience.\textsuperscript{154} Overall, the Policy Adminis-

\begin{thebibliography}{99}
\bibitem{148}See Guttman, \textit{supra} note 30, at 344.
\bibitem{150}See Schooner, \textit{supra} note 18, at 564–65.
\bibitem{153}Singer, \textit{supra} note 15.
\bibitem{154}Dorgan Hearing, \textit{supra} note 1, at 6 (comment by Rep. Waxman describing Bechtel); \textit{id.} at 8 (testimony of Sheryl Tappan, former Bechtel Proposal Manager) (“In my 12 years doing government proposals, I had never seen anything as arrogant, as egregious as the ways in which Pentagon officials, in particular Corps of Engineer contracting staff at
trator of the Office of Federal Procurement observed as she left office in September 2003, “[t]here is still not a lot of oversight in some areas of our contracting system, and I think it will haunt us.”

The contracts involved in the Abu Ghraib abuses proceed through an end-run around standard contracting procedures—an end-run that turns out to be well-trod. Agencies can streamline the purchasing process by paying a fee to a program manager in another agency, which in turn can select a favored contractor without a competitive bid. This shortcut permits sole-source awards and is supposed to separate programmatic authority from procurement authority, but also permits an agency to bypass procedural restrictions within its own department. Steven Schooner worries that, as a result, procurement is often accomplished by an agency working for the fee rather than results. The Inspector General for the Department of Interior concluded that chasing fees distorted the judgment of procurement officials.

These interagency task-order service contracts were routed through the Department of Interior and the General Service Administration (the “GSA”) in contracts for interrogation, intelligence, and security services in Iraq without establishing clear content or monitoring for those contracts. This approach may exempt the contractors from exposure to criminal liability that otherwise would attach through the Military Extraterritorial Jurisdiction Act. Alerted to these departures from checks and controls in the contracting process, the Defense Inspector General reviewed twenty-four contracts awarded between February and August 2003 by DOD and found that eighteen were awarded through the GSA supply schedule approach, bypassing procurement rules, and thirteen lacked adequate surveillance.
Experiments in streamlined contracting processes also have allowed the military to pay contractors for a flexible delivery order rather than specifying detailed agreement by contract.\textsuperscript{164} The Department of Interior’s Inspector General investigating the Abu Ghraib situation reported that CACI, a private contractor, was given six orders predominantly for interrogation, intelligence, and security services in Iraq.\textsuperscript{165} Reviewing this situation, Dan Guttman at the Center for Public Integrity concluded that neither his own research nor review by the GSA “could find any existing schedule that provided for these services.”\textsuperscript{166} Apparently, no one had checked to determine whether the contract work at Abu Ghraib was permitted or authorized, no one monitored the steps leading to hiring those individuals and directing their behavior, and no one determined whether those tasks were supposed to be contracted out at all.\textsuperscript{167}

Yet involving private individuals in interrogations and intelligence activities crosses the line into centrally governmental work.\textsuperscript{168} A longstanding executive policy, now expressed in Office of Management and Budget (the “OMB”) Circular No. A-76, directs that inherently governmental functions should not be outsourced.\textsuperscript{169} In 1998, Congress adopted a statute requiring agencies to inventory civil service work and to identify jobs as commercial or inherently governmental in order to assist the privatization effort.\textsuperscript{170} A GAO report in unrelated services. See Jon D. Michaels, Beyond Accountability: The Constitutional, Democratic, and Strategic Problems with Privatizing War, 82 Wash. U. L.Q. 1001, 1008 n.226 (2004); Mary H. Cooper, Private Affairs: New Reliance on America’s Other Army, 62 Cong. Q. Wkly Rep. 2194 (2004).

\textsuperscript{164} Schooner, supra note 18, at 564.
\textsuperscript{165} Guttman, supra note 21.
\textsuperscript{166} Id.
\textsuperscript{167} Id.; see also Guttman, supra note 30, at 340.
\textsuperscript{168} Guttman, supra note 21.
\textsuperscript{169} See generally OMB Cir. No. A-76 (Revised), supra note 72. This document includes a categorization of governmental activities into commercial and inherently governmental. Inherently governmental jobs would be those “so intimately related to the public interest as to mandate performance by government personnel.” Id. at A-2. Additionally, functions may be “inherently governmental” where they involve discretion and sovereign authority that could bind the U.S. and “[s]ignificantly affect[] the life, liberty, or property of private persons.” Id.
\textsuperscript{170} Federal Activities Inventory Reform (“FAIR”) Act of 1998, 31 U.S.C. § 501 (2000) (stating that in using the private sector for needed commercial services, officials are to identify savings and also identify noninherently governmental functions to enable cost comparisons between private bids and public budgets); see Maj. Mary E. Harney, The Quiet Revolution: Downsizing, Outsourcing, and Best Value, 158 Mil. L. Rev. 48, 61–92 (1998) (describing FAIR, Circular A-76, and the competitive cost comparison process); see also Diebold v. United States, 947 F.2d 787, 789–90 (6th Cir. 1991) (finding the Army’s decision to
2002 indicated that the government should retain for its own workers certain tasks of “warfighting, judicial, enforcement, regulatory, and policymaking functions.”

Nevertheless, disagreements over precisely what an inherently governmental activity is and gaps in governmental capacity contribute to ambiguity over what can or should be outsourced. For example, a private contractor may be entrusted with providing security, as well as driving trucks to transport soldiers, equipment, and food, but when does this function move from civilian support to core military activity? The Department of Defense has not adopted nor consistently applied measures to identify and track what functions should remain within the government and what can be outsourced. The DOD increasingly spends billions of dollars in private contracts without clarifying or monitoring this policy.

In its recent consideration of these issues, the government of Great Britain concluded that “[t]he distinction between combat and non-combat operations is often artificial. The people who fly soldiers and equipment to the battlefield are as much a part of the military operation as those who do the shooting.” A contractor may be

contract out food services was required by federal procurement law because the private company could provide the services at a lower cost than the DOD cost).


172 See Schooner, supra note 18, at 556 n.22 (describing fights within the government over outsourcing what may be inherently governmental functions). The official Pentagon statement is that private security companies “are not being used to perform inherently military functions.” Bergner, supra note 39, at 32 (quoting an “officially approved written statement”).

173 See Schooner, supra note 18, at 553–54. “Just as the distinction between combat arms and non-combat arms has become blurred during operations, the distinction between ‘advising’ and ‘doing’ for these contractors is similarly blurred,” states Major Thomas Milton of the Foreign Area Officer Association. Silverstein, supra note 35, at 166. Ken Silverstein writes: “[t]he reality is that most of these corporations’ operations become an integral part of the foreign government’s military capability.” Id. Although Silverstein was writing about private military companies working for foreign governments, the points would seem to hold for the relationship between such companies and our government and military, too. Eroding or elusive, the distinctions between noncombat and combat and between advising and doing are difficult to track and monitor. See Singer, supra note 15.

174 Schooner, supra note 18, at 555–56.

175 See id. at 555–57; Singer, supra note 15.

hired to support a complex technological weapons system, which may be an unmanned surveillance aircraft and support for it may include its deployment. If so, has the surveillance and deployment of that aircraft become an inherently governmental function? Is interrogation an inherently governmental task? What if the government lacks sufficient knowledge of the subject’s language to engage in effective interrogation while a contract employee may have that knowledge?

We do not even have to reach—at least not yet—the difficult philosophic question about what should be viewed as an inherently governmental activity, sheltered from contracting out, in order to conclude that the current governmental practices fail to provide sufficient monitoring even to attend to that question. Whatever the line is or should be, there must be sufficient government resources and attentiveness to monitor it and to review outsourced functions. That, at minimum, is an inherently governmental task.

II. FURTHER JEOPARDY TO THE MILITARY, THE NATION, AND DEMOCRACY

The reliance on contractors—even if subject to adequate oversight and appropriate procurement practices—risks jeopardy to the quality of military activities, the national interest, and democratic values.

A. JEOPARDY TO THE MILITARY

Investigation of the abuses in Abu Ghraib revealed that private contractor employees “wandered about with too much unsupervised free access in the detainee area.” The situation in Abu Ghraib, one hopes, was anomalous in many respects. Abu Ghraib did, however, also expose problems that arise with the command structure, discipline, accountability, security, and predictability when personnel working together include members of the military and private contract employees. This is precisely the situation established when the military contracts out logistical planning and support as it has in

177 Dan Guttman reports that although the Army concluded that intelligence work is inherently governmental, it did not bar contractors from that work in the Contractors on the Battlefield field manual used in Abu Ghraib. Guttman, supra note 21.


180 See Dorgan Hearing, supra note 1, at 18.
Iraq. Marie deYoung, former U.S. Army captain and former employee of Halliburton, observed the loss of control produced by layers of contracting and subcontracting, and the confusion over lines of authority and measures of accountability.\(^\text{182}\)

Congress adopted the Military Extraterritorial Jurisdiction Act to allow the prosecution of contractors in the United States for offenses occurring while they are deployed abroad, but only if the conduct would be a felony if committed in the United States.\(^\text{183}\) Congress adopted this law after U.S. military and local Bosnian law enforcement found they had no ability to prosecute employees of DynCorp apparently engaged in human sex trafficking.\(^\text{184}\) Through the prostitution ring, employees of DynCorp “purchased” young women and children to serve as their sexual slaves.\(^\text{185}\) Once the practice was exposed, the company fired the individuals involved—but there were no prosecutions for statutory rape, human trafficking, or anything under military, Bosnian, or U.S. law.\(^\text{186}\) The U.S. military uses this same company to train the Iraqi police.\(^\text{187}\) Members of the military can face swift court marshals, but civilian contractors fall outside that jurisdiction and elude any domestic legal system as well.\(^\text{188}\) A whistleblower lost her job for exposing the scandal—and later won a damages award on that basis in Britain.\(^\text{189}\) If left entirely to the military, it is not obvious that such scandals would get swift treatment. At least one civilian officer in Bosnia was told by his military commanders to lie about the

\(^{181}\) See id. at 10–13.

\(^{182}\) Id. at 12 (testimony of Marie deYoung).


\(^{185}\) See Capps, Crime Without Punishment, supra note 184; Capps, Outside the Law, supra, note 184. As Laura Dickinson has noted, this statute would not reach contractors working with the CIA or Department of Interior unless they are running U.S. facilities overseas as specified in a recent statutory provision. Laura A. Dickinson, Government for Hire: Privatizing Foreign Affairs and the Problem of Accountability Under International Law, 47 WM. & MARY L. REV. (forthcoming 2005) (manuscript at 52, on file with author) (discussing the USA PATRIOT Act of 2001, 18 U.S.C.A. § 7 (West 2005)).

\(^{186}\) See Capps, Crime Without Punishment, supra note 184.

\(^{187}\) See id.; Capps, Outside the Law, supra, note 184.


\(^{189}\) John Crewdson, Contractor Tries to Avert Repeat of Bosnia Woes, Sex Scandal Still Haunts DynCorp, Chi. Trib., Apr. 19, 2003, at C3.
DynCorp sex scandal, and if he had been a member of the armed services, he would not even have been able, as he was, to quit.\(^{190}\)

Yet even with the Military Extraterritorial Jurisdiction Act, procedures to move against civilian contractors remain unclear.\(^{191}\) Its scope may not apply to civilians who do not work directly for the government.\(^{192}\) It offers only limited control on the scene for the military commanders or even civilian supervisors.\(^{193}\) Military lawyers have been writing for a decade about the ambiguity over what law applies to contractors working in military settings.\(^{194}\) The contract employees are not governed by military discipline or norms; nor are they regulated by rules that apply only to government actors, such as the FOIA, limits on political activities, and conflict of interest rules.\(^{195}\) Ambiguities remain over what law applies if contract employees are captured or injured in confrontations with enemies.\(^{196}\) Both in terms of legal authorization and actual competence and training, can contract employees defend themselves with force—or does that exceed their role and jeopardize the safety of the military members working alongside


\(^{191}\) See Katie Fairbank, *Who Investigates Private Interrogators in Iraq? Use of Contractors to Gather Intelligence Raises Concerns, DALLAS MORNING NEWS*, May 7, 2004, at 22A.

\(^{192}\) Miller, supra note 188, at A8.


\(^{194}\) See, e.g., id.; Major Mark R. Ruppert, *Criminal Jurisdiction over Environmental Offenses Committed Overseas: How to Maximize and When to Say “No,”* 40 A.F. L. REV. 1, 17–19 (1996); Turner & Norton, supra note 77, at 25 (discussing lack of clarity over whether contractors, unlike military, are subject to the employment, tax, and customs laws of the host nation); see also Charles E. Cantu & Randy W. Young, *The Government Contractor Defense: Breaking the Boyle Barrier*, 62 ALB. L. REV. 403, 406 (1998) (exploring liability issues); Washburn & Bigelow, supra note 88, at A1 (stating that it remains unclear whether military, U.S., or Iraqi law applies to crimes committed by civilian contractors, and whether these civilian contractors are subject to FOIA and direct congressional oversight).


them? Do civilian contractors compromise the security of a mission when they discuss troop movements in a restaurant in Baghdad? The Third Geneva Convention would seem to cover combatants who are civilians if they are under the command of a superior, wear distinctive fixed signs recognizable at a distance, carry arms openly, and conduct themselves in accord with the laws of armed conflict, although each of these elements may be ambiguous in the case of particular contract employees. But Protocol I to the Geneva Convention deprives mercenaries of the privilege of serving as lawful combatants or immunity as prisoners of war upon capture. The line between “contract employee” and “mercenary” for this or other purposes in international law remains unclear. Confusion about precisely these legal questions can lead to disorder and ineffectiveness in actual operations, harming military effectiveness.

The military can be harmed in a different way if the option of retiring to a new career with a better-paid private contractor appeals to talented officers. Margaret Stock, a professor at the United States Military Academy at West Point and a reserve officer, explains how retiring officers can benefit from entering the private sector:

Military officers may be very tempted to retire in order to work for private military companies because the day after retiring, an individual can collect his or her retired pay, which is typically 50% of an active duty salary, and at the same time return to work as a consultant in essentially the same capacity, but with a new salary. Such individuals perform the same function while receiving both their retired pay and the consultant salary.

197 See Dorgan Hearing, supra note 1, at 13 (testimony of Marie deYoung).
198 Id. The for-profit entities working under contract may avoid disclosure by using the trade secret exemption. See Freedom of Information Act, 5 U.S.C. § 552(b)(4) (2000); Dickinson, supra note 185 (manuscript at 86).
199 Vernon, supra note 102, at 405–08.
201 Even the definition of “mercenary” in international law is unclear. See infra note 211 and accompanying text.
202 E-mail from Margaret Stock, Associate Professor of Law, United States Military Academy, to Martha Minow, Jeremiah Smith, Jr. Professor of Law, Harvard Law School (Oct. 20, 2005, 16:54:21 EST) (on file with author). Westerners earn between $400 and $700 each day they work in Iraq for private security companies, whereas employees from Fiji, Chile, and other non-Western nations earn between $40 and $150 per week. Bergner, supra note 39, at 34.
Reliance on private sources for important tasks, such as logistics and maintenance of advanced technological weapons, may in addition relieve the military of developing those capacities internally—to the long-term detriment of military strength.\(^\text{203}\) Similarly, depending upon contracts for the leases of trucks and equipment without ensuring appropriate maintenance plans can leave the military vulnerable at crucial movements to failures beyond their control to remedy.\(^\text{204}\) If the contractor in turn does not pay subcontractors—as was the case apparently with Halliburton—vendors may grow resentful or even collapse under bankruptcy.\(^\text{205}\)

Alternatively, the military also may compromise its strength by relying through layers of subcontracts on people it would never use directly, or by relying on individuals from third-country nations who are paid little and shift loyalties based on who pays them.\(^\text{206}\) Global military companies have recruited members of defeated armed groups and militias as mercenaries.\(^\text{207}\) Some of the companies do little to screen employees.\(^\text{208}\) One contract employee turned out to be a fugitive charged with embezzlement and previously convicted of assault in the United States.\(^\text{209}\) The work can attract volatile individuals. Mercenary involvement in the Congo over decades is a notable dimension of the area’s violence and instability.\(^\text{210}\)

Here the use of contractors raises the enduring question about mercenaries.\(^\text{211}\) Nicolo Machiavelli argued against mercenaries in his

---

\(^\text{203}\) Davidson, supra note 183, at 265.
\(^\text{204}\) Id.; see Vernon, supra note 102, at 393–94 (discussing risks of contracting for maintenance and repair of weapon system and other military equipment).
\(^\text{205}\) Dorgan Hearing, supra note 1, at 13 (testimony of Marie deYoung).
\(^\text{206}\) See id.
\(^\text{207}\) See Singer, supra note 15.
\(^\text{208}\) See, e.g., Bergner, supra note 39, at 50.
\(^\text{209}\) See id.
\(^\text{211}\) There is no settled definition of a mercenary in international law. See Milliard, supra note 35, at 19–69 (discussing alternative meanings under different international sources). The British government has considered regulation for private military companies and tries to distinguish mercenaries from servicemen in foreign armies and defense industrial companies. Ninth Report, supra note 176, at 1–3; Return to an Address of the Honourable the House of Commons, supra note 176, at 7–8, 22–24. The distinction might be drawn between an individual (or group) selling combat services and an individual (or group) selling support services to government militaries, but the increasing roles of technological and complex logistical operations make this distinction increasingly of little
classic work of politics, *The Prince*, because they work for pay. Illustrating Machiavelli’s warning that soldiers working for pay would not take the kind of life-risking action that can turn the tide of battle, some contractors during the Gulf War fled from a possible chemical weapons attack. Perhaps if the contractors build a team of retired military officers, the ethos of loyalty to the country and the military can be sustained even among these civilian employees. Yet Machiavelli’s warnings become more powerful for other employees, and especially for low-paid employees brought in from other countries under subcontracts.

### B. Jeopardy to the United States and Democracy

Reliance on private contractors also risks exposure to war profiteering: the exploitation of the chaos and fear of wartime by significance in terms of advancing military effort. See *Return to an Address of the Honourable the House of Commons*, supra note 176, at 7–8.


213 Vernon, supra note 102, at 394 (“During the Persian Gulf War, a small number of contractors fearful of chemical weapon attacks fled from an air base in Saudi Arabia. While the contractor’s [sic] departure did not disrupt the operation, it highlighted potential weaknesses. The contractor [sic] decided that financial gain was simply not worth the risk. This highlights the major difference between military personnel and contractors: one is present to serve his country, the other to make a profit.”).

214 See id. at 394–95 (discussing contractors’ arguments).

215 Machiavelli wrote:

Mercenaries and auxiliaries are useless and dangerous; and if one holds his state based on these arms, he will stand neither firm nor safe; for they are disunited, ambitious and without discipline, unfaithful, valiant before friends, cowardly before enemies; they have neither the fear of God nor fidelity to men, and destruction is deferred only so long as the attack is; for in peace one is robbed by them, and in war by the enemy. The fact is, they have no other attraction or reason for keeping the field than a trifle of stipend, which is not sufficient to make them willing to die for you.

*Machiavelli*, supra note 212, at 18; see *Protocol Additional to the Geneva Conventions of 12 Aug. 1949, and Relating to the Protection of Victims of International Armed Conflicts* (Protocol I), June 8, 1977, art. 47, 1125 U.N.T.S. 3, available at http://www.ohchr.org/english/law/protocol1.htm (defining mercenary as one recruited to and taking part in armed hostilities, motivated essentially by desire for material compensation in excess of what combatants of similar ranks earn in armed forces, and not a national of a party to the conflict or member of armed forces of a party to the conflict).
suppliers of materials to the military. Resisting war profiteering has been a governmental goal as long as this nation has existed.\textsuperscript{216} War profiteering is a serious problem not only because it diverts public monies—the money of the citizens—to private hands through overcharging and fraud, but also because it can jeopardize peacemaking and broader confidence in government. These issues overshadow but should not obscure the problems of former government officials finding employment with contractors after helping them build connections with the government.\textsuperscript{217}

Meanwhile, the revenues pouring into private military companies—the stock in publicly traded private military companies jumped 50\% after 9/11—are funneled into lobbying.\textsuperscript{218} Iraq contractors DynCorp, Bechtel, and Halliburton donated more than $2.2 million to political causes—mainly Republican—between 1999 and 2002, according to the Center for Responsive Politics.\textsuperscript{219} Lobbying efforts by private contractors have documented effects on policies regarding weapons systems development.\textsuperscript{220} It does not seem out of bounds to wonder about the influence contractor lobbying has on foreign policy. Political scientists have studied the relationship between defense spending and domestic political goals.\textsuperscript{221}

But even more troubling is the possibility that by using private contractors, the government can avoid checks and balances in a democratic system. This is the caution pressed in articles by Deborah


\textsuperscript{218} Publicly traded stocks in private military companies jumped roughly fifty percent, making this one of the few industries whose economic outlook improved after 9/11. Singer, supra note 15; see Nelson D. Schwartz, Pentagon’s Private Army, FORTUNE, Mar. 17, 2003, at 100, 102.

\textsuperscript{219} See Rupert Cornwell, et al., How the Allies Won the War; But Then Lost the Battle for Peace, INDEP. (London), Jan. 28, 2005, at 4; Jim Krane, Private Armies Also Fight, ST. LOUIS POST-DISPATCH, Oct. 30, 2003, at A8. Companies with strong ties to the Bush Administration have received contracts in the efforts to rebuild after Hurricane Katrina. Bush Associates Win Disaster Relief Contracts; Special Report: After Katrina, INDEP. (London), Sept. 11, 2005, at 18.

\textsuperscript{220} See Markusen, supra note 147, at 484–85.

Avant and Jon Michaels.\textsuperscript{222} Because private contractors are obliged to share far less information with Congress than required of the DOD or the military, Avant argues that the administration can effectively limit congressional checks on foreign policy.\textsuperscript{223} She claims that the United States can also advance its interests indirectly by licensing a private military company to assist another government, so that the United States itself can deny that it is actually pursuing foreign policy.\textsuperscript{224} As an example, she notes that “in 1994, the United States licensed U.S. company Military Professional Resources International (MPRI) to provide advice and training to the Croatian government. The country’s president, Franjo Tuđman, received the advantages of U.S. military assistance, but through a private entity.”\textsuperscript{225} Jon Michaels similarly warns that democratic accountability can be bypassed with private contractors doing military work.\textsuperscript{226}

The lack of clear lines of authority and sanctioning power over civilian contractors also potentially impairs the nation’s reputation internationally. The Iraqis do not distinguish between the civilian contractors and the U.S. military in judging the conduct of the U.S. occupation.\textsuperscript{227}

Foreign policy can be shaped even more insidiously by reliance on private contractors. As Jim Krane put it for the Associated Press, “the use of contractors also hides the true costs of war. Their dead aren’t added to official body counts.”\textsuperscript{228} With an estimated thirty to forty private contractor employees killed due to fighting in Iraq in 2004, and many more killed in accidents, including these private employees would notably increase the total casualties and injuries from the war.\textsuperscript{229}

\textsuperscript{222} Avant, supra note 42; see Michaels, supra note 163, at 1040–41.
\textsuperscript{223} See Avant, supra note 42.
\textsuperscript{224} Id.
\textsuperscript{225} Id. At least one reporter saw a connection between the private military company aid and later U.S. support. See Richard Whittle, U.S. Loan to Fund Bell Aircraft Deal, DALLAS MORNING NEWS, Oct. 31, 1996, at 12D.
\textsuperscript{226} Michaels, supra note 163, at 1011, 1050–52.
\textsuperscript{228} Krane, supra note 219, at A8.
\textsuperscript{229} Some estimates indicate that more than 200 private contract employees have been killed in Iraq. Risk and Reward in Iraq, N. TERRITORY NEWS (Austl.), May 7, 2005, at 15.
Use of contractors contributes to a lack of transparency in the conduct of military activities regarding not only casualties and injuries, but also total numbers of people deployed, and, indeed, the total size of the government-sponsored effort.\textsuperscript{230} This puts the scale of the initiative outside of public awareness and full political discussion, obscuring choices about military needs and human implications.\textsuperscript{231} Congressional interest in private contracting may emerge, but full oversight will be hampered by the insulation of the private companies from public review.\textsuperscript{232} Even information about procurement decisions and practices has been privatized, placing them further out of public reach.\textsuperscript{233} Previously, the Federal Procurement Data Center made available through the Internet information about the allocation of defense contracts to private firms, which could be searched by the name of the firm.\textsuperscript{234} In the past year, however, the government has outsourced this service and now charges for access to the data.\textsuperscript{235} The ability of the government to bypass public debate may also make it easier, on occasion, for the government to plan and launch either an aggressive war or a humanitarian intervention. Even if I agree with the ends, should I not worry if the means require bypassing democratic review?\textsuperscript{236}

I began with a caveat about the limitations of my entire discussion given the curbs on information available about outsourcing by the military.\textsuperscript{237} Even if it had the political will to try to exercise oversight, Congress would be largely constrained in reviewing the actions and practices of private military contractors.\textsuperscript{238} Media and ordinary citizens—even competitors—also face constraints, including curbs on otherwise available tools of disclosure.\textsuperscript{239} In the past, courts have con-
The Trade Secrets Act and the FOIA exemption for trade secrets and confidential or privileged commercial or financial information, courts have rejected release of prices paid by the government for servicing planes to be flown by military personnel on dangerous missions. At stake in these rather technical cases is nothing less than the ability of citizens to know what the government is doing, and yet outsourcing veils its conduct. The Court of Appeals for the District of Columbia Circuit asserted that the prices at issue concern the internal workings of a contractor, not the Government. As Judge Garland wrote in partial dissent, this nondisclosure interferes with the public’s ability to evaluate “whether the government is receiving value for taxpayer funds, or whether the contract is instead an instance of waste, fraud, or abuse of the public trust.” Yet the public interest is even more extensive when the military outsources logistics, services, and security to a contractor. It remains to be seen how transparent any of those actions will become.

**Conclusion**

Besides jeopardizing internal democratic monitoring and besides failing to control costs or even the performance of private employees, the expanded governmental use of private military companies erodes the control of force represented by the ascendancy of the nation-state. The nation-state itself had an interest in demonstrating its neutrality vis-à-vis other states as part of the Westphalian Pact: the leaders of each nation would respect the borders and sovereignty of another state while expecting other states to respect their own bor-

---


240 E.g., McDonnell Douglas Corp. v. U.S. Dep’t of Air Force, 375 F.3d 1182, 1193 (D.C. Cir. 2004); see id. at 1194 (Garland, J., dissenting in part) (characterizing the contract at issue); see also McDonnell Douglas Corp. v. Nat’l Aeronautics & Space Admin., 180 F.3d 303, 307 (D.C. Cir. 1999) (rejecting disclosure of line item prices in contract awarded to private company by NASA); McDonnell Douglas Corp. v. Widnall, 57 F.3d 1162, 1164 (D.C. Cir. 1995) (noting that items falling within trade secret exemption from FOIA may not be disclosed).

241 McDonnell Douglas Corp., 375 F.3d at 1193.

242 Id.

243 See Dorgan Hearing, supra note 1, at 12–13; Guttman, supra note 21.

244 See Thomson, supra note 42, at 36–39.
ders and sovereignty. In turn, the leaders would regulate the conduct of their own citizens to make sure they would not join foreign armies or otherwise jeopardize this system. And deployment of a non-national by a foreign sovereign would itself be viewed as infringement of the home nation’s sovereignty.

Thomas Jefferson argued that “the granting of military commissions, within the United States, by any other authority than their own, is an infringement on their sovereignty, and particularly so when granted to their own citizens, to lead them to commit acts contrary to the duties they owe their own country.”245 As the system developed, it supported national self-interest by ensuring that “no one can raise an army within the jurisdiction of the United States with the intention to commit hostile acts against a state friendly to the U.S. state.”246 By controlling private desires for money or adventure, the nation-state would build a more secure world.

Signs of the decline of the nation-state come in many contexts besides the rising use of private military companies.247 Nonetheless, the growing role of private military companies is a symptom of a larger, dangerous challenge to the aspirations of order in the world represented by the system of nation-states and the rule of law.248

245 Id. at 37 (citing Charles G. Fenwick, The Neutrality Laws of the United States 19 (1913)).
246 Id. at 38.
248 See McGinnis, supra note 247, at 905; Kaplan, supra note 247, at 46, 73–74. See generally Ohmae, supra note 247; Slaughter, supra note 247.
OH CANADA!:
ANTITRUST GEOGRAPHIC MARKET DEFINITION AND THE REIMPORTATION OF PRESCRIPTION DRUGS

Abstract: In recent years, public attention has focused on the need for affordable prescription drugs. Although Congress has recently enacted a Medicare Prescription Drug Plan, many private citizens and state and local governments continue to reimport prescription drugs from Canada to take advantage of the lower drug prices available in Canada. Many pharmaceutical companies have responded to this phenomenon by cutting off supplies of their drugs to Canadian pharmacies engaging in reimportation. As a result, some state and local governments have initiated litigation alleging state antitrust violations. This Note addresses one of the first questions raised in U.S. antitrust litigation under the Rule of Reason—the definition of the geographic market. First, this Note surveys the current caselaw standard and the academic approaches to geographic market definition. This Note then applies these approaches, concluding that Canada may be included in any geographic market when addressing the legality of reimportation in an antitrust context.

Introduction

Although global drug manufacturing companies with operations in the United States ship and sell their products all over the world, including Canada, certain drug manufacturing companies have recently threatened to reduce their drug supplies to Canadian distributors because the distributors are redirecting the products to U.S. consumers.¹ For example, in a letter dated January 14, 2005, Merck & Co. officials threatened to block supplies of their drugs to Canadian pharmacies that continued to directly or indirectly sell Merck products to U.S. residents.² Other pharmaceutical companies, such as Pfizer, AstraZeneca International, and Wyeth have also stated similar intentions.³

² Id.
³ Id.
The present trend of levying supply threats to Canadian pharmacies and wholesale suppliers comes in the wake of continued evidence that numerous U.S. customers are turning to Canada, rather than to domestic suppliers, to purchase their prescription drugs. This practice has come to be known as reimportation, as it involves individuals purchasing prescription drugs made in the United States that are sent to Canada and then brought back into the United States. The allure of reimportation lies in the differing pricing structures for prescription drugs in the United States and other countries. Because of the Canadian government’s price regulations, prices for the same prescription drugs are often thirty to fifty percent cheaper in Canada than in the United States. Drug manufacturers’ recent threats are an effort to stem the flow of drugs across the Canadian border and back into the United States. For pharmaceutical companies, the continued supply of the Canadian market, which is one-twentieth the size of the United States market and regulated at lower prices, could lead to the loss of billions of dollars in U.S. sales without any comparable gains.

Although levying threats to Canadian suppliers might seem to be a rational business practice for the drug companies, such behavior could create potential antitrust liability, if such actions are shown to be an unreasonable restraint on trade that lack any redeeming virtue—and, more importantly, if such actions can be reached by U.S. antitrust laws. In fact, the Minnesota Attorney General is currently

5 Id. at 3.
7 Id.
8 See Merck Tightens Sales to Canada, supra note 1, at C2.
10 See Sherman Anti-Trust Act § 1, 15 U.S.C.A. § 1 (West 2005). Antitrust liability would stem from the Sherman Anti-Trust Act, which prohibits agreements that unreasonably restrain trade:

Every contract, combination in the form of trust or otherwise, or conspiracy, in restraint of trade or commerce among the several States, or with foreign nations, is declared to be illegal. Every person who shall make any contract or engage in any combination or conspiracy hereby declared to be illegal shall be deemed guilty of a felony, and, on conviction thereof, shall be punished by fine not exceeding $10,000,000 if a corporation, or, if any other person,
seeking to obtain internal documents from GlaxoSmithKline in relation to the office’s investigation of such conduct as a potential violation of Minnesota’s state antitrust laws. One potential allegation is that GlaxoSmithKline engaged in vertical restraints—between firms at different levels in the production and distribution network, such as between a wholesaler and retailer—in violation of section 1 of the Sherman Anti-Trust Act. Alternatively, drug companies could also be exposed to antitrust liability if they colluded with each other, as a horizontal restraint, to stop the importation of drugs from Canada by collectively threatening to cut off supplies to Canadian distributors. Such actions could constitute a concerted refusal to deal, also known as a group boycott, in violation of section 1 of the Sherman Act. In seeking documents from GlaxoSmithKline, the Minnesota Attorney

$1,000,000, or by imprisonment not exceeding 10 years, or by both said punishments, in the discretion of the court.

*Id.* The underlying purpose of U.S. antitrust laws is to promote and protect competition. *Phillip E. Areeda & Herbert Hovenkamp, ANTITRUST LAW: AN ANALYSIS OF ANTITRUST PRINCIPLES AND THEIR APPLICATION ¶ 100 (2d ed. 2000).* Optimal competition occurs when market price equals manufacturers’ marginal costs. *Id.* Thus, optimal competition exists when the market reaches equilibrium. *Id.* Market equilibrium is achieved when market supply equals market demand. *E. Thomas Sullivan & Herbert Hovenkamp, ANTITRUST LAW, POLICY AND PROCEDURE: CASES, MATERIALS, PROBLEMS 51–54 (5th ed. 2004).*

11 *See generally In re GlaxoSmithKline PLC, 699 N.W.2d 749 (Minn. 2005).*

12 See Sherman Anti-Trust Act § 1; *N. Pac. Ry. v. United States, 356 U.S. 1, 7 (1958)* (prohibiting the defendant’s tying arrangement, which required lessees or grantees of its land to ship all commodities produced on the land over the defendant’s railroad lines); *Lorain Journal Co. v. United States, 342 U.S. 143, 152–55 (1951)* (prohibiting refusals to deal as violating section 2 of the Sherman Anti-Trust Act); *Dr. Miles Med. Co. v. John D. Park & Sons, 220 U.S. 373, 408 (1911)* (prohibiting resale price maintenance, also known as vertical price restrictions, as per se illegal because such agreements, “having for their sole purpose the destruction of competition and the fixing of prices, are injurious to the public interest and void”). An example of a vertical restraint is when a manufacturer limits distributors to an exclusive territory or allocates customers to distributors, thereby precluding the distributors from selling outside their designated areas or to customers of another distributor. *See Cont’l T.V., Inc. v. GTE Sylvania Inc., 433 U.S. 36, 59 (1977)* (holding a franchise agreement between a manufacturer of television sets and a retailer, which barred the retailer from selling franchised products from locations other than those specified in the agreement, to be an unreasonable restraint of trade as analyzed under the Rule of Reason).

13 *See Klor’s Inc., v. Broadway-Hale Stores, Inc., 359 U.S. 207, 208–09 (1959)* (concluding that a horizontal refusal to deal, or group boycott, existed, based on the fact that manufacturers and distributors had conspired with each other).

14 *See United States v. Socony-Vacuum Oil Co., 310 U.S. 150, 218 (1940)* (reaffirming the idea that horizontal price fixing is per se illegal). If so, concerted collusion would raise additional section 1 concerns as an example of a horizontal restraint—agreements between competitors within a market. *Id.; see also ABA SECTION OF ANTITRUST LAW, ANTITRUST LAW DEVELOPMENTS 104 (4th ed. 1997).*
General has claimed there is some evidence of collusion among the major pharmaceutical companies.\textsuperscript{15}  

Assuming evidence exists either of collective involvement by drug companies in limiting supplies to Canadian markets or of particular drug companies engaging in vertical territorial restraints by prohibiting distributors from selling to particular categories of customers, such actions still do not violate antitrust laws under the Rule of Reason analysis unless they have an anticompetitive effect.\textsuperscript{16}  Before a court can analyze whether such actions have an anticompetitive effect, the court must define the geographic market that is potentially affected by drug manufacturers’ actions.\textsuperscript{17}  This Note argues that the nature of the pharmaceutical industry requires courts analyzing antitrust claims against drug manufacturers to define the relevant geographic market to include Canada and the United States.\textsuperscript{18}  If courts determine the pharmaceutical companies violated antitrust laws, then drug manufacturers would have to discontinue the challenged practice—in this case the horizontal collective refusal to deal with Canadian pharmacies or the vertical territorial restraints.\textsuperscript{19}  

Part I of this Note provides a brief overview of the present status of the prescription drug industry in the United States and Canada, including their respective pricing schemes.\textsuperscript{20}  Part I also outlines some of the trends in consumption of prescription drugs by the U.S. market

\textsuperscript{15} America’s Seniors, Today’s Seniors Network.com, Documents Show Glaxo Antitrust Violations, Attorney General Hatch Says (Dec. 15, 2004), http://todaysseniorsnetwork.com/glaxo_evidence.htm. In the state’s appeal after a trial court initially decided the documents had to remain confidential, the Minnesota Attorney General argued that “[t]he 45 documents at issue . . . contain direct evidence of unlawful concerted action by GSK and other drug companies to block the importation of prescription drugs from Canada.” \textit{Id.}  

\textsuperscript{16} See Nat’l Soc’y of Prof’l Eng’rs v. United States, 435 U.S. 679, 691 (1978) (explaining that the Rule of Reason is limited to whether the restraint “is one that promotes competition or one that suppresses competition”).  

\textsuperscript{17} See infra notes 72–76 and accompanying text.  

\textsuperscript{18} See infra notes 215–328 and accompanying text.  

\textsuperscript{19} See FTC v. Ind. Fed’n of Dentists, 476 U.S. 447, 459 (1986) (holding a horizontal refusal to deal, evinced by an agreement between dentists and insurance companies, was illegal under section 1); Nat’l Soc’y of Prof’l Eng’rs, 435 U.S. at 691–95 (holding that a professional association’s canon of ethics prohibiting competitive bidding was illegal under section 1); Cont’l T.V., Inc., 433 U.S. at 46 (holding that a manufacturer’s territorial restrictions on retailers were illegal under section 1). The term “horizontal” refers to agreements among actual or potential competitors, \textit{Areeda & Hovenkamp}, supra note 10, ¶ 1901b, whereas the term “vertical” refers to agreements among those in different levels of the chain of production, such as producers, distributors, wholesalers, and retailers. \textit{Id.} ¶ 1902d.  

\textsuperscript{20} See infra notes 29–77 and accompanying text.
in comparison to the worldwide market. Part II begins by providing an overview of the antitrust legal framework. Part II then surveys three main theoretical approaches to geographic market definition in antitrust law. This Part also illustrates how the federal courts utilize each of these approaches. Finally, drawing on this background, Part III argues that courts should include Canada as part of the relevant geographic market in antitrust litigation arising when pharmaceutical companies limit drug supplies to Canadian pharmacies that participate in reimportation to U.S. customers. Thus, the relevant geographic market for antitrust purposes should consist of at least Canada and the United States, with the possibility of expansion to a worldwide scale. By this argument, the pharmaceutical companies’ refusal to deal with Canadian wholesalers and distributors could constitute action violative of Section 1 of the Sherman Act, either as a horizontal or vertical restraint of trade. Without the inclusion of Canada into an antitrust analysis, the geographic market would be limited to the United States—precluding oversight of drug manufacturers’ potentially anticompetitive behavior.

I. Present Status of the Prescription Drug Industry in the United States and Canada

A. Spending, Price, and Prescription Drug Markets

In the modern global community, the pharmaceutical industry has undergone a series of dramatic changes. For example, several large-scale international mergers have created global drug companies. Despite the expanded reach of drug manufacturing companies, these companies have narrowed their focus to the U.S. markets

21 See infra notes 31–42, 50, and accompanying text.
22 See infra notes 78–95 and accompanying text.
23 See infra notes 96–195 and accompanying text.
24 See infra notes 96–195 and accompanying text.
25 See infra notes 215–328 and accompanying text.
26 See infra notes 215–328 and accompanying text.
27 See Ind. Fed’n of Dentists, 476 U.S. at 459 (1986) (holding that a horizontal refusal to deal between dentists and insurance companies was illegal under section 1); Cont’l T.V., Inc., 433 U.S. at 46 (holding that a manufacturer’s territorial restrictions on retailers were illegal under section 1).
28 Areeda & Hovenkamp, supra note 10, ¶ 530a.
30 Id.
as a major source of overall revenue and profits generation.\textsuperscript{31} The U.S. share of drug makers’ worldwide revenues rose from almost one-third of global revenues in 1996 to almost one-half in 2002—an increase of fifty percent.\textsuperscript{32} Sales in the United States were roughly ninety-five percent of the North American total.\textsuperscript{33} Attempting to capitalize on this reality, the global drug manufacturers have adopted a global perspective to their businesses.\textsuperscript{34} The pharmaceutical companies’ reliance on global business plans is most evident in two aspects of their businesses: their financing structure of research and development for new drugs, and their global structure for manufacturing drugs.\textsuperscript{35} With the U.S. drug market now comprising one-half of the drug manufacturers’ total revenues, the pharmaceutical industry relies on U.S. revenues to support the expensive process of researching and developing new drugs for use throughout the world.\textsuperscript{36}

While pharmaceutical companies reaped greater profits from the U.S. market in the late 1990s, U.S. consumers also found themselves spending more and more on prescription drugs.\textsuperscript{37} Between 1990 and 2003, spending on prescription drugs by the U.S. public increased

\textsuperscript{31} Alan Sager & Deborah Socolar, Health Reform Program, Lower U.S. Prescription Drug Prices Are Vital to Both Patients and Drug Makers—But Instead, U.S. Prices Have Been Rising Rapidly Relative to Those in Other Wealthy Nations 5–6 (2003), http://dcc2.bumc.bu.edu/hs/pdfs/Lower%20drug%20prices.pdf (revealing that U.S.-generated revenues for drug manufacturers during the period from 1996 to 2002 increased from 34.7% to 50.8%).

\textsuperscript{32} Id. at 6.

\textsuperscript{33} Id. at 5.


\textsuperscript{35} See, e.g., Pierce, supra note 34, at 22, 25–26 (quoting GlaxoSmithKline representatives as having a “global” manufacturing and packaging outfit); Careers with Pfizer Global Manufacturing (PGM), supra note 34 (describing Pfizer’s global manufacturing plants).

\textsuperscript{36} Sager & Socolar, supra note 31, at 5–6 (estimating that drug manufacturers secure about two-thirds to three-fourths of their profits from U.S. citizens and that some industry sources have found drug manufacturers’ investment in research and development to be roughly equal to their U.S. profits). In fact, drug manufacturers’ reliance on the U.S. market for providing the pharmaceutical industry with the sufficient funds to maintain current levels for research and development of new drugs is one of the main justifications provided by the industry for the price differentials. Pharm. Research & Mfrs. of Am., The Minnesota Attorney General’s Report on Pharmaceuticals: Correcting the Record 2–3 (2003), available at http://www.phrma.org/publications/policy/2003-10-31.862.pdf.

\textsuperscript{37} Sager & Socolar, supra note 31, at 4–5.
almost five-fold, from $40.3 billion to $189.1 billion.\footnote{38} The reasons for this increase are varied, ranging from higher prices on existing drugs, the introduction of new drugs with high retail prices, and changes in the rate of drug usage.\footnote{39}

In addition to increased domestic spending on prescription drugs, U.S. consumers also pay higher prices than do foreign consumers for the same pharmaceutical products.\footnote{40} The fundamental cause of pricing differentials between the U.S. market and other foreign markets relates to the respective countries’ wholesale drug pricing systems.\footnote{41} In other industrialized countries, such as Canada and European Union nations, the national governments regulate the wholesale price charged by the drug manufacturers.\footnote{42} For example, in Canada, the national Patented Medicine Prices Review Board (the “PMPRB”) regulates the prices of patented drugs.\footnote{43} The PMPRB establishes the maximum prices that drug manufacturers can charge in Canada for their patented drugs.\footnote{44} To arrive at the wholesale price for each drug, the PMPRB considers several factors, including the median price charged in other specified industrial countries and a comparison of price increases to the Canadian Consumer Price Index (the “CPI”).\footnote{45} In addition to setting a target wholesale price, the PMPRB also polices the wholesale prices that drug manufacturers quote to wholesalers, hospitals, and pharmacies, in an effort to ensure that the price is not excessive in comparison to the PMPRB’s calculated price.\footnote{46}

\footnote{38} Id. at 4.  
\footnote{39} Id. at 5.  
\footnote{40} Id.  
\footnote{42} Id.  
\footnote{45} Id. CPI is the measure of consumer goods and services. Prescription Drug Coverage for Seniors: Hearing Before the House Comm. on Commerce, Subcomm. on Health and Environment, 106th Cong. 2 (1999) (statement of the American Academy of Actuaries), available at http://www.actuary.org/pdf/medicare/rxstatement.pdf. In the United States, the CPI increased 2.3 percent between 1989 and 1999. Id. In comparison, the CPI for prescription drugs and medical supplies increased 5.9 percent during the same period. Id.  
In the United States, however, Congress simply requires each drug manufacturer to report their average wholesale price (the “AWP”) to third-party compilers, such as the Drug Topics Red Book, American Druggist First Database Annual Director of Pharmaceuticals, or the Essential Director of Pharmaceuticals (Blue Book), which in turn report AWP data for use by healthcare professionals, Medicare, and Medicaid for reimbursement calculations.  

47 Each prescription drug’s AWP represents the average price at which manufacturers sell drugs to physicians, pharmacies, and other customers.  

48 Despite its name, however, the AWP is not an accurate reflection of actual market prices for drugs; rather, it is a price derived from self-reported manufacturer data for both patented and generic drugs, with no external oversight or determination that such prices accurately reflect drug manufacturers’ actual costs.  

49 There are no governmental requirements or industry-wide conventions requiring the AWP to reflect the price of any actual sale of drugs by a manufacturer.  

50 Thus, according to the U.S. Government Accountability Office, the AWP may be neither “average” nor “wholesale.”  

51 In a recent investigation, the U.S. Department of Justice (the “DOJ”) and the National Association of Medicaid Fraud Control Units further debunked the illusion of accurate wholesale price reporting.  

52 By comparing actual wholesale pricing information, based upon wholesalers’ price lists, with the AWP price reported by the drug manufacturers, investigators found numerous instances in which the AWP price was considerably higher than the actual wholesale price.  

53 The study’s findings of inflated pricing were utilized in 2003, in In re Pharmaceutical Industry Average Wholesale Price Litigation, when the


48 Id.  

49 Id. at 3.  

50 Id.  


53 Id.
U.S. District Court of Massachusetts evaluated a class action complaint alleging fraud and violation of section 1 of the Sherman Anti-Trust Act.\textsuperscript{54} The complaint detailed specific discrepancies between the actual wholesale price, as calculated by the DOJ, and the AWP indicated by the drug manufacturer.\textsuperscript{55} At times, the data revealed the drug manufacturer’s AWP was thousands of percentage points higher than the DOJ-calculated wholesale price.\textsuperscript{56}

This difference between the AWP and the actual wholesale price, also known as the “spread,” allows drug manufacturers to realize profits in the United States that elude them in other markets, where governmental regulations force drug manufacturers to charge wholesale prices closer to actual cost, thereby decreasing the spread in those markets.\textsuperscript{57} As a result, U.S. consumers face higher retail costs for prescription drugs than do foreign consumers whose governments prevent such pricing abuses—giving U.S. consumers reason to view reimportation as a way to benefit from the lower prices charged in foreign countries.\textsuperscript{58}

The increased U.S. demand for prescription drugs, coupled with higher relative prices as compared to other countries, has led the U.S. public to seek means of obtaining cheaper prescription drugs from foreign countries.\textsuperscript{59} One approach taken by U.S. consumers is to pressure their state and local governments to propose legislation and policies allowing for reimportation; state and local communities, such as Illinois, Iowa, Maine, Michigan, Minnesota, and Springfield, Massachusetts, have already proposed measures to establish drug reimport-
tation programs for state employees and retirees. Meanwhile, another approach taken by U.S. consumers is to reimport prescription drugs personally. Typically, these individuals obtain these prescription drugs either by physically traveling across the border or by ordering from online Canadian pharmacies. The majority of reimported drugs are mailed from Canada into the United States, with recent estimates placing the total volume at approximately twelve million prescription drug products, valued at approximately $700 million in 2003 alone. Additionally, about the same amount of prescription drugs enter the United States from the rest of the world through mail and courier services offered by traditional Canadian pharmacies. In recent years, estimates suggest that up to one million Americans a year travel across the border to purchase their prescription drugs from Canada for a fraction of the U.S. domestic price.


62 See HHS Task Force on Drug Imp., supra note 4, at 11–12 (stating that in 2003, about $408 million of the total prescription drug sales to U.S. consumers from Canada was derived from Internet pharmacy sales, while the remaining $287 million was due to personal travel). Many Canadian online pharmacies actively cater to U.S. consumers by allowing a consumer to place an order that will be delivered to the consumer’s home simply by inputting a drug name and quantity and filling out an order form. See generally Canada Pharmacy Home Page, http://www.canadapharmacy.com (last visited Sept. 17, 2005); Canadian Pharmacy Trust Home Page, http://www.canadianpharmacytrust.com (last visited Sept. 17, 2005); CanadaRX Home Page, http://www.canadarx.com (last visited Sept. 17, 2005). Recent speculation has suggested that the Canadian government may alter its domestic laws to prohibit individuals from receiving prescription medication unless they have first met with a Canadian doctor. Joel B. Finkelstein, Drug Reimportation Situation Is Shifting as Canada Could Cut Availability, Am. Med. News (Am. Med. Ass’n, Chi., Ill.), Jan. 24, 2005, at 5, available at http://www.ama-assn.org/amednews/2005/01/24/gvs30124.htm. If so, then those currently reimporting prescription drugs via the Internet may lose their current sources. Id.

63 See HHS Task Force on Drug Imp., supra note 4, at 11–12.

64 See id. at 12.

B. Prescription Drugs Under the Federal Antitrust Statutory Framework

The pricing structure of pharmaceutical companies in the U.S. market has spurred both political and social outrage and consumer attempts to circumvent U.S. pricing. To impose legal antitrust liability for the companies’ behavior, however, courts must first determine that the pharmaceutical companies engaged in conduct that unlawfully restrains trade. As mentioned previously, such conduct could take the form either of a horizontal collective agreement among all the pharmaceutical companies to refuse to deal with Canadian wholesalers who sell to U.S. customers, or of a vertical refusal deal in which each company independently refuses to deal with particular Canadian wholesalers who sell to U.S. customers.

Although a classic horizontal group boycott is traditionally deemed per se illegal under U.S. caselaw interpreting the Sherman Act, modern courts have narrowed the per se category in favor of applying a Rule of Reason analysis. Furthermore, vertical territorial restraints are ana-

66 Rigdon, supra note 61, at 22, 24, 27.
67 See Nat’l Soc’y of Prof’l Eng’rs v. United States, 435 U.S. 679, 691 (1978) (stating that under the Rule of Reason, courts will determine whether the restraint “promotes competition or [is] one that suppresses competition”); see also Levine v. Cent. Fla. Med. Affiliates, 72 F.3d 1538, 1552 (11th Cir. 1996) (stating that “[t]he rule of reason analysis is concerned with the actual or likely effects of defendants’ behavior, not with the intent behind that behavior”); SCFC ILC, Inc. v. Visa USA, Inc., 36 F.3d 958, 969 (10th Cir. 1994) (finding intent alone insufficient to invoke antitrust laws). See generally Found. for Interior Design Educ. Research v. Savannah Coll. of Art & Design, 244 F.3d 521 (6th Cir. 2001) (holding that without proof of an antitrust injury, malicious intent alone is an insufficient basis for antitrust liability).
68 See FTC v. Ind. Fed’n of Dentists, 476 U.S. 447, 459 (1986) (holding that the horizontal refusal to deal by a group of dentists with insurance companies was illegal under section 1 of the Sherman Anti-Trust Act); Cont’l T.V., Inc. v. GTE Sylvania Inc., 433 U.S. 36, 46 (1977) (holding that a manufacturer’s territorial restrictions on retailers were illegal under section 1).
69 Compare United States v. Gen. Motors Corp., 384 U.S. 127, 145 (1966) (holding an agreement between a car manufacturer and dealers to refuse to deal with discount dealers was per se illegal), and Klor’s, Inc. v. Broadway-Hale Stores, 359 U.S. 207, 212 (1959) (holding an agreement among retailers, wholesalers, and manufacturers to refuse to deal with retail competitor was per se illegal), with Nw. Wholesaler Stationers, Inc. v. Pac. Stationery & Printing Co., 472 U.S. 284, 296 (1985) (holding the Rule of Reason applied to a wholesale purchaser cooperative’s decision to expel a member), and Ind. Fed’n of Dentists, 476 U.S. at 459–65 (holding the “quick look” Rule of Reason applied to a refusal to deal with insurance companies on the part of a group of dentists). The Court has stated that horizontal refusals to deal will remain per se illegal if the boycotting party “possesses market power or exclusive access to an element essential to effective competition.” Nw. Wholesalers Stationers, 472 U.S. at 296.
lyzed under the Rule of Reason as well.\textsuperscript{70} Thus, regardless of whether drug manufacturers are targeted under antitrust statutes for collectively refusing to deal with Canadian distributors or for imposing vertical territorial restraints, courts will most likely engage in a Rule of Reason analysis.\textsuperscript{71}

Essentially, the Rule of Reason analysis seeks to determine whether the restraint’s anticompetitive effects substantially outweigh the procompetitive justifications offered by the defendant.\textsuperscript{72} In most Rule of Reason cases, the plaintiff must prove that the identified restraint is likely to have a substantial, adverse impact on competition, as indicated through a market analysis of the restraint’s effects.\textsuperscript{73} Such market analysis begins with a definition of the relevant market to determine the relative impact of the contested action on the overall market.\textsuperscript{74} In the antitrust context, the relevant market is defined from two perspectives: the market for the product and the geographic market.\textsuperscript{75} The remainder of this Note describes the differing approaches to defining geographic markets and then applies these approaches to the issue of reimportation of prescription drugs.\textsuperscript{76} The Note concludes that Canada should be included as part of any geographic market considered in an antitrust claim against drug manufacturers operating in the U.S.\textsuperscript{77}

\textsuperscript{70} See Cont’l T.V., Inc., 433 U.S. at 57–59 (holding the Rule of Reason applied to vertical nonprice restraints).

\textsuperscript{71} See supra notes 69–70 and accompanying text. This Note does not discuss the “quick look” Rule of Reason, in which the plaintiff does not have to prove every aspect of anticompetitive effect, including a market power analysis. See Ind. Fed’n of Dentists, 476 U.S. at 460–61 (applying the “quick look” Rule of Reason to a group of dentists’ refusal to submit patient’s X-rays to insurance companies); NCAA v. Bd. of Regents, 468 U.S. 85, 100–03, 109 (1984) (applying the “quick look” Rule of Reason to the NCAA’s restrictions on the number of football games that members could televise and its agreement on the minimum price member schools would receive for broadcasting rights). Rather, it is sufficient to note that the Court has deemed the “quick look” version of the Rule of Reason appropriate “when the great likelihood of anticompetitive effects can easily be ascertained.” Cal. Dental Ass’n v. FTC, 526 U.S. 756, 777 (1999). As such, because it is uncertain whether the “quick look” version would be applied to actions of drug manufacturers, this Note will analyze the topic assuming the case falls under the full Rule of Reason. See infra notes 215–328 and accompanying text.

\textsuperscript{72} See Nat’l Soc’y of Prof’l Eng’rs, 435 U.S. at 691.


\textsuperscript{74} See, e.g., Times-Picayune Publ’g Co. v. United States, 345 U.S. 594, 611–13 (1953); United States v. Microsoft Corp., 252 F.3d 34, 51–52 (D.C. Cir. 2001).

\textsuperscript{75} See Brown Shoe Co. v. United States, 370 U.S. 294, 324 (1962).

\textsuperscript{76} See infra notes 81–206, 215–328, and accompanying text.

\textsuperscript{77} See infra notes 215–328 and accompanying text.
II. CURRENT APPROACHES TO GEOGRAPHIC MARKET DEFINITION

A. General Theory of U.S. Antitrust Laws

The underlying philosophy of U.S. antitrust laws is to protect free and fair market competition. Thus, when anticompetitive practices by private businesses hinder open competition to a substantial degree, antitrust laws allow courts to step in to regulate practices to protect competition.

As stated above, before a court can apply the Rule of Reason analysis and evaluate a practice’s potentially anticompetitive effects, a plaintiff first must define a legally sufficient market. Market definition has both a product and geographic aspect. Thus, the relevant product market consists of all products that are “reasonably interchangeable” with—that is, are economic substitutes for—the product at issue. The geographic market consists of the geographic area within which competition occurs.

Market definition is a critical first step in an analysis of an antitrust claim because courts determine if the defendant has market power according to the defendant’s role in the defined market. If an antitrust plaintiff fails to identify a relevant market, then the defendants are entitled to a dismissal as a matter of law.

---

78 See, e.g., United States v. Pabst Brewing Co., 384 U.S. 546, 549 (1966); United States v. Aluminum Co. of Am., 148 F.2d 416, 428–29 (2d Cir. 1945) (drawing on legislative intent to preserve competition and limit the aggregation of capital in select hands because such behavior hinders free markets).


80 See supra notes 16–17 and accompanying text.

81 Bathke v. Casey’s Gen. Stores, Inc., 64 F.3d 340, 344–45 (8th Cir. 1995) (citing Spectrum Sports, Inc. v. McQuillan, 506 U.S. 447, 459 (1993)). Although this Note deals solely with geographic market analysis, the complementary product market definition is recognized as being crucial to the Rule of Reason analysis. Id.

82 Id.


84 See Sherman Anti-Trust Act § 1; Cont’l T.V., Inc. v. GTE Sylvania Inc., 433 U.S. 36, 59 (1977) (holding franchise agreement between a manufacturer of television sets and a retailer, which barred the retailer from selling franchised products from locations other than those specified in the agreement, was an unreasonable restraint of trade under section 1 of the Sherman Anti-Trust Act as analyzed under the Rule of Reason).

85 See Am. Online, Inc. v. GreatDeals.Net, 49 F. Supp. 2d 851, 858–59 (E.D. Va. 1999); see also ABA SECTION OF ANTITRUST LAW, supra note 14, at 544–45 (citing TV Commc’ns
The Supreme Court has defined market power as “the ability to raise prices above those that would be charged in a competitive market.” Without market power, defendants cannot be liable for anticompetitive behavior, on the rationale that any anticompetitive conduct will not negatively affect the competitive marketplace. In assessing market power, most courts refer to the defendant’s market share (within the defined geographic and product markets) as a proxy for market power. The relevant geographic market area translates into the denominator out of which the particular firm’s market share is calculated. The larger the market share held by a firm, the more market power it is said to hold.

**Network v. Turner Network Television**, 964 F.2d 1022, 1028 (10th Cir. 1992), where the Tenth Circuit affirmed the defendant’s motion to dismiss because the plaintiff “did not allege a relevant product market which TNT was capable of monopolizing, attempting to, or conspiring to monopolize”).


87 See, e.g., **Gen. Leaseways, Inc. v. Nat’l Truck Leasing Ass’n**, 744 F.2d 588, 596 (7th Cir. 1984) (“[P]laintiff [must] first prove that the defendant has sufficient market power to restrain competition substantially. . . . If not, the inquiry is at an end; the practice is lawful.”) (citations omitted); **Davis-Watkins Co. v. Service Merch.**, 686 F.2d 1190, 1202 (6th Cir. 1982) (“Without market power, a firm cannot have an adverse effect on competition.”).

88 **Flegel v. Christian Hosp., Ne.-Nw.**, 4 F.3d 682, 689 (8th Cir. 1993) (“To establish . . . market power, . . . [it must be] show[n] that the defendants have a dominant market share in a well-defined relevant market.” (internal quotation and citation omitted)). For example, if a defendant has a low market share (typically below thirty percent), courts are precluded from concluding that the defendant maintains market power. See, e.g., **Capital Imaging Assocs. v. Mohawk Valley Med. Assocs.**, 996 F.2d 537, 547 (2d Cir. 1993) (holding that a market share of only 1.15% was so “de minimis” as to lack proof of market power); **Hassam v. Indep. Practice Assocs., P.C.**, 698 F. Supp. 679, 694–95 (E.D. Mich. 1988) (finding the market power held by a firm was insufficient, where the firm had a market share of only twenty percent in a market with low barriers to entry and no evidence of the defendant’s ability to impose above-market prices). On the other hand, if a defendant has a high market share, it is an indication that the defendant may command market power. See, e.g., **Graphics Prods. Distribrs., Inc. v. ITEK Corp.**, 717 F.2d 1560, 1570–71 (11th Cir. 1983) (holding that the defendants’ average seventy-percent market share was sufficient to prove market power in a non-price, vertical restraint on trade); **Barrett v. Fields**, 924 F. Supp. 1063, 1075 (D. Kan. 1996) (finding the defendants’ fifty-percent market share was sufficient to prove market power in case regarding an attempted conspiracy to monopolize).


90 See id. Market power is the ability of a firm to obtain higher profits by reducing output and selling at a higher price. Areeda & Hovenkamp, *supra* note 10, ¶ 501. A firm’s high market power corresponds to its competitors’ inability to constrain the firm’s pricing above market equilibrium, skewing the market away from perfect competition. Id.
Geographic market definition is also a key strategic decision for the parties. Typically, a large geographic area decreases the probability of finding an anticompetitive practice by the targeted firm because it will have little market power to affect the overall market. Alternatively, a smaller geographic market increases the firm’s market power, which raises the potential that the reviewing court will find that the market can be manipulated for anticompetitive purposes. For this reason, parties often attempt to manipulate the boundaries of the geographic market as a means of dictating the direction of the rest of the controversy. Thus, it is crucial that the court define the relevant geographic market as accurately as possible; failure to do so could result in a misinterpretation of the antitrust claim.

B. Courts’ Various Approaches to Geographic Market Definition

In 1961, in *Tampa Electric Co. v. Nashville Coal Co.*, the U.S. Supreme Court articulated the standard for determining the relevant geographic market area as “the market area in which the seller operates, and to which the purchaser can practically turn for supplies.” Using this standard to identify a geographic region is difficult because it requires both capturing the relevant forces of supply and demand

---

91 See, e.g., *Pabst Brewing Co.*, 384 U.S. at 549 (defendant urged the court to adopt a national geographic market to gain clearance for the proposed merger, whereas the government urged a localized geographic market, limited to the state of Wisconsin, in an attempt to block the merger); *United States v. Eastman Kodak Co.*, 63 F.3d 95, 103, 105 (2d Cir. 1995) (defendant urged the court to adopt a global geographic market, so as to dilute its market power, whereas the government urged the court to adopt a national geographic market so as to increase the defendant’s market power). The geographic market defines the group of competitors that consumers may turn to as substitutes if a supplier were to raise prices. *Tampa Elec.*, 365 U.S. at 327. Thus, the geographic market represents the competitors that operate as a pricing constraint on the firm in question. *Id.*

92 *Eblen*, *supra* note 89, at 54–55.

93 *Id.*

94 *Id.*

95 See *ABA Section of Antitrust Law*, *supra* note 14, at 532–43. In their attempts to define the geographic market as accurately as possible, courts tend to focus on six major factors. *Id.* at 533–39. First, actual sales patterns are used to determine whether two areas are within the same market. *Id.* at 533–34. Second, evidence of parallel price movements may suggest two areas are part of the same market. *Id.* at 536. Third, transportation costs, in relation to the price of the product, are an important consideration in determining whether two areas are part of the same market. *Id.* at 536–37. Fourth, governmental barriers to trade, such as licenses, quotas, and tariffs, can limit competition between areas, signifying two separate geographic markets. *Id.* at 537–38. Fifth, industry and firm practices also help courts define the geographic market. *Id.* at 538. Finally, the nature and scope of the anticompetitive effect at issue can also determine the likely market. *Id.* at 538–39.

96 365 U.S. at 327.
and, at the same time, effectively limiting the market so that the actual market power of each firm is measured.\textsuperscript{97} Therefore, a court applying this standard must undertake a detailed examination of the particular supply and demand forces at play in each case.\textsuperscript{98} Although each case must be decided according to its unique factual circumstances, the Court has been inconsistent in identifying and weighing the economic factors to be considered within the \textit{Tampa Electric} framework.\textsuperscript{99}

Such analytic inconsistency has led the Supreme Court in opposing directions at times.\textsuperscript{100} For example, in \textit{Tampa Electric} itself, even though the Court articulated a standard that accounts for both sides of the market, the Court’s ensuing analysis focused solely on the supplier’s side.\textsuperscript{101} \textit{Tampa Electric} involved a utility company, Tampa Electric, which sued the Nashville Coal Company for failure to perform its portion of a requirements contract.\textsuperscript{102} Among other provisions, the contract required Nashville Coal to supply Tampa Electric’s entire demand for coal for two utility stations for twenty years.\textsuperscript{103} Near the expected performance date of the contract, Nashville Coal informed Tampa Electric that it would not deliver the coal because it believed the contract violated antitrust laws.\textsuperscript{104} The Supreme Court diverged from the district court’s finding that the geographic market consisted

\textsuperscript{97} \textit{See} Brown Shoe Co. v. United States, 370 U.S. 294, 338–39 (1962) (defining the geographic market to include suppliers in adjacent suburban areas of large metropolitan areas, but to exclude suppliers outside of the immediate reach of the metropolis). In \textit{Brown Shoe Co. v. United States}, the Court stressed that Congress envisioned a “pragmatic, factual approach to the definition of the relevant market and not a formal, legalistic one.” \textit{Id.} at 336. Furthermore, “[t]he geographic market selected must, therefore, both correspond to the commercial realities of the industry and be economically significant. Thus, although the geographic market in some instances may encompass the entire Nation, under other circumstances it may be as small as a single metropolitan area.” \textit{Id.} at 336–37 (internal quotations and citations omitted).

\textsuperscript{98} \textit{See id.} at 331 (describing the identified geographic market as including Tampa-area consumers and national coal producers).


\textsuperscript{100} Compare United States v. Grinnell Corp., 384 U.S. 563, 575–76 (1966) (identifying a national market due to the defendants’ business operations despite the fact that they provided localized services), \textit{with} Phila. Nat’l Bank, 374 U.S. at 358–61 (identifying a local four-county market due to the localized nature of providing banking services to consumers despite the fact that national banking competitors operated in the same area).

\textsuperscript{101} 365 U.S. at 331–32 (“[T]he relevant competitive market . . . is of course the area in which respondents and the other 700 producers effectively compete.”).

\textsuperscript{102} \textit{Id.} at 324.

\textsuperscript{103} \textit{Id.} at 322.

\textsuperscript{104} \textit{Id.} at 323.
of only peninsular Florida. Instead, the Court expanded the geographic market area to include several other states that were home to additional producers of coal, available to Tampa Electric. By focusing on supply and expanding the geographic market on that basis alone, the Court only implicitly recognized a national demand and thereby rendered demand analysis seemingly irrelevant to defining a geographic market.

Similarly, in 1963, in *United States v. Philadelphia National Bank*, the Supreme Court again held that supply forces were critical in identifying the relevant geographic market. In *Philadelphia National Bank*, the government sought to enjoin a proposed merger between Philadelphia National Bank, the second largest bank in the Philadelphia metropolis, and Girard Trust Corn Exchange Bank, the third largest bank, as illegal under section 1 of the Sherman Act and section 7 of the Clayton Act. The Court held the proposed merger void under the Clayton Act because it would have substantially lessened competition. In doing so, the Court concluded that the relevant geographic market was a four-county area, reasoning that the banks (the suppliers) transacted very little business with customers outside of that area.

By contrast, in 1966, in *United States v. Pabst Brewing Co.*, the Supreme Court defined the geographic market by focusing on the market demand forces instead of supply forces. In that case, the defendant beer company defined the proper geographic market as the entire United States, whereas the government argued for a smaller market of only Wisconsin or a three-state region of Wisconsin, Illinois, and Michigan. The Court held that due to the high demand for Pabst beer among Wisconsin state residents, the proper geographic market should be limited to Wisconsin.

Given the Court’s inconsistency in focusing alternately on demand or on supply forces when defining the relevant geographic market in a

105 *Id.* at 331.
106 *Tampa Elec.*, 365 U.S. at 331–32.
107 *See id.*
111 *Id.* at 359. The Supreme Court did not define the geographic market as where the banks do business or where they competed with each other, but rather as where the effect of the merger would have been “direct and immediate.” *Id.* at 357.
112 *See* 384 U.S. at 550–52.
113 *Id.* at 550.
114 *Id.* at 550–52.
given case, scholars have taken the lead in attempting to formulate workable solutions that properly and uniformly capture the economic supply and demand forces. Three distinct approaches to defining the relevant geographic market area have emerged: the “shipments” approach, the “diversion” approach, and the approach articulated in the U.S. Department of Justice’s 1992 Merger Guidelines.

1. Shipments Approach

Guided by the principle in *Tampa Electric* that the geographic market is comprised of the area in which buyers and sellers operate, scholars developed the shipments approach to focus on the physical locations to which and from which suppliers send shipments. Proponents of this school of thought believe that all economic factors affecting price also correlate to the quantity shipped. Thus, they use shipping figures as an appropriate proxy indicator to estimate the geographic market.

Under this methodology, the shipment data is first classified into two categories: destination and origin. From this starting point, the analysis identifies the areas from where the majority of the product is shipped and where the majority of goods are shipped to consumers. Proponents suggest that measuring these patterns captures both the demand and the supply for the product within a geographic area. Once that area is defined, the total consumption of shipment in the entire geographic area is calculated to arrive at the total market volume. The total market volume then functions as the denominator from which to calculate a particular firm’s market share.

---


117 *Tampa Elec.*, 365 U.S. at 327; Elzinga & Hogarty, *supra* note 115, at 73.

118 Elzinga & Hogarty, *supra* note 115, at 73–76.

119 *Id.*

120 *Id.* at 73.

121 *Id.*

122 *Id.*

123 Elzinga & Hogarty, *supra* note 115, at 73–76.

124 *Id.* at 74.
Some federal appellate courts have utilized this approach to define the relevant geographic market.\textsuperscript{125} For example, in 1990, in \textit{United States v. Rockford Memorial Corp.}, the Seventh Circuit Court of Appeals held that the shipments approach should be used to define the geographic market.\textsuperscript{126} In defending a proposed merger between two of the largest hospitals in Rockford, Illinois, the defendant hospitals defined the geographic market as a ten-county region based on the areas in which their patients lived.\textsuperscript{127} The Seventh Circuit, however, defined the relevant geographic market area as the sole county where the hospitals were located.\textsuperscript{128} In doing so, the Seventh Circuit cited the district court’s finding that the supply of hospital services provided by the defendant went to approximately eighty-seven percent of the patients in Rockford and Winnebago counties.\textsuperscript{129} Likewise, approximately eighty-three percent of the patient demand in Rockford and Winnebago counties was directed to these hospitals.\textsuperscript{130} In focusing on the physical locations of the supply and demand forces of the industry, the court thus adopted the shipments approach.\textsuperscript{131}

More recently, in 1995, in \textit{United States v. Eastman Kodak Co.}, the Second Circuit Court of Appeals also adopted the shipments approach to define the relevant geographic market.\textsuperscript{132} In that case, Kodak brought suit to modify or terminate consent decrees that it had entered into in 1921 and 1954 to rectify antitrust violations that existed at that time.\textsuperscript{133} In holding that a worldwide geographic market existed for the sale of photographic film, the Second Circuit noted that foreign manufacturers supplied one-third of U.S. film.\textsuperscript{134} Characterizing this amount as a “significant” foreign presence in the sale of film, the court turned to the rationale of the shipments approach to

\textsuperscript{125} See, e.g., \textit{Eastman Kodak Inc.}, 63 F.3d at 102–05; \textit{United States v. Rockford Mem’l Corp.}, 898 F.2d 1278, 1284–85 (7th Cir. 1990).
\textsuperscript{126} 898 F.2d at 1284–85.
\textsuperscript{127} Id.
\textsuperscript{128} Id.
\textsuperscript{129} Id.
\textsuperscript{130} Id.
\textsuperscript{131} See \textit{Rockford Mem’l Corp.}, 898 F.2d at 1284–85; Elzinga & Hogarty, supra note 115, at 74 (defining a market based on where the services were directed to, in addition to from where the services were supplied).
\textsuperscript{132} \textit{Eastman Kodak Co.}, 63 F.3d at 103 (citing the Elzinga & Hogarty shipments approach in its geographic market analysis).
\textsuperscript{133} Id. at 97.
\textsuperscript{134} Id. at 104.
hold that shipment patterns of both domestic and foreign manufacturers supported a worldwide geographic market.\footnote{Id. at 103. The court also uses the diversion approach to reach the same conclusion.}

Although the shipments approach is useful to capture historic and current market behavior, some critics fault this methodology for failing to consider the future effects on the geographic market if a firm attempts to gain further market power.\footnote{See, e.g., FTC v. Freeman Hosp., 69 F.3d 260, 269 (8th Cir. 1995); Morgenstern v. Wilson, 29 F.3d 1291, 1296 (8th Cir. 1994); A.A. Poultry Farms v. Rose Acre Farms, 881 F.2d 1396, 1403 (7th Cir. 1989); cf. Illinois ex rel. Hartigan v. Panhandle E. Pipe Line Co., 730 F. Supp. 826, 901 (C.D. Ill. 1990) (market defined narrowly because pipeline’s captive customers could not readily switch suppliers), aff’d sub nom. Illinois ex rel. Burris v. Panhandle E. Pipe Line Co., 935 F.2d 149 (7th Cir. 1991).} Because the shipments approach measures the elasticity of demand solely by the demand of current consumers, the methodology fails to capture the effect of the existence of practicable alternatives that would be available to consumers should prices increase.\footnote{Freeman Hosp., 69 F.3d at 269.} Acknowledging this criticism, courts have held that practicable alternatives to the potential suppliers should be included in any attempt to define the geographic market.\footnote{Tampa Elec., 365 U.S. at 327.} Thus, although some lower courts have adopted the shipments approach, others have rejected it, citing this deficiency.\footnote{See, e.g., Casey’s Gen. Stores, Inc., 64 F.3d at 344–45; United States v. Mercy Health Servs., 902 F. Supp. 968, 978 (N.D. Iowa 1995).}

For example, in 1994, in \textit{Morgenstern v. Wilson}, the Eighth Circuit Court of Appeals held that the relevant geographic market should be defined to include areas beyond those evident from the current market structure.\footnote{29 F.3d at 1296.} In that case, Dan Morgenstern, a cardiac surgeon, alleged actual monopolization in violation of section 2 of the Sherman Anti-Trust Act by the defendants, other cardiologists and cardiac surgeons who were members of a professional corporation that referred patients internally.\footnote{Id. at 1294; see Sherman Anti-Trust Act § 2, 15 U.S.C.A. § 2 (West 2005).} Morgenstern focused on where patients actually traveled for such medical treatment and therefore defined the relevant geographic market as Lincoln, Nebraska, and twenty-six surrounding counties, excluding Omaha, Nebraska.\footnote{Morgenstern, 29 F.3d at 1296. Morgenstern’s private practice was located in Lincoln, Nebraska, which is why he centered his geographic market around that area. Id. at 1293, 1296.} The defendants, on the other hand, defined the geographic market to include, at a minimum,
Omaha, Nebraska.\textsuperscript{143} The court determined that Omaha was a practicable alternative for patients in Lincoln, even if these consumers were not currently turning to Omaha.\textsuperscript{144} After expanding the market to include Omaha, the court reasoned that the defendants lacked sufficient market power to function as a monopoly.\textsuperscript{145}

Likewise, in 1995, in \textit{Federal Trade Commission v. Freeman Hospital}, the Eighth Circuit reaffirmed its holding that the geographic market must include potential suppliers, even if they are not currently utilized by consumers.\textsuperscript{146} The Federal Trade Commission (the “FTC”) challenged the proposed merger between two hospitals in Joplin, Missouri, as a violation of section 7 of the Clayton Act.\textsuperscript{147} The Eighth Circuit affirmed the district court’s conclusion that the FTC’s proposed geographic market definition, which was limited to a twenty-seven mile radius around Joplin, was an unreasonably static approach.\textsuperscript{148} The court acknowledged that adopting the FTC’s approach would mean the geographic market was limited to only a snapshot of the current market.\textsuperscript{149} Because these potential suppliers exerted pressure on the defendant by forcing the defendant to maintain competitive prices, these potential suppliers were, in reality, a decisive factor in the market.\textsuperscript{150}

2. Diversion Approach

Because of the failure of the shipments approach to account for future effects caused by a firm’s current practices, some scholars advocate adopting a diversion approach.\textsuperscript{151} This approach defines market power as “the ability to set price above marginal cost.”\textsuperscript{152} The underlying economic premise of this approach is that any firm with significant market share would be able to set prices above perfect

\textsuperscript{143} Id.
\textsuperscript{144} Id. The court included Omaha as part of the geographic market by relying on the ease of patients in Lincoln to reach Omaha for cardiac procedures, vigorous competition between health care providers in Lincoln and Omaha, patient referrals by health care professionals in Lincoln to professionals in Omaha for better care, and Morgenstern’s own practice of performing cardiac surgery in Omaha by commuting from Lincoln. \textit{Id.} at 1297.
\textsuperscript{145} Id. at 1297.
\textsuperscript{146} 69 F.3d at 269.
\textsuperscript{147} Id. at 262–63; see \textit{Clayton Act} § 7, 15 U.S.C. § 18 (2000).
\textsuperscript{148} \textit{Freeman Hosp.}, 69 F.3d at 269.
\textsuperscript{149} Id.
\textsuperscript{150} See \textit{Landes & Posner, supra} note 115, at 947.
\textsuperscript{151} See \textit{id.} at 938.
\textsuperscript{152} Id. at 939. This concept is formalized in traditional economic theory as the Lerner Index, which measures the “proportional deviation of price at the firm’s profit-maximizing output from the firm’s marginal cost at that output.” \textit{Id.}.
market competition circumstances because consumers are unable to turn to cheaper competitors. The diversion approach illustrates how the elasticity of demand can affect a firm’s supply decisions. A high elasticity of demand indicates greater ease with which consumers are able to switch to a competitor’s products at the slightest increase in price, forcing firms to maintain price levels on par with those of their competitors. Alternatively, if a firm faced an inelastic demand for its product, it would act as a rational profit-maximizing entity by raising prices and reducing the quantity produced. Such actions would lower the firm’s total costs while increasing profits through higher prices charged to consumers. Consequently, the smaller the market share held by a firm, the greater elasticity of demand facing the firm. A high elasticity of demand may also correlate to a higher elasticity of supply. In terms of market power, a high elasticity of supply manifests itself as small increases in price by any particular firm generating large increases in output production by competing firms, because consumer demand will switch from the high-priced product to the competitor’s low-priced product.

Although most market share analyses measure the elasticities of supply and demand, the unique aspect of the diversion approach is that it incorporates the firm’s entire sales when calculating its market power in the local market. The rationale behind this approach is that if a firm has entered a local market with even a single sale, it has done so after calculating the associated costs and determining it to be profitable. Thus, if the local market’s price should rise, then according to the diversion theory, the firm would simply divert its production away from its other markets and to the local market. Under this theory, a distant firm’s total production can affect the local geo-

153 Id. at 949–50.
154 Id. at 941–42.
155 Landes & Posner, supra note 115, at 941–42.
156 Id. at 942.
157 Id.
158 Id.
159 Id.
160 See Landes & Posner, supra note 115, at 942, 945.
161 Id.
162 Id. at 962–65.
163 Id. at 962.
164 Id. at 963.
graphic market’s elasticity of supply because the distant firm can divert its total production to the local market at the first sign of a profitable price change.\textsuperscript{165} Because the distant seller has overcome transportation or other distance-related costs to sell even one unit in the local market, it follows that the distant seller would not incur additional prohibitive costs if it chose to sell additional quantities in the local market.\textsuperscript{166} Following this diversion theory, foreign production often is included in calculating the domestic firm’s market share.\textsuperscript{167}

In addition to using the shipments approach in the 1995 \textit{Eastman Kodak} case, the Second Circuit also drew upon the diversion theory’s principles.\textsuperscript{168} In looking at the photographic film sale market, the court identified five major manufacturers, naming Kodak as the sole domestic producer.\textsuperscript{169} The court reasoned that because the supply of film is elastic and foreign film manufacturers had already established a presence in the U.S. market, foreign film manufacturers would respond to any attempt by Kodak to raise prices by diverting their production to the U.S. market from other locations, allowing them to absorb the increase in consumer demand for cheaper film.\textsuperscript{170} With Kodak’s major competition coming from foreign film manufacturers, the court held that a worldwide market, as opposed to a national market, best captured the market pressures facing Kodak.\textsuperscript{171} Thus, the diversion approach defines the relevant geographic market by taking into account the elasticity of demand faced by the particular firm in order to incorporate potential competition faced by the firm.\textsuperscript{172}


In the 1992 Merger Guidelines, the DOJ and the FTC offered yet another approach to geographic market definition, which centers on the opportunity for price discrimination.\textsuperscript{173} Similar to the diversion theory, this approach heavily weighs the effect of price changes on a market, although it does not focus as much on the cross-elasticity of

\textsuperscript{165} Landes & Posner, \textit{supra} note 115, at 964.
\textsuperscript{166} \textit{Id.}
\textsuperscript{167} \textit{Id.} at 968.
\textsuperscript{168} 63 F.3d at 103 (citing Landes & Posner, \textit{supra} note 115, at 960–66).
\textsuperscript{169} \textit{Id.} at 98.
\textsuperscript{170} \textit{Id.}
\textsuperscript{171} \textit{Id.} at 105.
\textsuperscript{172} \textit{See supra} notes 151–71 and accompanying text.
The relevant geographic market includes firms with the ability, actual or potential, to increase prices without losing so many buyers that the price increase is unprofitable.\textsuperscript{174} To arrive at this definition of the market, a court would estimate the effects of a hypothetical increase in price.\textsuperscript{175} If enough buyers would turn to a cheaper alternative, then the location of that alternative supplier is included in the geographic market, and additional alternatives are analyzed and included until an area is identified where a slight increase would not affect the purchasers’ choice of supplier.\textsuperscript{177} Although this definition leaves open the possibility of a worldwide geographic market, the reality of finding such an expansive market is diminished because the 1992 Merger Guidelines do not allow for geographic price discrimination.\textsuperscript{178} Therefore, if the company can differentiate prices on a geographic basis, then the smaller regions in which the company has power to set the prices will be recognized as separate geographic markets.\textsuperscript{179}

Courts are increasingly adopting this approach.\textsuperscript{180} For instance, in 1989, in \textit{FTC v. Elders Grain, Inc.}, the Seventh Circuit Court of Appeals affirmed an injunction preventing the acquisition of one industrial dry corn manufacturer by another.\textsuperscript{181} The defendants argued that the targeted company was not in the same geographic market because their plants were located in Kansas and the acquirer’s plants were in Indiana.\textsuperscript{182} The defendants argued that each plant supplied a different demand market by virtue of their locations on different sides of the Mississippi River.\textsuperscript{183} The court, however, viewed the geographic market as covering the entire nation.\textsuperscript{184} In defining the geographic market as nationwide, the court followed the 1992 Merger

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{174} See id. § 1.0.
\item \textsuperscript{175} See id.
\item \textsuperscript{176} Id. § 1.21.
\item \textsuperscript{177} Id.
\item \textsuperscript{178} See 1992 Merger Guidelines, supra note 116, § 1.22, reprinted in 4 Trade Reg. Rep. (CCH) ¶ 13,104.
\item \textsuperscript{179} Id.
\item \textsuperscript{180} See, e.g., \textit{Eastman Kodak}, 504 U.S. at 455, 459–79; FTC v. Elders Grain, Inc., 868 F.2d 901, 906–07 (7th Cir. 1989); FTC v. Owens-Ill., Inc., 681 F. Supp. 27, 51 (D.D.C.) (explaining the relevant geographic market did not include foreign producers because although foreign producers may have met domestic inelastic demands, foreign production would not have been significant in the short term because foreign producers were not a significant reservoir of capacity), \textit{vacated as moot} 850 F.2d 694 (D.C. Cir. 1988).
\item \textsuperscript{181} Elders Grain, Inc., 868 F.2d at 907–08.
\item \textsuperscript{182} Id. at 906–07.
\item \textsuperscript{183} Id.
\item \textsuperscript{184} Id. at 906 (“The defendants and everyone else in their industry ship industrial dry corn all over the United States.”).
\end{itemize}
\end{footnotesize}
Guidelines by emphasizing the price-setting power of competitors. Because there were only five suppliers in the United States, the court reasoned that when faced with increased prices from any one manufacturer of dry corn, consumers would turn to any alternative, regardless of that supplier’s location. As further evidence of a national market, the court noted that minimal shipping costs allowed suppliers on either side of the Mississippi to meet national demand. The court thus viewed both companies as operating in the same geographic market because either could supply the same product to buyers nationwide if the other increased prices. The court held that if the acquisition were allowed, then an alternate supply source would be eliminated for buyers.

Yet courts have not uniformly accepted the 1992 Merger Guidelines’ approach to geographic market definition. For example, in Eastman Kodak, the Second Circuit rejected this approach. Attempting to follow the 1992 Merger Guidelines’ approach, the government in that case had pointed out Kodak’s ability to engage in price discrimination within the U.S. market by charging a premium. The government alleged Kodak’s ability to engage in price discrimination meant that a global market definition would be too broad. The court rejected this argument, reasoning that although Kodak’s domestic wholesale prices were higher than its foreign wholesale prices, there was no evidence to suggest that Kodak’s costs were uniform throughout the world. The court concluded that without uniformity of costs, price differentials among geographic regions could not definitively represent multiple markets; additional entry costs experienced in one region but not the next could account for higher prices from region to region.

---

185 See id. at 907; 1992 Merger Guidelines, supra note 116, § 1.21, reprinted in 4 Trade Reg. Rep. (CCH) ¶ 13,104.
186 Elders Grain, Inc., 868 F.2d at 907.
187 Id. at 906.
188 Id. at 906–07.
189 Id. at 907.
192 Eastman Kodak Co., 63 F.3d at 106.
193 Id.
194 Id. at 106–07.
195 See id.
4. Other Factors Contributing to Geographic Market Definition

In addition to the aforementioned three approaches, which focus on actual sales patterns, elasticity of supply and demand, and pricing power, courts have identified several other factors that they reason contribute to the definition of the geographic market. Among these additional factors are the nature of the product or service, governmental barriers to trade (such as tariffs, quotas, or differing governmental regulatory practices), transportation costs, parallel price movements between two areas, and industry standards. Courts often consider these factors when defining a geographic market, regardless of the overall approach they use—a practice that can lead to opposite conclusions regarding the proper geographic definition, even within the same approach, depending on the weight courts accord to the various factors.

In 1966, in *United States v. Grinnell Corp.*, the Supreme Court held that industry operating standards can be used to arrive at a national geographic market definition. The majority reasoned that in providing burglary and fire protection services, the overall business structure and plan of the defendant corporations was coordinated at a national level. Pointing to the recorded national schedule of prices, rates, and terms, the majority concluded that to reflect properly the national scope in which the defendants conducted business, a national geographic market must be adopted. The dissent, however, argued that the business structure of the defendants was not as essential as the nature of the services they provided. In protecting clients

---

197 *Id.*; see also *United States v. Marine Bancorporation, Inc.*, 418 U.S. 602, 628 (1974) (noting that entrance into the banking industry depends on governmental authorization); *Eastman Kodak Co.*, 63 F.3d at 108 (concluding a worldwide market existed because of parallel pricing movements between U.S. and foreign markets); Hornsby Oil Co. v. Champion Spark Plug Co., 714 F.2d 1384, 1394 (5th Cir. 1983) (looking to transportation costs, delivery limitations, and customer convenience to define the relevant geographic market); *United States v. Hammermill Paper Co.*, 429 F. Supp. 1271, 1278 (W.D. Pa. 1977) (stating that the fact that there was “no separate delivered pricing zone” established a national market).
198 Compare *Grinnell Corp.*, 384 U.S. at 575–76 (concluding a national market existed because of the defendants’ nationwide planning and contracting systems), *with id.* at 587–90 (Fortas, J., dissenting) (arguing a local market existed because of the localized nature of the services provided by defendants).
199 *Id.* at 575–76.
200 *Id.*
201 *Id.*
202 *Id.* at 587–90 (Fortas, J., dissenting).
from fire and burglary, the defendants maintained local central service stations that received notification of alarms and in turn notified local police or fire departments. Because each protected house was stationary and the fire and burglary protection ultimately had to be provided at each home’s individual location, the dissent reasoned that the proper geographic market should be limited to these discrete local areas. Furthermore, buyers contracted directly with the local providers on the basis of local conditions, and each service center only covered a radius of twenty-five miles. Calling the business aspects “incidental” to the services the defendants provided, the dissent contended that the majority simply had ignored the economic realities underlying the market at hand.

C. The Effect of Internet Sales in Geographic Market Definition

Despite a lack of consensus about which economic theory to adopt as the reasoning behind the application of *Tampa Electric* and the identification of other factors outside of those offered by the three scholarly analyses, the one commonality regarding geographic market definition has been adherence to the need to delineate the physical location of the market. This practice may be problematic when courts are faced with cases involving the Internet and e-commerce. Through the World Wide Web, a company can establish a truly global market for any product. Unsurprisingly, when courts have addressed instances of Internet trade, they admittedly have encountered much difficulty in defining the geographic market. Without defined boundaries or a physical place, the Internet allows millions of people, from throughout the world, to reach each other.

---

203 *Grinnell Corp.*, 384 U.S. at 588.
204 See *id.* at 589–90 (Fortas, J., dissenting).
205 *Id.* at 588.
206 *Id.* at 589.
207 *Tampa Elec.*, 365 U.S. at 327; see *Eastman Kodak Co.*, 63 F.3d at 103 (defining a global geographic market because producers all over the world held significant market presence); *Morgenstern*, 29 F.3d at 1296–97 (defining the geographic market as the state of Nebraska); *Rockford Mem’l Corp.*, 898 F.2d at 1284–85 (defining the geographic market as Winnebago County, Illinois); *Elders Grain, Inc.*, 868 F.2d at 906–07 (defining the geographic market as the belt of states running from Indiana on the east side of the Mississippi River to Kansas and Nebraska on the west side of the Mississippi River).
210 See *GreatDeals.Net*, 49 F. Supp. 2d at 858.
211 See *CyberPromotions, Inc.*, 948 F. Supp. at 459.
From a commercial market perspective, the Internet allows local suppliers to reach beyond their physical localities, expanding their “geographic” markets.\textsuperscript{212} The obstacle for courts, however, lies in measuring properly the impact of Internet sales—which can occur throughout the world—without diluting the market share of firms by automatically assuming the firm competes within a global geographic market.\textsuperscript{213} Given that more and more suppliers are turning to the Internet for the very purpose of reaching a broader market of buyers, as seen in the case of reimportation of prescription drugs, the courts soon will have to determine how to incorporate Internet sales into a geographic market analysis.\textsuperscript{214}

III. Application of the Traditional Approaches of Geographic Market Definition to Antitrust Litigation Involving the Pharmaceutical Industry

Drug manufacturers may be subject to antitrust claims, based on section 1 of the Sherman Act, if sufficient evidence exists of coordination among them to limit drug supplies to the Canadian market, in an overall attempt to prevent U.S. consumers from purchasing cheaper drugs.\textsuperscript{215} Because section 1 of the Sherman Anti-Trust Act prohibits \textit{unreasonable} vertical or horizontal price agreements, a court analyzing such a claim would have to find the restraint unreasonable in order to

\textsuperscript{212} Eblen, \textit{supra} note 89, at 80–81.
\textsuperscript{213} See ABA Section of Antitrust Law, \textit{supra} note 14, at 494 (stating that the starting point of any antitrust analysis is determining the market share of the firm, which can be done only after the overall market has been defined).
\textsuperscript{214} Eblen, \textit{supra} note 89, at 80–81.
\textsuperscript{215} See Sherman Anti-Trust Act § 1, 15 U.S.C.A. § 1 (West 2005); United States v. Trans-Mo. Freight Ass’n, 166 U.S. 290, 312 (1897) (interpreting section 1 of the Sherman Act to prohibit horizontal agreements among competitors that affect price or price-related features). It is beyond the scope of this Note to analyze whether there is sufficient evidence to conclude if drug manufacturers have committed an antitrust violation. Rather, this Note presupposes that there is sufficient evidence for a court to engage in a market analysis. The applicability and focus of this Note is limited to the sole issue of geographic market definition. Once the geographic market is defined, a court would continue the market analysis by defining the product market and then calculating the market share held by the defendants. \textit{See} Wilk v. Am. Med. Ass’n, 895 F.2d 352, 359–60 (7th Cir. 1990) (stating that “\textit{whether market power exists in an appropriately defined market is a fact-bound question” and continuing to define geographic and product markets to determine the defendants had a fifty-percent market share). The market analysis then would allow the court to gauge the anticompetitive effect of the disputed action. \textit{Id.} (agreeing with the district court’s findings that the defendant’s boycott had anticompetitive effects).
impose legal liability under the Rule of Reason.\textsuperscript{216} Under this analysis, the plaintiffs first must prove that such restrictions on supply have anticompetitive effects.\textsuperscript{217} Then, the defendants would have the opportunity to prove that the restraint in question offers offsetting procompetitive effects.\textsuperscript{218} Should the defendants be successful in presenting such evidence, the burden would shift back to the plaintiffs to demonstrate that the restraint was not reasonably necessary to achieve the purported benefits.\textsuperscript{219}

In this Rule of Reason analysis, one of the first issues the plaintiff would face would be defining the relevant market, which includes the relevant geographic market.\textsuperscript{220} This Note argues that for the drug manufacturers to have violated the Sherman Act in their attempt to prevent drug reimportation, Canada must be included as part of the relevant geographic market definition analyzed in an antitrust claim.\textsuperscript{221} Excluding Canada from the geographic market would separate the drug companies’ practices into two distinct markets, precluding an antitrust violation because the disputed action would occur in the Canadian market and would therefore have no anticompetitive effect on the U.S. market and would remain unreachable by U.S. oversight—even though the drug manufacturers’ actions affect U.S. consumers.\textsuperscript{222} In other words, the level of anticompetitive effects due to the targeted conduct can only be measured if the threshold issue of geographic market is defined to include Canada.\textsuperscript{223}

This Part applies the various theoretical approaches of geographic market definition to the drug manufacturing industry, in particular, to the practice of drug reimportation.\textsuperscript{224} After illustrating how the judicial test, articulated in \textit{Tampa Electric Co. v. Nashville Coal Co.},

\textsuperscript{216} See United States v. Topco Assocs., 405 U.S. 596, 606 (1972) (referring to Congress’s intent that practices that “in some insignificant degree” restrain competition would not be held to violate antitrust laws).


\textsuperscript{219} See Sherman Anti-Trust Act § 1; \textit{Cont’l T.V., Inc.}, 433 U.S. at 49 (holding that the vertical restraint between a manufacturer and retailer of television sets is illegal, as analyzed under the Rule of Reason based on section 1 of Sherman Act); see also Nat’l Soc’y of Prof’l Eng’rs, 435 U.S. at 692–96 (applying first the Rule of Reason to find anticompetitive effects and then continuing to analyze the defendant’s purported justifications).


\textsuperscript{221} See id.

\textsuperscript{222} See id. (inferring that if conduct occurs in a separate market, then it would be outside the area to which “the purchaser can practicably turn” for alternatives).

\textsuperscript{223} See supra notes 73, 81–90 and accompanying text.

\textsuperscript{224} See infra notes 228–75 and accompanying text.
would support inclusion of Canada as part of the geographic market, this Part details how the other theoretical approaches—the shipments approach, the diversion approach, and the 1992 Merger Guidelines approach—also support this conclusion.\textsuperscript{225} Finally, this Part addresses some of the additional factors that must be accounted for when defining the relevant geographic market for the unique situation of drug reimportation.\textsuperscript{226} This Part argues that these additional factors are essential to any court’s geographic market analysis and that any analysis must account for these unique factors.\textsuperscript{227}

A. Pharmaceutical Drug Reimportation Under the Tampa Electric Standard for Geographic Market Definition

In 1961, in *Tampa Electric*, the U.S. Supreme Court held that the relevant geographic market is the area to which consumers can practically turn for alternatives if suppliers increased prices.\textsuperscript{228} Applying the *Tampa Electric* standard to an antitrust claim against drug manufacturers reveals that Canada should be included as part of the geographic market.\textsuperscript{229} Current industry practices and evidence of arbitrage are strong indicators that any geographic market should recognize Canada for purposes of applying U.S. antitrust laws to this practice.\textsuperscript{230} The first prong of the *Tampa Electric* test—the area in which the seller operates—supports the inclusion of Canada based on the conduct of the major drug companies.\textsuperscript{231} The most compelling evidence for this argument is the fact that companies such as Merck, GlaxoSmithKline, Roche, and Pfizer maintain manufacturing plants throughout the world to support their global distribution systems.\textsuperscript{232}

\textsuperscript{225} See infra notes 239–75 and accompanying text.
\textsuperscript{226} See infra notes 276–328 and accompanying text.
\textsuperscript{227} See infra notes 276–328 and accompanying text.
\textsuperscript{228} See 365 U.S. at 327.
\textsuperscript{229} See id.
\textsuperscript{230} See Elzinga & Hogarty, supra note 115, at 73; Landes & Posner, supra note 115, at 942, 964. Compare United States v. Grinnell Corp., 384 U.S. 574, 575 (1966) (recognizing that the security industry planned and maintained its production process and management on a national level), with Careers with Pfizer Global Manufacturing (PGM), supra note 34 (showing that Pfizer maintains a global perspective when manufacturing its products, including those sold in the United States), and Pierce, supra note 34 (recognizing that the pharmaceutical industry maintains a national perspective in its business planning and production process).
\textsuperscript{231} See 365 U.S. at 327.
\textsuperscript{232} See, e.g., Pierce, supra note 34 (quoting GlaxoSmithKline representatives as stating they have a “global” manufacturing and packaging outfit); Careers with Pfizer Global Manufacturing (PGM), supra note 34 (describing its global manufacturing plants).
In fact, by its very definition, reimportation is limited to only those drugs that are first manufactured in the United States, shipped to the Canadian market in accord with the manufacturer’s plans, and then brought back into the United States by sales from Canadian pharmacies. Thus, Canada is clearly a primary area in which drug manufacturers conduct business.

The second prong of the Tampa Electric test—the area in which the U.S. consumer (as the purchaser) can practicably turn for supply alternatives—narrows the geographic market from a potentially global market to include only Canada and the United States. An annual $695 million commercial trade, with consumer pressure for even greater growth, clearly indicates that U.S. consumers are turning to Canadian suppliers for their prescription drugs. Regarding the ease of obtaining prescription drugs from Canada, the presence of online pharmacies has expanded the portion of consumers who can practicably turn to Canadian suppliers, a key qualification in the Tampa Electric test. A supply option that previously was limited to only those living near the Canadian border or with enough expendable resources to travel to Canada has now become a reality for millions of Americans with access to a computer and a mailing service.

B. Academic Approaches to Geographic Market Definition and Reimportation of Prescription Drugs

Beyond the general Tampa Electric test, this section applies the three theoretical approaches of geographic market definition to the issue of the reimportation of prescription drugs. Like a court employing the standard Tampa Electric approach, Courts using either the shipments and diversion approaches also would conclude that Canada should be included as part of any proper geographic market analysis. The geographic market definition under the 1992 Merger Guidelines,

---

233 HHS Task Force on Drug Imp., supra note 4, at 3.
234 See supra notes 228–33 and accompanying text.
235 See Tampa Elec., 365 U.S. at 327; HHS Task Force on Drug Imp., supra note 4, at 11–12.
236 HHS Task Force on Drug Imp., supra note 4, at 11–12.
237 See Tampa Elec., 365 U.S. at 327 (defining the second prong of the test as “the market area . . . to which the purchaser can practicably turn for supplies”); Eblen, supra note 89, at 80–81 (arguing that the Internet has allowed for greater contact between consumers and sellers regardless of the distance of separation).
238 See Eblen, supra note 89, at 80–81.
239 See infra notes 243–75 and accompanying text.
240 See infra notes 243–64 and accompanying text.
however, is less definitive.\textsuperscript{241} After this section analyzes drug reimportation under the three scholarly approaches, Part III.C focuses more closely on the unique aspects of reimportation, such as the heavy presence of Internet sales and the existence of governmental barriers, and ultimately concludes that the diversion approach is best suited for defining the geographic market in which drug manufacturers compete.\textsuperscript{242}

1. Prescription Drug Reimportation Under the Shipments Approach

The shipments approach emphasizes the quantity shipped by the suppliers to various destinations as a proxy indicator of the market share held by the supplier in each destination market.\textsuperscript{243} For purposes of the shipments approach, the pharmaceutical industry’s international business operation highlights the fact that any geographic market analysis related to this industry cannot be restricted by traditional geographic demarcations because that is simply not how the industry operates.\textsuperscript{244} In fact, one of the industry’s primary arguments against legalizing reimportation has been that it would reduce drug manufacturers’ U.S. profits and thereby limit the expenditures invested into the research and development of new pharmaceutical drugs.\textsuperscript{245} Industry representatives Pharmaceutical Research and Manufacturers of America released a report in 2003 that cited the findings of several economic studies, which indicated that lower prices in the United States due to reimportation from Canada would lead to a 36.1\% to 47.5\% reduction in overall research and development intensity.\textsuperscript{246} Thus, there is a close tie and almost an interdependence between the U.S. and Canadian prescription drug supplies.\textsuperscript{247}

Not only does this aspect of the pharmaceutical industry’s research and development funding support a broad geographic definition, but the industry’s widely dispersed manufacturing system also supports a global geographic market definition.\textsuperscript{248} Drugs currently supplied in the

\textsuperscript{241} See infra notes 265–75 and accompanying text.
\textsuperscript{242} See infra notes 276–328 and accompanying text.
\textsuperscript{243} See Elzinga & Hogarty, supra note 115, at 73–76.
\textsuperscript{244} See HHS Task Force on Drug Imp., supra note 4, at 3 (describing how reimported drugs are actually manufactured in the United States and then shipped to Canada by the pharmaceutical companies).
\textsuperscript{245} See id.
\textsuperscript{246} See Pharm. Research & Mfrs. of Am., supra note 36, at 15.
\textsuperscript{247} See supra notes 32–36 and accompanying text.
\textsuperscript{248} See HHS Task Force on Drug Imp., supra note 4, at 37 (discussing the current means through which foreign medicines enter the United States).
United States frequently are manufactured in plants located in foreign countries and shipped to the United States for sale. In the specific instance of reimportation, the drugs at issue are manufactured in U.S. plants and then shipped to Canada for retail sales. Therefore, the industry’s own production process considers Canada and the United States as part of the same geographic market.

The coordination and business structure of the pharmaceutical industry thus mirrors that of the burglary and fire protection industry, which was analyzed by the U.S. Supreme Court in an antitrust claim in 1966. In United States v. Grinnell Corp., the Court defined a national market based on the burglary and fire protection industry’s national schedule of prices, rates, and terms. Similarly, under the shipments approach here, courts should define the geographic market to include Canada based on the pharmaceutical industry’s international operation of manufacturing plants, shipping patterns, and allocation of research and development funds.

2. Prescription Drug Reimportation Under the Diversion Approach

To apply the diversion approach, one must determine the elasticity of supply and demand in the pharmaceutical industry. Although calculation of each specific numeric elasticity is beyond the scope of this Note, the evidence of arbitrage in the pharmaceutical industry lends credence to the proposition that domestic demand is fairly elastic. Internet sales from Canadian pharmacies have created a truly global market, in which consumers are able to divert their consumption from domestic markets to cheaper alternatives from foreign markets. At the same time, retailers in the pharmaceutical industry also exhibit a

---

249 Id.
250 See id. at 3.
251 See supra notes 29–39 and accompanying text.
252 Compare Grinnell Corp., 384 U.S. at 575 (recognizing that the security industry planned and maintained its production process and management on a national level), with Careers with Pfizer Global Manufacturing (PGM), supra note 34 (showing that Pfizer coordinates among its global facilities to manufacture its products), and Pierce, supra note 34 (recognizing that the pharmaceutical industry maintains a national perspective in its business planning and production process).
253 See Grinnell Corp., 384 U.S. at 575–76.
254 Id. at 575.
255 See id.
257 See id. at 942.
258 See Eblen, supra note 89, at 80–81.
relatively high elasticity of supply.\textsuperscript{259} This phenomenon is manifested by the fact that many Canadian pharmacies are increasing their output to satisfy the demands of U.S. consumers as U.S. consumers switch from the higher-priced domestic product to the cheaper Canadian product.\textsuperscript{260} Thus, the Canadian pharmacies’ elastic supply of prescription drugs can absorb the excess U.S. consumer demand created as a result of the higher prices of domestic prescription drugs.\textsuperscript{261}

Consequently, the prescription drug market exhibits characteristics similar to those of the photographic film market in \textit{United States v. Eastman Kodak Co.} in 1995.\textsuperscript{262} In that case, the Second Circuit Court of Appeals utilized the diversion approach to define a global geographic market for the photographic film industry.\textsuperscript{263} Just as in the pharmaceutical industry with relation to Canadian suppliers, the photographic film industry at the time of the court’s decision consisted of international competitors who could easily absorb excess consumer demand as a result of a price increase by Kodak, the defendant firm.\textsuperscript{264}

3. Drug Reimportation Under the 1992 Merger Guidelines

The final major theory of geographic market definition, the 1992 Merger Guidelines, might prove to be the least useful in reference to the pharmaceutical industry, though it still lends some support to including Canada as part of the U.S. market.\textsuperscript{265} On the one hand, the 1992 Merger Guidelines suggest the inclusion of Canada within the geographic market because empirical evidence seems to show that as domestic prices increase, a greater number of U.S. consumers turn to

\begin{footnotes}
\item[259] See Landes & Posner, \textit{supra} note 115, at 945.
\item[260] See \textit{id.} at 947.
\item[261] United States \textit{v.} Eastman Kodak Co., 63 F.3d 95, 104–05 (2d Cir. 1995) (holding that the ability of foreign competitors to absorb the excess demand led to their inclusion as part of the relevant geographic market definition.) By including Kodak’s competitors, the court recognized a global geographic market because Kodak’s major competitors were located throughout the world, coming from Japan, Europe, and the United States. \textit{Id.}
\item[262] \textit{Id.} (applying the \textit{Tampa Electric} standard to define a worldwide market because Kodak sold on a worldwide basis, its competitors were foreign companies, and its purchasers turned to foreign competitors’ film products as alternatives to Kodak’s products).
\item[263] \textit{Id.} at 104.
\item[264] \textit{Id.}
\item[265] See 1992 MERGER GUIDELINES, \textit{supra} note 116, § 1.22, \textit{reprinted in} 4 Trade Reg. Rep. (CCH) ¶ 13,104 (incorporating the ability of the firm to set prices higher than perfect competition levels, while at the same time refusing to recognize price discrimination within the geographic market).
\end{footnotes}
cheaper alternatives, including reimportation from Canada.\(^\text{266}\) In 1988, in *Federal Trade Commission v. Elders Grain, Inc.*, the Seventh Circuit Court of Appeals concluded that a national geographic market existed in the industrial dry corn industry based on evidence that consumers were willing to purchase from any supplier, regardless of their location in the United States.\(^\text{267}\) Similarly, in the pharmaceutical industry, a court could draw on the evidence that U.S. consumers are uniformly willing to purchase their prescription drugs from Canadian suppliers, regardless of their proximity, or lack thereof, to Canada.\(^\text{268}\)

Although the above analysis would support the inclusion of Canada within the geographic market, the 1992 Merger Guidelines suggests the exclusion of Canada from any geographic market because the Guidelines do not recognize geographic price discrimination.\(^\text{269}\) The Second Circuit Court of Appeals rejected the government’s argument for geographic market definition in *Eastman Kodak*.\(^\text{270}\) In doing so, the government relied on the 1992 Merger Guidelines to suggest that because of the different prices charged by Kodak in the United States and in foreign markets, the geographic market should be limited to the United States.\(^\text{271}\) Without evidence that Kodak incurred equal costs in the domestic and foreign markets, the court refused to identify the U.S. market as a separate geographic market for purposes of the antitrust analysis.\(^\text{272}\) Following the court’s analysis in *Eastman Kodak*, one could argue that the price differentials between the Canadian and U.S. markets is simply the result of differences in costs experienced by the drug manufacturing industry.\(^\text{273}\) For example, the approval process for new drugs, differences in governmental regulations, and per capita income differentials between the U.S. and

\(^{266}\) See id.

\(^{267}\) FTC v. Elders Grain, Inc., 868 F.2d 901, 906–07 (7th Cir. 1989).

\(^{268}\) See HHS Task Force on Drug Imp., supra note 4, at 11, fig.1.1. A report from the U.S. Department of Health and Human Services found 355 points of entry into the United States for reimported drugs, including mail, courier, and port services. *Id.* Each state, except for South Dakota, had at least one point of entry through which prescription drugs could reach U.S. citizens from Canada. *Id.*

\(^{269}\) See 1992 Merger Guidelines, supra note 116, § 1.22, reprinted in 4 Trade Reg. Rep. (CCH) ¶ 13,104 (arguing that the presence of geographic price discrimination supports the identification of multiple geographic markets, as opposed to a single all-inclusive geographic market).

\(^{270}\) See 63 F.3d at 103, 105.

\(^{271}\) See id. at 106; 1992 Merger Guidelines, supra note 116, § 1.22, reprinted in 4 Trade Reg. Rep. (CCH) ¶ 13,104.

\(^{272}\) See Eastman Kodak, 63 F.3d at 106.

\(^{273}\) See id. at 103–06.
Canadian markets all potentially contribute to the industry’s differential pricing.274 Because the application of the 1992 Merger Guidelines generates these contradictory outcomes, this approach seems to offer the least insight of the scholarly approaches into defining the relevant geographic market.275

C. Unique Factors of Prescription Drug Reimportation and Its Geographic Market Definition

Besides offering another opportunity for courts to assess the merits of the various approaches to geographic market definition, the factual circumstances of the reimportation issue provide a unique opportunity for the courts to address the difficult and novel issues related to geographic market definition: those arising from the sale of prescription drugs on the Internet and the unpredictable status of governmental barriers to reimportation.276

1. Effect of Internet Sales

Unlike many practices of other industries that are challenged as being potentially anticompetitive, the practice of drug reimportation is fueled by the Internet—meaning the Internet plays a larger role in geographic market definition here than it would for most other industries.277 The entire practice of reimportation is a practicable alternative to most U.S. consumers only because the Internet allows distant consumers (those located within the United States) easy access to distant suppliers (those located within Canada).278 In other words, the presence of the Internet has effectively integrated two otherwise separate and distinct geographic markets for prescription drugs—that of the United States and that of Canada—into a single geographic market.279

2. Effect of Governmental Barriers

In addition to the effect of Internet sales, governmental barriers to trade also must be accounted for in any legal analysis regarding the

274 See Baker, supra note 61, at 2–3.
275 See id.
277 See supra notes 61–65 and accompanying text.
278 See supra notes 61–65 and accompanying text.
279 See supra notes 207–14 and accompanying text.
pharmaceutical industry.\textsuperscript{280} Currently, Canada allows its pharmacies to fill patient prescriptions without requiring the patient first to have visited a Canadian doctor.\textsuperscript{281} It is because of this nonregulation that Canadian Internet pharmacies are able to operate.\textsuperscript{282} There may be reason to believe, however, that the Canadian government will soon erect barriers to the operation of Internet pharmacies in an effort to limit reimportation.\textsuperscript{283} Canadian officials have recently begun discussing requiring Canadian physicians to see U.S. patients in person before prescribing their medication.\textsuperscript{284} Such a change would make it practically impossible for many Americans to receive medications from Canada via the Internet.\textsuperscript{285} Currently lacking such regulations, however, the Canadian government poses no governmental barriers to reimportation.\textsuperscript{286}

The greater governmental challenge to reimportation comes from the U.S. government, which technically maintains that commercial reimportation is illegal.\textsuperscript{287} There are signs, however, that this governmental barrier may not be as absolute as it initially appeared.\textsuperscript{288} The 2003 Medicare Act has in name legalized reimportation; all that remains to effectuate this change is the approval of the Secretary of Health and Human Services.\textsuperscript{289} In addition, the FDA, the federal agency in charge of enforcing the U.S. laws against reimportation, has not been enforcing these laws.\textsuperscript{290} Furthermore, increasing public pressure for more affordable prescription drugs has motivated many public representatives to advocate change.\textsuperscript{291} In fact, many state and

\begin{flushleft}
\textsuperscript{280} See infra notes 281–328 and accompanying text.  \\
\textsuperscript{281} See Finkelstein, supra note 62, at 6.  \\
\textsuperscript{282} See id. (describing how Canadian doctors currently copy and then issue prescriptions written by U.S. doctors without requiring an in-person consultation).  \\
\textsuperscript{283} See id. at 5.  \\
\textsuperscript{284} Id.  \\
\textsuperscript{285} See Tampa Elec., 365 U.S. at 327 (including practicable consumer alternatives as part of the calculus for determining the relevant geographic market).  \\
\textsuperscript{286} See Finkelstein, supra note 62, at 5.  \\
\textsuperscript{287} See HHS Task Force on Drug Imp., supra note 4, at 4–5.  \\
\textsuperscript{288} See id. at 1–5.  \\
\textsuperscript{289} Id.  \\
\textsuperscript{290} Id. at 20.  \\
\textsuperscript{291} See Pharmaceutical Reimportation from Canada, supra note 65 (reporting to the House Government Reform Subcommittee on Wellness and Human Rights that prescription drug costs should be lowered and that U.S. Rep. Bernie Sanders had introduced legislation that would prohibit pharmaceutical companies from retaliating against Canadian pharmacies selling to Americans); McLaughlin & Davidson, supra note 59 (urging union members to pressure their senators to take action on drug reimportation legislation); Reimportation Update, supra note 59 (quoting Sen. Grassley as stating, “‘American consumers are demanding lower prices on prescription drugs . . . .’”).
\end{flushleft}
local governments have already implemented programs to facilitate the purchase of drugs from Internet sources.\footnote{See Pharmaceutical Reimportation from Canada, supra note 65; McLaughlin & Davidson, supra note 59; Reimportation Update, supra note 59.} Finally, and perhaps most indicative of the need to address the situation, a variety of legal challenges raised under state antitrust laws are forcing this issue before the courts for resolution.\footnote{See In re Pharm. Indus. Average Wholesale Price Litig., 307 F. Supp. 2d 196, 213–15 (D. Mass. 2004).}

Under current understanding, governmental barriers, on either the Canadian or the U.S. side, would suggest the exclusion of Canadian sales from any geographic market analysis for purposes of U.S. antitrust laws.\footnote{See id.} The traditional understanding rests upon the assumption that such barriers usually indicate separate geographic markets because it is assumed that suppliers would cater their business to the unique regulatory demands of each market.\footnote{See United States v. Marine Bancorporation, Inc., 418 U.S. 602, 628 (1974) (stating that entrance into the banking industry depends on governmental authorization); ABA Section of Antitrust Law, supra note 14, at 537–38.} The practice of reimportation of pharmaceutical drugs, however, explicitly recognizes that the suppliers’ products are the same, regardless of whether the supplier is located in the United States or in Canada.\footnote{See HHS Task Force on Drug Imp., supra note 4, at 3.} In other words, to reimport drugs to the United States, the drugs are first produced in the United States (along with drugs that are sold in the United States), exported to Canada, and then sold by Canadian pharmacies back to U.S. consumers.\footnote{Cf. Marine Bancorporation, 418 U.S. at 628 (citing the fact that entrance into the banking industry depends on governmental authorization as a factor in delineating the relevant geographic market area).} Thus, the traditional assumption that suppliers are catering to different markets due to governmental regulations is inapplicable in the context of drug reimportation.\footnote{See supra notes 41–51, 280–99 and accompanying text.} Rather, in the case of drug reimportation, drug manufacturers produce the prescription drugs that are sold to both the U.S. and Canadian markets in the same factories located in the United States.\footnote{See HHS Task Force on Drug Imp., supra note 4, at 3.} This practice suggests that governmental regulations are not barriers to the industry practice that views both Canada and the United States as a single North American market.\footnote{See supra notes 41–51, 280–99 and accompanying text.}
An additional governmental barrier that is likewise not to be viewed as an absolute divide is related to the wholesale pharmaceutical pricing schemes between the United States and Canada.\footnote{See supra notes 41–58 and accompanying text.} The Canadian government, through the PMPRB, imposes a cap on the wholesale prices pharmaceutical drug manufacturers can charge.\footnote{See supra notes 42–46 and accompanying text.} By contrast, the U.S. government has established only a self-reporting system that allows drug manufacturers to set wholesale prices.\footnote{See supra notes 47–58 and accompanying text.} Although the Canadian system might seem to be the more anticompetitive model because it injects the heavy hand of government into the free market, the U.S. model, upon closer inspection, does not operate on a free market basis either.\footnote{For examples of the U.S. government’s regulation of free markets in the United States, see, e.g., Sherman Anti-Trust Act § 1, 15 U.S.C.A. § 1 (West 2005); Sherman Anti-Trust Act § 2; Clayton Act § 13(a), 15 U.S.C. § 23 (2000); Clayton Act § 14; Clayton Act § 18; Federal Trade Commission Act § 5, 15 U.S.C. § 45 (2000).}

For example, in 2003, in In re Pharmaceutical Industry Average Wholesale Price Litigation, a Federal District Court in Massachusetts was presented with evidence revealing monopolistic pricing tendencies in the United States by the major pharmaceutical companies.\footnote{Complaint ¶¶ 699–703, In re Pharm. Indus. Average Wholesale Price Litig., 263 F. Supp. 2d 172 (D. Mass. 2003).} The plaintiffs’ complaint alleged that the exorbitant “spread” between the actual wholesale price and the charged wholesale price violated federal antitrust laws.\footnote{See id. ¶¶ 187, 208, 280, 311, 466, 501, 534.} For multiple prescription drugs, the DOJ calculated that the AWP (the drug manufacturer’s reported “wholesale” price) exceeded the actual wholesale price by hundreds to thousands of percents.\footnote{See id.} Such evidence suggests that the pharmaceutical manufacturers are actually operating as a monopolistic cartel, setting prices higher than would be expected in a purely competitive free market.\footnote{See id.}

Regardless of whether the U.S. pricing mechanism is an antitrust violation, the pricing scheme does support the idea that the pharmaceutical industry views the Canadian and U.S. markets as a single market.\footnote{See Careers with Pfizer Global Manufacturing (PGM), supra note 34 (describing its global manufacturing plants).} As already discussed, the majority of global research and development undertaken by pharmaceutical companies is funded by
revenues generated in the United States.\footnote{310} Thus, the regulatory nature of the Canadian market does not operate as a wall between Canada and the United States; rather, it highlights the fact that the pharmaceutical companies augment their Canadian sales with profits earned by having a high price spread in the U.S. markets.\footnote{311}

Although the current regulatory scheme does not prohibit sales between Canada and the United States, potential new legislation could dramatically reduce the commerce between the two countries in the future.\footnote{312}

3. The Best Methodology: The Diversion Approach

Given the deeply political nature of this issue, the diversion approach is the best analytical framework to address the antitrust issues raised by drug manufacturers’ attempts to end the practice of reimportation of prescription drugs.\footnote{313} The shipments approach’s focus on the current status of the market means it could misinterpret the actual market if governmental regulations eventually limit transactions between Canadian pharmacies and U.S. drug manufacturers.\footnote{314} Unlike the shipments approach, which includes only current shipment figures in its analysis of the geographic market, the diversion approach is able to measure both the current and future status of any market.\footnote{315} Thus, even if legislation eventually limits the trade between Canadian pharmacies and U.S. drug manufacturers, the diversion approach still will include Canada as part of a single market based on the idea that the U.S. drug companies have entered the Canadian wholesale market through their willingness to continue sales to Canadian wholesales who do not sell to U.S. customers.\footnote{316} The shipments approach, in contrast, would simply conclude that two markets exist because significant units of product no longer would be shipped to

\footnote{310} See id.  
\footnote{311} See supra notes 31–33, 57, and accompanying text.  
\footnote{312} See Finkelstein, supra note 62, at 5.  
\footnote{313} See supra notes 151–72 and accompanying text.  
\footnote{314} See supra notes 117–50 and accompanying text.  
\footnote{315} See supra notes 151, 162–67, and accompanying text.  
\footnote{316} See supra notes 151, 162–67, and accompanying text (explaining that under the diversion approach, if a firm has entered local market with even a single sale, it has done so after calculating the associated costs and thus, that is sufficient evidence to support the claim that if the local market were to increase in profitability, the firm would then divert additional units of product to that local market).
Canada. Thus, the diversion approach is able to remain flexible in its measurements of the elasticity of supply.

In addition to its ability to deal with future governmental barriers, the diversion approach remains flexible in measuring the elasticity of demand. Because U.S. pharmaceutical companies often own monopolies in the form of patents over their drugs, the demand for their products is relatively inelastic. Thus, in keeping with general economic thinking, the U.S. pharmaceutical companies, as rational profit-maximizing entities, will reduce the quantity of goods sold to Canadian pharmacies. This is because sales by Canadian pharmacies to U.S. consumers cut into the profits drug manufacturers gain when U.S. pharmacies sell the same product to the same U.S. consumer, but at a higher, unregulated price. By focusing on the elasticity of demand, the diversion approach is able to capture the desire of Canadian pharmacies to purchase prescription drugs from U.S. pharmaceutical companies, even if the pharmaceutical companies refuse to sell (and therefore ship) any product.

In addition, the diversion approach is best able to account for demand from U.S. consumers. By incorporating all Canadian suppliers that supply even a single unit of prescription drugs to a U.S. consumer, the diversion approach incorporates all sales of such suppliers and thus more accurately assesses the cross-border flow of prescription drug sales. In this fashion, the diversion approach also properly accounts for the large impact of online sales from Canadian pharmacies to U.S. consumers. By contrast, the shipments approach, which accounts for only those suppliers that currently ship a majority of their products to the defined market, would preclude inclusion of most Canadian pharmacies because their main consumers are not U.S. customers—a rationale that ignores the dynamic reality of prescription drug reimportation.

317 See supra notes 117–24 and accompanying text.
318 See supra notes 161–67 and accompanying text.
319 See supra notes 155–61 and accompanying text.
320 See Landes & Posner, supra note 115, at 942.
321 See id.
322 See id.
323 See supra notes 156, 162–67, and accompanying text.
324 See supra notes 155–61 and accompanying text.
325 See supra notes 162–67 and accompanying text.
326 See supra notes 61–65 and accompanying text.
327 See supra notes 120–24 and accompanying text.
Because it maintains flexibility in accounting for both supply and demand forces, the diversion approach is best capable of dealing with the unique role of the Internet, a global market, and governmental barriers in the reimportation of prescription drugs between Canada and the United States.\textsuperscript{328}

\textbf{Conclusion}

Although prescription drug reimportation, in the end, might not be a long-term sustainable option for U.S. consumers aiming to reduce the cost of their prescriptions, the current practice has caused enough concern among the pharmaceutical industry and consumers that both sides have brought their concerns before U.S. courts. As such, numerous questions regarding the current practice of geographic market definition exist when antitrust claims are raised with respect to reimportation. In addition to the ambiguity surrounding the existing methodologies for geographic market definition, the advancement of modern technologies—especially the Internet—is changing how society views its “geography” and thus altering the legal sense of market share based on a physical geographic definition. Rooted in this unique setting, prescription drug reimportation provides an opportunity for courts to recognize the flexibility that the diversion approach offers to geographic market definition. Unlike other approaches, the diversion approach accurately captures the relevant market forces while accommodating any potential governmental barriers subsequently imposed by governments. In short, although the issue of drug reimportation raises some unique considerations under antitrust law, the traditional antitrust methodology and analysis still can be adapted to ensure drug manufacturers’ actions to limit drug reimportation are subjected to antitrust oversight. In turn, such oversight safeguards free competition to the benefit of all U.S. consumers.

\textsc{Maryan M. Chirayath}

\textsuperscript{328} See \textit{supra} notes 155–71 and accompanying text.
Abstract: The current trend of local parish suppressions raises fundamental questions regarding ownership of gifted property. Disputes regarding ownership arise when a donor’s intent conflicts with the institution’s hierarchy claiming ownership of the gifted property. Hierarchical religious institutions like the Roman Catholic Church can assert such claims under two legal principles. First, because the First Amendment requires the separation of church and state, civil courts cannot interpret religious law. Second, because hierarchical religious institutions are often organized as corporations sole, local parishes are agents of the archdiocese. This Note argues for extending neutral principles of law currently used to settle intra-church property disputes to property disputes between religious institutions and individuals. In applying neutral principles of law, including the law of wills and trusts, courts can successfully adjudicate property disputes to ensure a testator’s intent is effectuated.

Introduction

The suppression, or closing, of local parishes as part of a hierarchical church system presents complex problems for the disposition of church property. Because hierarchical churches, like the Roman Catholic Church, are often organized as corporations sole, property that belongs to the local parish may also belong to the higher church body. Parishioners who set aside assets for the exclusive use by their local parish in wills or charitable trusts may be surprised to find out that upon the suppression of their local parish, the archdiocese often has title to the property.

In these situations, a fundamental tension exists between the disposition of church property and the actual intent of the donor. For the most part, these issues are not justiciable by civil courts because the

---

2 See Bassett, supra note 1, § 1:17.
First Amendment requires the separation of church and state.\(^5\) Depending upon the ecclesiastical framework of hierarchical churches, however, it is sometimes possible for civil courts to sever secular property disputes from religious doctrine and successfully apply neutral principles of law to dispose of the property.\(^6\)

This Note argues for the extension of the neutral principle approach from the context of intra-church disputes to disputes arising between religious institutions and individuals, and suggests how donors can protect their gifted property from transferring to the archdiocese. Part I of this Note provides a synopsis of the concepts and practical implications of hierarchical religious institutions organized as corporations sole.\(^7\) Part II of this Note explains the First Amendment hurdle that plaintiffs must clear in order to adjudicate this kind of property dispute.\(^8\) Part III of this Note explores the concepts of severability, neutral principles of law, and specific examples of such secular principles.\(^9\) Part IV of this Note endorses the extension of the neutral principle approach to property disputes between churches and individual donors and suggests how best to express one’s intent so as to prevent an archdiocese from becoming an unintended beneficiary.\(^10\)

I. The Corporation Sole and Religious Organizations

The corporate structure known as the corporation sole often shields religious institutions from civil court interference in property disputes.\(^{11}\) A corporation is an artificial person existing as an entity distinct from that of its members.\(^{12}\) The corporation is a legal entity created via state law designed to shield its individual members from liability while also providing a legally responsible voice for all who deal through and with the corporation.\(^{13}\) A corporation can hold title to property, sue, and be sued in its own name.\(^{14}\) By creating a separate corporate structure for religious organizations, individual members can avoid personal liability.\(^{15}\)

\(^5\) U.S. Const. amend. I.

\(^6\) See Jones v. Wolf, 443 U.S. 595, 603 (1979); Bassett, supra note 1, § 3:6.

\(^7\) See infra notes 11–40 and accompanying text.

\(^8\) See infra notes 41–88 and accompanying text.

\(^9\) See infra notes 89–170 and accompanying text.

\(^10\) See infra notes 171–223 and accompanying text.

\(^11\) See Bassett, supra note 1, § 1:17.

\(^12\) See id. § 3:1.

\(^13\) See id.

\(^14\) See id.

\(^15\) See id.
Religious organizations can create two types of corporations: aggregate and sole.\textsuperscript{16} Corporations aggregate are characterized by multiple people united in form—people who occupy the same office at different times—who are part of a perpetual succession of members that can continue indefinitely.\textsuperscript{17} In contrast, corporations sole consist only of one person and that person’s successors who are incorporated by law to give them certain legal capacities and advantages, particularly perpetuity.\textsuperscript{18} Examples of corporations sole include kings, bishops, parsons, and vicars.\textsuperscript{19} Through the corporation sole, the present incumbent and subsequent predecessors are theoretically one and the same, because anything given to one is also given to the others.\textsuperscript{20}

The modern corporation sole emerged from English common law.\textsuperscript{21} At that time, the crown, bishopric and vicarage, by operation of law, were considered continuing and independent entities that held and administered property through a vertical succession of office holders.\textsuperscript{22} Vertical succession allowed office holders to hold property for the duration of their time in office and to hold it for the benefit of and until replaced by their successor.\textsuperscript{23} Today about one-third of the Roman Catholic diocesan bishops in the United States are corporations sole.\textsuperscript{24} Episcopal dioceses, various Orthodox dioceses, and the bishop of the Church of Jesus Christ of Latter-Day Saints are also organized as corporations sole.\textsuperscript{25}

Corporations sole are creations of state law and must be established by charter or religious incorporation acts.\textsuperscript{26} Legislative action is required because corporations sole cannot arise by operation of law because the First Amendment, as applied to the states by incorporation through the Fourteenth Amendment, requires church disestab-
lishment. Seventeen states have statutes that explicitly recognize the corporation sole. At least eight other jurisdictions have utilized a special charter to create one or more corporations sole.

Exceptions to the general rule—that corporations sole no longer arise by implication of law and must be created by state law—exist in (1) Florida, where the state supreme court held that the common law corporation sole is still valid; (2) Arkansas, where the state supreme court held that without a special legislative act, the Roman Catholic Bishop of Little Rock is recognized as a corporation sole; and (3) Delaware, where the Roman Catholic Diocese of Wilmington is a one-person corporation under the state’s General Corporation Law instead of a corporation sole incorporated under the Delaware Code for Religious Societies and Corporations.

A corporation sole provides a simple and efficient method of centralized control. Moreover, and more importantly, disputes regarding the use and ownership of property as well as civil accountability for church assets are usually nonjusticiable issues for civil courts. Thus, a corporation sole protects churches from use and ownership disputes because these disputes often cannot be resolved by civil courts. The ability to remove church property disputes from a civil court’s jurisdiction is the major reason why many states do not allow the creation of corporations sole and prefer to provide for a trustee or membership corporation for religious organizations.

---

27 See U.S. Const. amends. I, XIV; Wright v. Morgan, 191 U.S. 55, 59 (1903). In 1903, in Wright v. Morgan, the U.S. Supreme Court noted that the law does not recognize the bishop as a corporation sole unless incorporated as such by statute. Id. Disestablishment of religion is the separation of church and state which prevents religious institutions from establishing their beliefs as law and from receiving financial support from the state. See Bassett, supra note 1, § 2:1.


29 See Bassett, supra note 1, § 1:17. Those jurisdictions include the District of Columbia, Illinois, Kentucky, Maine, Maryland, Massachusetts, Nebraska, and Rhode Island. See id.


31 See Bassett, supra note 1, § 1:17.

32 See id.

33 See id.

34 See id.
poration sole, church members must rely almost exclusively upon internal canons and church bylaws for accountability and recourse regarding resolution of property disputes with the church.\textsuperscript{35} Relying on religious documents containing these canons and bylaws often requires interpretation of canon law, something a civil court cannot do, thus removing the majority of church disputes from civil courts.\textsuperscript{36}

In addition to the distinction between corporations sole and aggregate, churches can also be incorporated according to different structural frameworks.\textsuperscript{37} Congregational churches are self-governed and are described as independent entities without obligation to a higher authority.\textsuperscript{38} Unlike churches that are congregational in structure and based on the will of the majority, hierarchical churches place the final decision-making authority in the ecclesiastical body that is superior to the local congregation.\textsuperscript{39} Local parishes are subordinate to the higher church unit and to the church tribunal, which retains ultimate authority and decision-making capabilities.\textsuperscript{40} It is this hierarchical structure—in which a higher church unit may retain decision-making authority over the local church unit—that is the source of tension when a parish is suppressed and ownership of church property is called into question.

II. ConstitutionAl Approaches to Resolving Property Disputes

In addition to relying on corporate structures, religious institutions can also insulate or protect their assets by creating and documenting the appropriate “church polity.”\textsuperscript{41} Church polity describes the inner workings and organizational framework of religious entities, their methods of governance, and the manner in which churches implement their doctrines and religious commitments.\textsuperscript{42} Church polity determines the relationship between local churches and the national denomination.\textsuperscript{43} Because church polity is usually ecclesiastical in nature, it is of-

\begin{itemize}
\item \textsuperscript{35} See id.
\item \textsuperscript{36} See Jones v. Wolf, 443 U.S. 595, 604 (1979); Bassett, supra note 1, § 1:17.
\item \textsuperscript{37} See Bassett, supra note 1, § 3.3.
\item \textsuperscript{38} See id.
\item \textsuperscript{39} See Watson v. Jones, 80 U.S. 679, 682 (1872); Bassett, supra note 1, § 3:3.
\item \textsuperscript{40} See Watson, 80 U.S. at 682; Bassett, supra note 1, § 3:3. This Note uses the terms “local parish” and “higher church unit” in a broad sense as opposed to specifically naming the organization of a single religion because this subject matter is applicable to numerous hierarchical religious institutions.
\item \textsuperscript{41} See Jones v. Wolf, 443 U.S. 595, 603–04 (1979).
\item \textsuperscript{42} See Bassett, supra note 1, § 3:2.
\item \textsuperscript{43} See id.
\end{itemize}
ten not subject to judicial determination. For this reason judicial determination of an ecclesiastical dispute would violate the First Amendment.

Religious protections provided by the First Amendment, as incorporated through the Fourteenth Amendment to be applicable to the states, present the initial obstacle that plaintiffs must overcome when seeking to adjudicate church property disputes. The First Amendment not only protects individuals from the federal government’s intrusion into religious doctrines, but also serves to protect the liberty of churches and religious organizations. The First Amendment forbids Congress from enacting any law respecting the establishment of religion and from prohibiting the free exercise of religion.

The First Amendment’s Establishment Clause guarantees individuals that the federal government will not use its resources to impose or advance religion. Additionally, the Free Exercise Clause guarantees that the federal government will not interfere with an individual’s right to pursue religious beliefs of their choosing. These prohibitions are designed to limit the federal government’s influence over an individual’s religious beliefs and its oversight of religious institutions. The U.S. Supreme Court, however, distinguishes between the absolute freedom of religious beliefs and the limited freedom to act upon those beliefs. Religious organizations cannot use the First Amendment to shield themselves from civil liability when disputed conduct falls outside the scope of religious beliefs and doctrine.

The activities of religious corporations are governed by their own internal laws. Because the religious corporations are creations of the state, canon law is necessarily and indirectly enforced in civil courts to the extent that such issues overlap with non-religious issues. Leading Supreme Court decisions hold that when issues arise regarding inter-

---

44 See Jones, 443 U.S. at 602; Bassett, supra note 1, § 3:2.
45 See Jones, 443 U.S. at 602.
46 See id.
47 See U.S. Const. amends. I, XIV; Jones, 443 U.S. at 603–04.
49 See U.S. Const. amend. I, cl. 1; Cantwell, 310 U.S. at 303; Bassett, supra note 1, § 7:20.
50 See U.S. Const. amend. I, cl. 1; Cantwell, 310 U.S. at 303; Bassett, supra note 1, § 7:20.
51 See Cantwell, 310 U.S. at 303; Bassett, supra note 1, § 2:1.
52 See Cantwell, 310 U.S. at 303.
53 See Jones, 443 U.S. at 605; Bassett, supra note 1, § 7:20.
54 See Jones, 443 U.S. at 603; Bassett, supra note 1, § 3:6.
55 See Jones, 443 U.S. at 602; Bassett, supra note 1, § 3:6.
nal canon law, organizational discipline of churches, or resolution of property disputes within churches, civil courts must defer to the judgments of the religious corporation’s own canon laws and regulations.\textsuperscript{56} When ecclesiastical controls are absent or do not clearly resolve the matter in dispute, however, secular courts are empowered to intervene with the application of neutral principles of civil law.\textsuperscript{57}

Neutral principles of law allow courts to apply secular legal theories, such as property or contract law, to disputes involving religious organizations without violating the First Amendment.\textsuperscript{58} Civil courts are an appropriate forum for resolving church property disputes when neutral principles of law can be applied.\textsuperscript{59} The reason for this is that when ecclesiastical issues are not at stake, neither the Establishment Clause nor the Free Exercise Clause of the First Amendment is violated by government resolution of these disputes.\textsuperscript{60} Because no constitutional barrier is present for state action in these situations, states may resolve disputes over church property through state courts by using neutral principles of law and need not adopt rules of compulsory deference to religious authorities in resolving such disputes.\textsuperscript{61}

Thus, there is a clear distinction between disputes involving the internal structure and organization of a church—over which civil courts have no jurisdiction—and disputes involving the disposition of property that belongs to a religious institution—over which civil courts do have jurisdiction because resolution is achieved through secular laws of property, trust, and contract.\textsuperscript{62}

In 1872, \textit{Watson v. Jones} was the U.S. Supreme Court’s first case regarding a dispute over church property.\textsuperscript{63} Although the schism that triggered the dispute was a result of opposing views on slavery and began as an internal theological debate over morals, the involved church factions each claimed entitlement to use of the church.\textsuperscript{64} The

\begin{footnotes}
\item[56] See \textit{Jones}, 443 U.S. at 603–04; \textit{Watson v. Jones}, 80 U.S. 679, 729–31, 733 (1872). \textit{Watson} and \textit{Jones} created the key concepts of judicial deference to ecclesiastical authority within hierarchical religious structures and severability of secular legal issues from religion without violating the First Amendment so that civil courts can adjudicate church property disputes. See \textit{Jones}, 443 U.S. at 603–04; \textit{Watson}, 80 U.S. at 729–31, 733; \textit{Bassett, supra} note 1, § 3:6.
\item[57] See \textit{Watson}, 80 U.S. at 602; \textit{Bassett, supra} note 1, § 3:6.
\item[58] See \textit{Jones}, 80 U.S. at 603.
\item[59] See id. at 602; \textit{Bassett, supra} note 1, § 7:18.
\item[60] See \textit{Jones}, 80 U.S. at 602; \textit{Bassett, supra} note 1, § 7:18.
\item[61] See \textit{Jones}, 80 U.S. at 604–05.
\item[62] See \textit{Jones}, 80 U.S. at 602–03; \textit{Bassett, supra} note 1, § 7:18.
\item[63] See \textit{Watson}, 80 U.S. at 705.
\item[64] See id. at 692.
\end{footnotes}
Supreme Court ultimately decided that it could not determine which of the two factions actually represented the church and therefore controlled the property.65

The reason for this decision was twofold.66 First, the Court explained that those who belong to a church implicitly consent to the church’s superior organization’s ability to make their own determinations without interference from civil courts.67 Second, the Court held that this was not a case where the disputed property was devised by a specific individual who wanted to retain control, but rather the property was purchased by the church for the congregation, so long as the congregation remained part of the church.68 The Court’s decision to allow the church tribunal to determine the rightful congregation represents the historical practice of compulsory deference to religious institutions.69 Compulsory deference was advocated because of the seeming inability to adjudicate religious institutions’ property issues without violating constitutional protections.70

In 1979, however, the Supreme Court established the current standard of civil justiciability in Jones v. Wolf, in which it declared civil courts competent in adjudicating any issue capable of analysis under neutral principles of law so long as it does not require a judgment regarding matters of theology or religious belief.71 Like Watson, Jones also involved a dispute over church property resulting from a schism in a local church that was part of a hierarchical church organization.72 In Jones, more than half of the congregation voted to separate from the larger church government and unite with another denomination.73 After the superior church body formed a commission to investigate the schism within the local church, it issued a ruling that the minority faction—which opposed the break—was the true congregation.74 The conflict over property ownership surfaced when the superior church body withdrew authority from the majority faction.75 Ultimately, the current standard for resolving controversies about the

65 Id. at 727.
66 Id. at 726, 729.
67 Id. at 729.
68 Watson, 80 U.S. at 726.
69 Id. at 727.
70 Id.
71 See Jones, 443 U.S. at 603–05; Bassett, supra note 1, § 7:22.
72 Jones, 443 U.S. at 598.
73 Id.
74 Id.
75 See id. at 598–99.
ownership of church assets simply requires a determination of who is the legal owner of the property in question.\textsuperscript{76}

This current standard replaced the previous standard of compulsory deference to church bodies as seen in \textit{Watson}.\textsuperscript{77} The earlier practice of compulsory deference was based on the notion that because a religious institution was involved in a property dispute, the dispute was inherently religious in nature and therefore nonjusticiable by civil courts.\textsuperscript{78} The modern standard represents the revised understanding that church property disputes can be settled with neutral principles of law.\textsuperscript{79}

In 2004, in \textit{Akoury v. Roman Catholic Archbishop of Boston}, a Massachusetts Superior Court contributed to the concept of neutral principles of law.\textsuperscript{80} The \textit{Akoury} decision specifies that the application of neutral principles of law will be most applicable in those cases where plaintiffs can show a strong property interest.\textsuperscript{81} The court concluded that the plaintiffs’ challenge regarding ownership of church property was a pretext for a challenge against the Archbishop’s decision to suppress a local parish.\textsuperscript{82} The court determined that the decision to suppress a parish is inherently an ecclesiastical decision and therefore nonjusticiable by civil courts.\textsuperscript{83} The court reasoned that when the plaintiff has a strong property interest, the court will more likely address the underlying property dispute and less likely claim the dispute involves nonjusticiable issues.\textsuperscript{84} Thus, \textit{Akoury} adds to the foundation of neutral principles in two ways.\textsuperscript{85} First, it highlights that the way an argument is framed is key to whether a court will adjudicate the issue.\textsuperscript{86} For example, in this case the court determined that the plaintiff’s argument sounded less like a property dispute and more like a challenge against parish suppression.\textsuperscript{87} Second, it indicates that the extent of the plain-

\textsuperscript{76} See id. at 610; \textit{Bassett, supra} note 1, \S 7:22.

\textsuperscript{77} See \textit{Jones}, 443 U.S. at 605; \textit{Watson}, 80 U.S. at 727.

\textsuperscript{78} See \textit{Jones}, 443 U.S. at 604; \textit{Watson}, 80 U.S. at 706.

\textsuperscript{79} See \textit{Jones}, 443 U.S. at 603.


\textsuperscript{81} See id. at 272. In \textit{Akoury}, the plaintiffs were claiming title to the church grounds, the church, furniture, religious items, and $200,000 held in the parish accounts donated by the parish members. \textit{Id.} at 271.

\textsuperscript{82} See id. at 272.

\textsuperscript{83} See id.

\textsuperscript{84} See id.

\textsuperscript{85} See \textit{Akoury}, 18 Mass. L. Rptr. at 272.

\textsuperscript{86} See id. at 271–72.

\textsuperscript{87} See id. at 272.
tiff’s property interest also affects the court’s decision of whether to adjudicate the issue—the stronger the plaintiff’s claim to the property, the more likely the court will solve the dispute.\textsuperscript{88}

III. \textsc{Secular Principles Applied to Property Disputes}

A. \textit{Deference, Neutral Principles of Law, and Severability}

When it comes to the resolution of church disputes, civil courts have developed two basic approaches.\textsuperscript{89} The first approach requires the court to adhere to the notion of judicial deference whereby it protects and implements the decision reached by the church-designated ecclesiastical tribunal.\textsuperscript{90} The second approach involves the court adjudicating the given property dispute by applying a secular, that is, religiously neutral, analysis.\textsuperscript{91}

Courts followed the compulsory deference approach as spelled out by the Supreme Court in 1872 in \textit{Watson v. Jones} for over one hundred years.\textsuperscript{92} At issue in \textit{Watson} was whether the governing body of the Presbyterian Church had the power to prescribe qualifications for local church offices and determine which faction of a local church was entitled to the church property in dispute.\textsuperscript{93} The Supreme Court in \textit{Watson} held that the governing body of the church had such power, thereby creating the doctrine of judicial deference to the internal decision-making body of a church.\textsuperscript{94}

The Supreme Court indicated that the rule of judicial deference prevented the entanglement of church and state by providing religious institutions with the power to decide questions of religious belief, church discipline, and ecclesiastical government without state interference.\textsuperscript{95} Additionally, whenever questions of faith and custom had been determined by the ecclesiastical tribunal, civil courts had to accept the decisions as final and binding upon them.\textsuperscript{96} The rule of deference requires courts to defer to the internal decisions of church

\textsuperscript{88} See \textit{id.}
\textsuperscript{89} See \textit{Jones v. Wolf}, 443 U.S. 595, 604–05 (1979); \textsc{Bassett, supra} note 1, § 7:23.
\textsuperscript{90} See \textit{Jones}, 443 U.S. at 604; \textsc{Bassett, supra} note 1, § 7:23.
\textsuperscript{91} See \textit{Jones}, 443 U.S. at 604; \textsc{Bassett, supra} note 1, § 7:23.
\textsuperscript{92} See \textit{Watson v. Jones}, 80 U.S. 679, 706 (1872); \textsc{Bassett, supra} note 1, § 7:23.
\textsuperscript{93} See \textit{Watson}, 80 U.S. at 717; \textsc{Bassett, supra} note 1, § 7:23.
\textsuperscript{94} See \textit{Watson}, 80 U.S. at 706; \textsc{Bassett, supra} note 1, § 7:23.
\textsuperscript{95} See \textit{Watson}, 80 U.S. at 733; \textsc{Bassett, supra} note 1, § 7:23.
\textsuperscript{96} See \textit{Watson}, 80 U.S. at 733; \textsc{Bassett, supra} note 1, § 7:23.
authorities when such decisions are exclusively linked to the fundamental beliefs and interpretation of church law.\footnote{97}{See Jones, 443 U.S. at 604; Bassett, supra note 1, § 7:23.}

This deferential approach by courts changed, however, when, in 1979, the Supreme Court adopted the neutral principles of law approach in \textit{Jones v. Wolf}.\footnote{98}{See Jones, 443 U.S. at 603–05; Bassett, supra note 1, § 7:32.} In \textit{Jones}, the Court held that states may adopt such neutral principles in adjudicating the resolution of church property disputes without violating constitutional safeguards.\footnote{99}{See Jones, 443 U.S. at 602; Bassett, supra note 1, § 7:32.} The Court indicated, however, that in order for states to comply with the First Amendment’s religious clauses, civil courts must defer to ecclesiastical tribunals—if they exist—concerning matters of religious doctrine and polity.\footnote{100}{See U.S. Const. amend. I; Jones, 443 U.S. at 602; Bassett, supra note 1, § 7:32 (citing Peter M. Shannon, \textit{Deference or Neutral Principles: The Dual Approach by Civil Courts to Ecclesiastical Disputes}, 49 Proceedings CLSA 106 (1987)).}

Courts need not, however, follow a rule of compulsory deference to ecclesiastical authorities in resolving church property disputes where there is an absence of religious doctrine or polity concerning the issue in dispute.\footnote{101}{See Jones, 443 U.S. at 602; Bassett, supra note 1, § 7:32.} Where no doctrinal controversy is involved, civil courts are entitled to follow neutral principles of law to interpret relevant provisions of a religious organization’s governing documents—such as deeds, church constitutions, bylaws, and contracts—under current state law.\footnote{102}{See Jones, 443 U.S. at 603; Bassett, supra note 1, § 7:32 (citing Peter M. Shannon, \textit{Deference or Neutral Principles: The Dual Approach by Civil Courts to Ecclesiastical Disputes}, 49 Proceedings CLSA 106 (1987)).}

The neutral principles of law approach benefits those challenging a church because it is completely secular in operation and elastic enough to be applied to various religious organizations.\footnote{103}{See Jones, 443 U.S. at 603; Bassett, supra note 1, § 7:32.} Another advantage of this approach is the ease with which courts can determine whether or not it can be applied.\footnote{104}{See Jones, 443 U.S. at 603; Bassett, supra note 1, § 7:32.} For example, when the interpretation of deeds, charters, bylaws, or church constitutions would require a civil court to resolve disputes regarding religious doctrine, it is easily determined that neutral principles of law cannot be applied to interpret the case in accordance with the First Amendment.\footnote{105}{See U.S. Const. amend. I; Jones, 443 U.S. at 602; Bassett, supra note 1, § 7:32.}
such an event, the court must still defer to authoritative ecclesiastical bodies.\textsuperscript{106} Because the neutral principles of law approach relies exclusively upon objective, well-established concepts of property, trust, and contract law, subject matter familiar to lawyers and judges, the \textit{Jones} Court concluded that civil courts, in adhering to these norms, could avoid entanglement in religious doctrine, polity and practice.\textsuperscript{107}

As part of the concept of neutral principles of law, \textit{Jones} also discusses severability.\textsuperscript{108} The concept of severability relates to a civil court’s ability to address church property disputes through the lens of secular trust and property law rather than defer to ecclesiastical tribunals.\textsuperscript{109} So long as civil courts rely on objective and well-established concepts of law, which do not require judicial interpretation of religious doctrine, civil courts may identify the proper legal owner of church assets without running afoul of the First Amendment.\textsuperscript{110} After \textit{Jones}, civil courts have broad discretion to resolve church property disputes by relying on secular legal principles from other areas of the law.\textsuperscript{111}

Interestingly, the Court in \textit{Jones} explicitly outlined ways in which churches can avoid having church disputes settled in civil courts.\textsuperscript{112} The Court suggested redrafting church constitutions, charters, and other documents to make clear, in religious terms, who controls the property, thereby leaving civil courts no choice but to uphold church designations.\textsuperscript{113} The Supreme Court indicated that through explicit provisions, churches could define what happens to the disposition of property for any number of given doctrinal controversies.\textsuperscript{114} Additionally, the Supreme Court clarified that a civil court would have to defer to a hierarchical church’s tribunal’s jurisdiction if the church had property distribution policies in place and the tribunal was operating with procedural regularity.\textsuperscript{115}

For a church to shield successfully its decisions from civil adjudication, it must have documents that clearly present the procedures by which church decisions are made and express religious justifications

\begin{footnotes}
\textsuperscript{106} See \textit{Jones}, 443 U.S. at 604.
\textsuperscript{107} See \textit{id.} at 603.
\textsuperscript{108} See \textit{id.}; \textit{Bassett, supra} note 1, § 7:32.
\textsuperscript{109} See \textit{Jones}, 443 U.S. at 603; \textit{Bassett, supra} note 1, § 7:32.
\textsuperscript{110} See \textit{Jones}, 443 U.S. at 603; \textit{Bassett, supra} note 1, § 7:32.
\textsuperscript{111} See \textit{Jones}, 443 U.S. at 603; \textit{Bassett, supra} note 1, § 7:32.
\textsuperscript{112} See \textit{Jones}, 443 U.S. at 603; \textit{Bassett, supra} note 1, § 7:32.
\textsuperscript{113} See \textit{Jones}, 443 U.S. at 604; \textit{Bassett, supra} note 1, § 7:32.
\textsuperscript{114} See \textit{Jones}, 443 U.S. at 603.
\textsuperscript{115} See \textit{id.}; \textit{Bassett, supra} note 1, § 7:32.
\end{footnotes}
for the sources of authority or the purposes of particular decisions.\textsuperscript{116} In order to secure this constitutional safeguard, however, the basic documents must demonstrate a clear self-understanding as portrayed by the religious body.\textsuperscript{117} For the church to demonstrate its religious self-understanding, the church must show that its religious values are so entwined with the institution’s governance that civil adjudication is precluded.\textsuperscript{118} In order for the church to take advantage of First Amendment freedoms, the church must convince a civil court that religious overtones pervade the entire property dispute, preventing the court from severing the property dispute and applying secular principles of law.\textsuperscript{119}

Hierarchical religious institutions (such as the Roman Catholic and Episcopal Churches) organized as corporations sole thus have the ability to shield property disputes from civil courts.\textsuperscript{120} Such entities can do so by combining religious doctrine with organizational documents that address the adjudication and resolution of such issues.\textsuperscript{121} When church property is insulated in this manner, civil courts must defer to the religious decision-making body in order to comply with the First Amendment.\textsuperscript{122} Alternatively, when a hierarchical church organized as a corporation sole has neglected to indicate in its religious and organizational documents an intent to resolve church property disputes internally, a civil court may adjudicate the issue without violating the Establishment and Free Exercise Clauses by applying neutral principles of law and severability.\textsuperscript{123}

\textbf{B. Secular Principles}

Because civil courts can adjudicate some property disputes by applying secular principles of law, such as property and trust law, it is necessary to understand how courts apply these principles to resolve specific issues.\textsuperscript{124} This Note aims to help plaintiffs prevent, retrospectively or prospectively, property gifted to their local parish from being

\textsuperscript{117} See Jones, 443 U.S. at 603–04; Gaffney & Sorensen, supra note 116, at 99.
\textsuperscript{118} See Jones, 443 U.S. at 603–04; Gaffney & Sorensen, supra note 116, at 99–100.
\textsuperscript{119} See U.S. Const. amend. I; Jones, 443 U.S. at 604.
\textsuperscript{120} See Jones, 443 U.S. at 603–04; Bassett, supra note 1, § 7:23.
\textsuperscript{121} See Jones, 443 U.S. at 603–04, 606; Bassett, supra note 1, § 7:23.
\textsuperscript{122} See U.S. Const. amend. I; Jones, 443 U.S. at 605; Bassett, supra note 1, § 7:23.
\textsuperscript{123} See U.S. Const. amend. I; Jones, 443 U.S. at 602; Bassett, supra note 1, § 7:32.
\textsuperscript{124} See Jones, 443 U.S. at 603.
transferred to the archdiocese of a hierarchical church upon suppression of the local parish. This necessitates a discussion of the doctrine of cy pres, which allows courts to interpret a will so as to carry out the testator’s intentions expressed therein.\textsuperscript{125}

When a testator leaves property in trust for charitable purposes, the trust does not necessarily fail because the testator has failed to state the particular charitable purpose to which the property is to be applied.\textsuperscript{126} The trust fails, however, when the testator manifests an intention that the property is only to be used for a particular charitable purpose, yet fails to manifest properly that specific charitable intention.\textsuperscript{127} When a testator has not manifested an intention to devote property to charitable purposes generally, but has manifested an intention to devote it to a particular charitable purpose, the intended charitable trust fails if the particular charitable purpose cannot be accomplished.\textsuperscript{128}

For example, if a testator left property in trust for his or her local parish stating that he or she intended the property to benefit anybody interested in the teachings of Catholicism generally, the trust would not fail because the testator did not indicate a specific use for the property.\textsuperscript{129} The gift survives because the testator stated a general intent to benefit religion and did not condition the gift upon a particular use only.\textsuperscript{130} This is not the case, however, where a testator specifically leaves property to benefit his or her local parish.\textsuperscript{131} In the latter case, the testator has clearly manifested a specific intent—to benefit the local parish—limiting the scope of the charitable purpose, but failing to indicate the exact condition. Because the testator did not express a general intent to promote religion, but did express a specific intent to benefit his or her local parish, the gift must fail because of a lack of specification regarding the charitable purpose.\textsuperscript{132}

\textsuperscript{127} See Scott & Fratcher, supra note 126, § 395.
\textsuperscript{128} See Evans, 396 U.S. at 441; Scott & Fratcher, supra note 126, § 395.
\textsuperscript{129} See Evans, 396 U.S. at 440. The hypotheticals are the Author’s creation.
\textsuperscript{130} See id.
\textsuperscript{131} See id. at 441.
\textsuperscript{132} See id.
1. Cy Pres

In situations where it becomes impossible or impractical to carry out the testator’s wishes, cy pres can be used to modify the testator’s intention and achieve results that most closely resemble the testator’s intentions.\(^{133}\) Cy pres is an equitable doctrine that allows courts to modify a testator’s gift to mirror what the testator would have wanted when it becomes impossible to carry out the testator’s exact provision.\(^{134}\) Although a lack of specific charitable purpose may cause a gift to fail, courts may apply the doctrine of cy pres when attempting to save a charitable trust.\(^{135}\) Thus, the central analysis in a case where cy pres is applicable concerns the testator’s intent, which is a question of fact.\(^{136}\) Courts do not apply cy pres in every case when a testator’s wishes cannot be successfully fulfilled.\(^{137}\) For instance, in some situations, the testator had only one particular purpose in mind, and no expressed general charitable intent.\(^{138}\) In such cases, it is impossible to accomplish the particular purpose, leaving courts to presume that the testator would have preferred the trust to fail entirely, causing the assets to revert to the testator’s heirs.\(^{139}\)

When the testator includes contingency provisions in case the testator’s intent cannot be carried out, those provisions are controlling.\(^{140}\) When the testator does not, however, include any contingency language, courts must determine whether the gift fails entirely or whether the doctrine of cy pres should be applied.\(^{141}\) Courts have held that cy pres is inapplicable when a testator clearly intended the property to be put to a particular use or for the benefit of a particular organization that has since dissolved. In such instances, the property should revert to the testator’s estate.\(^{142}\) If a testator leaves property in trust, however, for a particular charitable purpose that is incapable of

\(^{133}\) See id. at 440.

\(^{134}\) See Evans, 396 U.S. at 440.

\(^{135}\) See id.; Obermeyer v. Bank of Am., 140 S.W.3d 18, 23 (Mo. 2004). While cy pres was originally used to save gifts in trust, it has been extended and applied to absolute gifts to charitable corporations or other organizations. See Obermeyer, 140 S.W.3d at 23 (citing George T. Bogert, Trusts and Trustees, § 431, 105 (2d ed. 1991)); Scott & Fratcher, supra note 126, § 399.

\(^{136}\) See Obermeyer, 140 S.W.3d at 22.

\(^{137}\) See Evans, 396 U.S. at 441.

\(^{138}\) See id.

\(^{139}\) See id. at 444.

\(^{140}\) See id.; Scott & Fratcher, supra note 126, § 399.2.

\(^{141}\) See Evans, 396 U.S. at 440.

\(^{142}\) See id. at 444.
being executed, the trust will not fail if the testator demonstrated a broad, general intention to devote the property to charitable purposes. In such a case, the court will utilize cy pres to dispose of the property in line with the testator’s ascertained intention.

For example, if a testator left money in trust to a local parish which was suppressed shortly after the testator’s death, a court could determine that the testator possessed a general intention to devote the property to charitable purposes, and therefore provide the money to a successor parish. The other possible outcome is that the court may assume the testator wanted to benefit the broader archdiocese in the event of a suppression of the testator’s local parish. If the court determines, however, that the testator only wanted the money to go to the local parish, then upon its suppression, and in the absence of a more general intent to promote religion, the gift must fail and the money must revert to the testator’s estate.

2. Ambiguities and Extrinsic Evidence

Other secular tools that courts apply when analyzing church property disputes include the doctrines of judicial construction and interpretation of wills. The fundamental legal principles of judicial construction and interpretation of wills are vital when it comes to in-

---

143 See id. at 441; Scott & Fratcher, supra note 126, § 399.2.

144 See Evans, 396 U.S. at 441.

145 See Obermeyer, 140 S.W.3d at 26 (determining where money left by testator should go after the original recipient, the Dental Alumni Development Fund, ceased to exist upon the closure of Washington University Dental School).

146 See id. This hypothetical is analogous to the court’s decision in Obermeyer v. Bank of America to use cy pres and give the money to Washington University generally, despite the fact that the conditional gift failed when the dental school closed and the specific Dental Fund ceased to exist. See id. The decision to give the money to Washington University hinged upon the court’s finding that the testator possessed general, rather than specific, charitable intent despite contrary language indicating his intent to benefit the Dental Fund specifically. See id. This situation, however, is distinguishable from the one proposed in this Note regarding a testator’s gift to their local parish for two reasons. First, the deceased in Obermeyer made unrestricted donations throughout his life to Washington University generally without any conditions regarding the dental school or the Dental Fund. See id. Second, the remainder of his estate was left to Washington University generally—a clear intention to benefit more than just the dental school. See id. Thus, the court’s ability to find general intent in the face of narrowly tailored language lies in the court’s analysis of the testator’s intent as determined from a lifetime of giving to the University. See id.

147 See Evans, 396 U.S. at 444 (holding that because testator gifted and demonstrated specific intent of leaving public park to white people only, testator would rather have gift fail and revert to his heirs than be a racially integrated park).

148 See id. at 440; Obermeyer, 140 S.W.3d at 25–26; George W. Thompson, Construction and Interpretation of Wills § 83 (1928).
interpreting testamentary documents. In cases where the testator’s intent is clear and can be inferred from the will itself, but the testator has omitted certain words necessary to express fully his intention, courts will supply the necessary words by implication to bring about the result the testator intended.

For example, if a testator bequeathed money to the local parish, conditioned on the fact that the parish remain in existence, courts may infer that, should the parish be suppressed, the testator intended the gift to revert back to the testator’s heirs based on the will itself. Under these circumstances, a court will supply the missing language. Courts may not read language into a will, however, unless the court finds that the testator’s intentions are clearly inferable from the will itself, although perhaps not appropriately expressed.

When applying a provision in the testator’s will to actual circumstances, events often occur after the testator’s death that call into question the testator’s intent. In cases where the testator’s intent cannot clearly be ascertained from the content of the will, a court may consider additional facts and circumstances existing at the time of the will’s execution to ascertain the testator’s intent. The court should thus step into the shoes of the testator and consider the circumstances influencing the testator when the will was executed.

When circumstances change after a testator’s death, ambiguities often result. There are two kinds of ambiguities—patent and latent. A patent ambiguity is one that appears on the face of the will itself, preventing a court from interpreting the testator’s intent. A latent ambiguity, in contrast, arises out of a document that is comprehensible on its face, but becomes unclear when its terms are applied

---

149 See Evans, 396 U.S. at 440; Obermeyer, 140 S.W.3d at 25–26; Thompson, supra note 148, § 83.
151 This is the Author’s own hypothetical.
152 See Henricksen, 46 F. Supp. at 836; Thompson, supra note 148, § 83.
153 See Henricksen, 46 F. Supp. at 834; Thompson, supra note 148, § 83.
154 See Obermeyer, 140 S.W.3d at 21; Thompson, supra note 148, § 217.
155 See Obermeyer, 140 S.W.3d at 26; Thompson, supra note 148, § 217.
156 See Evans, 396 U.S. at 441; Thompson, supra note 148, § 217.
157 See Evans, 396 U.S. at 437; Obermeyer, 140 S.W.3d at 22; Thompson, supra note 148, § 327.
158 See Bradley v. Wash., Alexandria & Georgetown Steam-Packet Co., 38 U.S. 89, 97 (1839); Thompson, supra note 148, § 327.
159 See Bradley, 38 U.S. at 97; Thompson, supra note 148, § 327.
to specific circumstances. Most courts hold that extrinsic evidence cannot be introduced to resolve patent ambiguities. In contrast, for latent ambiguities where the will itself is free from defect, extrinsic evidence can be introduced to clarify the testator’s intent.

Examples of extrinsic evidence that the court should take into consideration when clarifying the testator’s intent include the testator’s thought habits, the testator’s relationship to, or associations with, the bounty, and the testator’s motives. Even though the testator’s intention must be found in the will, courts are able to infer the meaning of the words in the will by taking into account various circumstances that affected and motivated the testator at the time of execution. Although extrinsic evidence of surrounding circumstances cannot be used to assert an intention differing from that expressed in the will, extrinsic evidence is properly used as an aid in proving the correct understanding of the will’s language.

Interpreting and applying extrinsic evidence is particularly relevant to the successful application of cy pres. When a testator’s will indicates an exact duty to be performed, and it then becomes impossible to perform that duty, the doctrine of cy pres is used to accomplish the testator’s intent as best as possible. There is often confusion and controversy in testamentary gifts to religious or charitable organizations, and for that reason, when applying the doctrine of cy pres to religious purposes, the general rule is to apply extrinsic facts and circumstances so as to give effect to the testator’s intent. Because it is critical to accomplishing the testator’s ultimate purpose, interpreting testamentary documents requires discerning the testator’s intent—that is, what the testator would have wanted if aware that the provision could not be carried out. Thus, a court should con-

160 See Bradley, 38 U.S. at 97; Thompson, supra note 148, § 327.
161 See Bradley, 38 U.S. at 97; Thompson, supra note 148, § 327. Extrinsic evidence is information from outside the will which is used to help construe the true intent of the testator. See Jesse Dukeminier & Stanley M. Johanson, Wills, Trusts, and Estates 420 (6th ed. 2000).
162 See Bradley, 38 U.S. at 97; Thompson, supra note 148, § 327.
163 See Obermeyer, 140 S.W.3d at 26; Thompson, supra note 148, § 217.
164 See Obermeyer, 140 S.W.3d at 26; Thompson, supra note 148, § 217.
165 See Obermeyer, 140 S.W.3d at 26; Thompson, supra note 148, § 217.
166 See Exans, 396 U.S. at 440; Thompson, supra note 148, § 142.
167 See Obermeyer, 140 S.W.3d at 23; Thompson, supra note 148, § 144.
168 See Obermeyer, 140 S.W.3d at 25.
sider all the relevant surrounding circumstances, as embodied by extrinsic evidence, in determining the testator’s intent.\textsuperscript{170} 

\textbf{IV. Analysis: Resolving Church Property Disputes Through Application of Neutral Principles}

The application of neutral principles is the accepted method for determining which faction involved in a church schism is the rightful owner of church property.\textsuperscript{171} This Note argues that this approach should be extended to situations involving gifts by individuals to churches, thereby preserving testator intent and preventing an archdiocese or similar body from claiming gifted property. In deciding which faction is the true representative of a hierarchical church, a judgment often requires ecclesiastical determinations, something a civil court cannot make.\textsuperscript{172} The Supreme Court, however, has stated that where church bodies have not created their own guidelines for property dispute resolutions, states are free to do so, so long as the investigation is free from religious determinations.\textsuperscript{173} Therefore, in cases involving individuals, rather than factions, where no or inadequate church guidelines exist for property dispute resolutions, neutral principles should apply.\textsuperscript{174} This Note also offers practical advice for donors who are retroactively trying to regain control of their property and for donors who prospectively want to guarantee their gifts are used according to their specific wishes.

\textit{A. Defense of Neutral Principles and Support for Its Extension}

1. The Policy Justification for Applying Neutral Principles in Intra-Church Disputes and for Extending It

The Supreme Court has given civil courts the ability to use neutral principles of law while steering clear of religious controversy in the context of determining which church faction is entitled to the use and control of church property.\textsuperscript{175} There are inherently religious overtones in determining which church faction actually controls the church in contrast to the lack of any religious overtones in determin-

\textsuperscript{170} See id. at 26.
\textsuperscript{171} See Watson v. Jones, 80 U.S. 679, 703 (1872).
\textsuperscript{172} See Jones v. Wolf, 443 U.S. 595, 607 (1979).
\textsuperscript{173} See id. at 602.
\textsuperscript{174} See id.
\textsuperscript{175} See Watson, 80 U.S. at 725.
ing whether a particular gift made by a testator’s will should fail.\textsuperscript{176} Analysis regarding who is entitled to use and enjoy church property in an intra-church dispute relies mostly upon the analysis of relevant church documents, which are used to decide which faction actually represents the church. In contrast, a determination of whether a church should be allowed to keep a bequest relies more heavily upon the law of wills and trusts.\textsuperscript{177}

The policy reason for enforcing the neutral principles of legal analysis in intra-church disputes—that is, the state’s interest in peaceful resolution of property disputes—applies even more forcefully to settling disputes between a religious body and an individual.\textsuperscript{178} This is because of the drastic reduction of religious inquiries required in a property dispute.\textsuperscript{179} In 1979, in \textit{Jones v. Wolf}, the Supreme Court indicated that states have a legitimate interest in providing a civil forum for the peaceful resolution of church property disputes while being mindful of resolving such disputes outside the auspices of religious doctrine and practice.\textsuperscript{180} The Supreme Court thus required deference when religious institutions have already designated that property disputes are to be resolved within the ecclesiastical body, but explicitly stated that states may otherwise adopt various approaches to resolving these property disputes.\textsuperscript{181} Applied to cases where gifts have been made, where no religious inquiries are necessary, neutral principles should apply.\textsuperscript{182}

2. Applying Neutral Principles Beyond Intra-Church Disputes

Opponents of the neutral principles approach believe that it is inappropriate to inquire into the proper ownership of property gifted to a soon-to-be suppressed local parish.\textsuperscript{183} This school of thought views the decision to suppress a parish as an internal church decision, inherently religious in nature.\textsuperscript{184}

\begin{flushright}
\textsuperscript{176} See Evans v. Abney, 396 U.S. 435, 440 (1970); Watson, 80 U.S. at 706.
\textsuperscript{177} See Evans, 396 U.S. at 439; Watson, 80 U.S. at 706.
\textsuperscript{178} See Watson, 80 U.S. at 726; Akoury v. Roman Catholic Archbishop of Boston, 18 Mass. L. Rptr. 271, 272 (Mass. Super. 2004).
\textsuperscript{179} See Evans, 396 U.S. at 439; Watson, 80 U.S. at 706.
\textsuperscript{180} See Jones, 443 U.S. at 602.
\textsuperscript{181} See id.
\textsuperscript{182} See Watson, 80 U.S. at 726; Akoury, 18 Mass. L. Rptr. at 272.
\textsuperscript{183} See Akoury, 18 Mass. L. Rptr. at 272.
\textsuperscript{184} See id.
\end{flushright}
In *Akoury v. Roman Catholic Archbishop of Boston*, the plaintiffs—members of a local parish that was part of the Roman Catholic Church—attempted to enjoin the Archbishop of Boston from liquidating, transferring, and conveying to the office of the Archbishop, property allegedly held in trust for the benefit of the local parish.¹⁸⁵ In focusing on the important role that the local parish played in the plaintiffs’ lives as a place to celebrate life and death, the Massachusetts Superior Court viewed the plaintiffs’ claim as an attempt to prevent parish suppression.¹⁸⁶ Because the court viewed this suit as an attempt to halt suppression, it focused on the religious nature of the decision to close a parish.¹⁸⁷ The court concluded that closing a parish is a matter best left to the Archbishop.¹⁸⁸ The Superior Court also indicated that because the Roman Catholic Church is hierarchical in structure and a corporation sole, the parish assets were held in custody for the benefit of the Archdiocese by its agent, the pastor of the local parish.¹⁸⁹

The *Akoury* court’s holding—that the decision to close a parish is inherently religious—appears at first to frustrate a plaintiff’s attempt to regain control of conditionally gifted property.¹⁹⁰ This holding, however, is not applicable to the situations this Note considers, which involve testators who want to prevent an archdiocese from claiming their property.¹⁹¹

For one, in *Akoury*, the court framed the legal issue in terms of the plaintiffs trying to prevent parish suppression, which is an inherently religious issue that civil courts cannot hear.¹⁹² The legal issue present in the case of a plaintiff who gifts property to their local parish for its use only is not a religious question touching suppression of the congregation in the first place.¹⁹³ Rather, the legal issue is the purely secular question of who has title to the property when the beneficiary of the gift ceases to exist as was required, either expressly or implicitly, in the gift provision itself.¹⁹⁴

¹⁸⁵ See id. at 271.
¹⁸⁶ See id.
¹⁸⁷ See id. at 272.
¹⁸⁸ See *Akoury*, 18 Mass. L. Rptr. at 272.
¹⁸⁹ See id.
¹⁹⁰ See id.
¹⁹¹ See id.
¹⁹² See id.
¹⁹³ See *Akoury*, 18 Mass. L. Rptr. at 272.
¹⁹⁴ See *Jones*, 443 U.S. at 603–04.
Additionally, there is an unspoken issue that pervades the *Akoury* case which distinguishes *Akoury* from cases involving gifted property.\(^{195}\) Nowhere in *Akoury* is it indicated that the plaintiffs had any interest in the disputed property aside from their previous use of it.\(^{196}\) In other words, potential plaintiffs have a truly legitimate interest in property they owned and conditionally gifted to the local parish, whereas the *Akoury* plaintiffs merely wanted to keep the property which they had become accustomed to using.\(^{197}\) For this reason, their attempt to frame the dispute as one over property failed.\(^{198}\)

There is thus a fundamental difference between the plaintiffs in *Akoury* and those who are attempting to regain control over property gifted to a church.\(^{199}\) Unlike the plaintiffs in *Akoury* who challenged the Archdiocese’s decision to suppress their local parish and tried to retain control over property that possibly did not belong to them, plaintiffs who are trying to regain control over property that is rightfully theirs, rather than challenging suppression, do not present a legal question that is religious in nature.\(^{200}\) Rather, they present a clear case for the application of neutral principles of the law of property, wills, and trusts.\(^{201}\)

**B. Preventing the Archdiocese from Claiming Gifted Property**

Because of the corporate structure of hierarchical churches and the religious freedoms guaranteed under the First Amendment of the Constitution, plaintiffs have limited room to maneuver when it comes to preventing the archdiocese from claiming gifted property upon suppression of their local parish.\(^{202}\) There is an inherent tension that exists between what a hierarchical church can do with certain property and the legal significance of interpreting the testator’s intent in gifting that same property.\(^{203}\)

\(^{195}\) See *Akoury*, 18 Mass. L. Rptr. at 272.

\(^{196}\) See id. at 271–72.

\(^{197}\) See id.

\(^{198}\) See id. at 272.

\(^{199}\) See id.

\(^{200}\) See *Akoury*, 18 Mass. L. Rptr. at 272.

\(^{201}\) See *Jones*, 443 U.S. at 603.

\(^{202}\) See U.S. CONST. amend. I; Bassett, supra note 1, § 3:2.

\(^{203}\) Compare Bassett, supra note 1, § 3:2 (stating that hierarchical institutions have the ability under the First Amendment to insulate decisions regarding property disputes by intertwining dispute-solving mechanisms with religious polity), with Thompson, supra note 148, § 83 (stating that, when determining the proper disposition of a testator’s property, the testator’s intent is the controlling authority).
Unfortunately for plaintiffs, where the hierarchical church has clearly chosen to designate in its religious charters, incorporation documents, and bylaws the method by which all property disputes are handled by the ecclesiastical tribunal, civil courts will probably not be able to adjudicate the matter by applying neutral principles of law.\textsuperscript{204} The Supreme Court has indicated that when religious institutions so commingle their religious doctrine with dispute-solving mechanisms, it would be a violation of the First Amendment for any civil court to attempt to alter the ecclesiastical tribunal’s resolution of the dispute.\textsuperscript{205}

It appears that challenges to the disposition of gifts already made to soon-to-be suppressed local parishes will be limited to those situations where the hierarchical church has not clearly indicated in its relevant documents the manner in which property disputes are to be adjudicated.\textsuperscript{206} In these cases, religion is not so intertwined with the disposition of the disputed property so as to prevent a civil court from applying neutral principles of law.\textsuperscript{207} It is within these confines that plaintiffs have the ability to argue the legal importance of intent so as to prevent the property from going to the archdiocese.\textsuperscript{208}

Under neutral principles of law, a civil court can turn to legal concepts with which it is familiar, mainly property, trusts, and wills.\textsuperscript{209} Because the testator’s intent is the single-most important determination of disposal of gifted property, plaintiffs should focus their arguments here.\textsuperscript{210} Donors who intend to benefit their local parish or a successor parish but not the archdiocese must know what to do both prospectively and retrospectively to ensure they succeed in accomplishing their goals.

1. Prospective Protection of Donor’s Intent

Although the doctrine of cy pres can be applied to a will to achieve results similar to those that the testator intended, use of unambiguous language is preferred.\textsuperscript{211} In order to avoid a situation in which a court finds broad charitable intent such that upon the suppression of a local parish the court assumes the testator would have wanted the bequest to

\textsuperscript{204} See Watson, 80 U.S. at 733; Bassett, supra note 1, § 3:2.
\textsuperscript{205} See U.S. Const. amend. I; Watson, 80 U.S. at 733; Bassett, supra note 1, § 3:2.
\textsuperscript{206} See Jones, 443 U.S. at 603; Bassett, supra note 1, § 7:32.
\textsuperscript{207} See Jones, 443 U.S. at 603–04; Bassett, supra note 1, § 7:32.
\textsuperscript{208} See Jones, 443 U.S. at 603–04; Bassett, supra note 1, § 7:32.
\textsuperscript{209} See Jones, 443 U.S. at 603–04; Bassett, supra note 1, § 7:32.
\textsuperscript{210} See Evans, 396 U.S. at 442.
\textsuperscript{211} See id. at 440–41.
transfer to the archdiocese or similar body, it is necessary that the testator indicate where the property is and is not to go.212

Potential testators seeking to benefit simultaneously a local parish and prevent the transfer of money to the archdiocese must be explicit in any gift provision.213 If a general charitable intent is ascertained in a gift provision to a local parish such that a court can apply the doctrine of cy pres, the testator must indicate that in the event of the parish’s suppression, the testator’s intent is limited to the benefit of a successor parish, not the archdiocese or similar body.214 Testators must accordingly carefully limit the extent of their intended generosity.215 Otherwise, a finding of general intent to promote religion might lead a court to award erroneously the money to the archdiocese in the absence of the testator’s clear intention not to do so.216

2. Retrospective Protection of Donor’s Intent

When a testator leaves money to a local parish that is later suppressed, it is proper for a court to use extrinsic evidence in order to prove under the circumstances that the testator would either have: (1) wanted the money to go to a successor parish; (2) wanted the money to revert to the testator’s heirs; or (3) not have wanted the money to go to the archdiocese.217 In any of these three situations, the use of extrinsic evidence does not infuse contradictory intention into the testator’s will different from that expressed by the document, but rather clarifies dispersal of the gift in the case of a suppressed parish.218

If a gift has already been made to a soon-to-be suppressed local parish and the archdiocese claims a right to the property, the ques-

---

212 See Obermeyer v. Bank of Am., 140 S.W.3d 18, 26 (Mo. 2004).
213 Compare Evans, 396 U.S. at 441–42 (holding that the testator left such detailed instructions in his will regarding the limits of his charitable purpose that he would prefer the gift to fail than be used otherwise), with Obermeyer, 140 S.W.3d at 26 (holding that the testator’s intent was not specific enough to require the gift to fail upon the recipient’s dissolution).
214 See Evans, 396 U.S. at 441; Obermeyer, 140 S.W.3d at 26.
215 See Obermeyer, 140 S.W.3d at 26.
216 Compare Evans, 396 U.S. at 441–42 (holding that where the testator clearly manifested a specific intent that could not be fulfilled, the gift must fail), with Obermeyer, 140 S.W.3d at 26 (holding that a testator’s stated intent to gift money to a university’s dental fund was not limiting and specific enough to prevent the court from giving the money to the University as a whole).
217 See supra notes 148–70 and accompanying text. This is a hypothetical situation, the content of which is supported by the previous discussion of the proper use of extrinsic evidence.
218 See Obermeyer, 140 S.W.3d at 26.
tion becomes how to protect retrospectively the donor’s intent. If the will does not expressly state what is to happen to the property upon suppression, a court may find the presence of a latent ambiguity. A latent ambiguity is an ambiguity that arises not from a flaw in the will itself, but from changed external circumstances since the time the will was executed. In the case of latent ambiguities, courts often look to extrinsic evidence, such as the testator’s motivations, so that it can step into the testator’s shoes and modify the will accordingly.

Although extrinsic evidence of surrounding circumstances cannot be used to assert an intention different from that expressed in the will, it is properly used as an aid in proving the correct understanding of the utilized language. Plaintiffs in this position must introduce evidence relating to the testator’s motivation in making the disposition and the relationship between the testator and the beneficiary of the bounty, so as to leave no doubt that the testator’s true intent was to benefit the local parish and not to promote religion generally.

Conclusion

Religious institutions such as the Roman Catholic Church that are hierarchical in structure and organized as corporations sole are powerful entities. Because the First Amendment mandates separation of church and state, churches have the ability to insulate the decisions made by their adjudicative bodies in resolving property disputes. Churches can carefully craft their religious documents in a way that entangles church polity with ordinary property disputes, requiring civil courts to defer to ecclesiastical tribunals in compliance with the Constitution.

In situations where hierarchical churches have neglected to make religious documents part of the adjudicative process, however, or in cases where such church bodies have not shown an entwined relationship between property disputes and religious polity, civil courts can adjudicate church property disputes by applying neutral principles of law. In applying neutral principles of law, courts will refer to areas of law with which lawyers and judges are familiar. By referring to the law of property, trusts, and wills, courts analyze a testator’s gift provisions

---

219 See Bradley v. Wash., Alexandria & Georgetown Steam-Packet Co., 38 U.S. 89, 97 (1839); THOMPSON, supra note 148, § 327.
220 See Bradley, 38 U.S. at 97; THOMPSON, supra note 148, § 327.
221 See Bradley, 38 U.S. at 97; THOMPSON, supra note 148, § 217.
222 See Bradley, 38 U.S. at 97; THOMPSON, supra note 148, § 217.
223 See Bradley, 38 U.S. at 97; THOMPSON, supra note 148, § 217.
to determine the testator’s intent when a particular provision is incapable of being satisfied.

The suppression of local parishes has brought the tension between a religious institution’s autonomy and the need to satisfy a testator’s intent to the public eye. There is limited room for potential plaintiffs to maneuver when it comes to preventing the archdiocese of a hierarchical church from capturing a local parish’s assets when the hierarchical church has explicitly provided for the manner in which it will resolve property disputes. Where these provisions are lacking, however, plaintiffs seeking to control the disposition of property retrospectively must rely on the doctrine of cy pres and the admission of extrinsic evidence to clarify any latent ambiguities regarding a testator’s intended beneficiaries. Those looking to control the disposition of property prospectively must explain the limited terms upon which the gift is being made, designate who are and are not the beneficiaries of the gift, and provide for the reversion of the gift to the testator’s heirs should circumstances arise such that the testator’s intent is frustrated.

Elizabeth Ehrlich
MEDICAL-MALPRACTICE REFORM: IS ENTERPRISE LIABILITY OR NO-FAULT A BETTER REFORM?

Abstract: This Note compares two medical-malpractice reforms: enterprise liability and no-fault. The Note compares the reforms for their relative ability to compensate injured patients and deter malpractice. The Note also examines the reforms’ economic and sociopolitical feasibility. The Note concludes that a no-fault medical-malpractice system would better compensate patients and deter malpractice, but enterprise liability is a more feasible reform that policymakers should pursue more aggressively.

INTRODUCTION

In November 2000, a man was admitted to the hospital for a broken hip.\(^1\) The doctor on duty inserted a feeding tube to ensure that the patient received proper nutrition.\(^2\) The doctor accidentally inserted the tube into the patient’s lung, which began to fill with liquid.\(^3\) Despite being transferred to the hospital’s critical care unit, the patient never recovered.\(^4\) A jury found that medical malpractice caused his death.\(^5\)

In the fall of 2004, a senior partner of an OB/GYN clinic in Chicago sent her patients letters informing them she would be moving to Madison, Wisconsin where medical-malpractice insurance premiums are lower.\(^6\) Although, a lifelong resident of Chicago, who had found the practice of medicine to be challenging and rewarding, the malpractice environment in Illinois made it a hostile place for the doctor to work.\(^7\)

Medical malpractice harmed this patient and doctor, but in two very different ways.\(^8\) One story illustrates that real lives are lost or ru-

\(^1\) Lagerstrom v. Myrtle Werth Hosp.-Mayo Health Sys., 700 N.W.2d 201, 206 (Wis. 2005).
\(^2\) Id.
\(^3\) Id.
\(^4\) Id. at 207.
\(^5\) Id. at 206.
\(^6\) Amelia Buragas, Priced Out, CAPITAL TIMES (Madison, WI), Sept. 27, 2004, at 8A.
\(^7\) Id.
\(^8\) See Lagerstrom, 700 N.W.2d at 206–07; Burages, supra note 6, at 8A.
ined because of medical malpractice. The other story paints a picture of a litigious America hurting the doctor-patient relationship by driving doctors out of business. Ideally, medical malpractice law compensates injured patients and deters doctors from injuring their patients. Although medical-malpractice law has helped raise the standard of care in the medical profession, it has become a source of frustration for physicians and inadequate protection for patients.

The problems with the current medical-malpractice system are numerous. Doctors are sued much more often than in the 1950s. Medical-malpractice liability insurance is less available and less affordable, forcing some physicians to leave the profession as a result. Those doctors continuing to practice sometimes provide unnecessary treatment to protect themselves from liability. This “defensive medicine” is dangerous and costly; 44,000 to 98,000 people die each year because of medical error. There are many more injuries caused by compensable medical errors than there are malpractice claims.

9 See Lagerstrom, 700 N.W.2d at 207; see also Watkins v. Cleveland Clinic Found., 719 N.E.2d 1052, 1057–59 (Ohio Ct. App. 1998) (involving a patient in a persistent vegetative state due to surgery to fix a deviated septum in her nose).

10 See Burages, supra note 6, at 8A; see also David A. Hyman, Medical Malpractice and the Tort System: What Do We Know and What (If Anything) Should We Do About It?, 80 Tex. L. Rev. 1639, 1639 (2002); Paul C. Weiler, The Case for No-Fault Medical Liability, 52 Md. L. Rev. 908, 909–10 (1993); Patrick B. Massey, M.D., Despite Top-Notch Physicians, Medical Field Showing Signs of Illness, Chicago Daily Herald, Jan. 5, 2004, at Health & Fitness 4.


12 See Helling v. Carey, 519 P.2d 981, 985 (Wash. 1974) (making glaucoma tests standard practice for ophthalmologists); Lagerstrom, 700 N.W.2d at 206–07; Burages, supra note 6, at 8A.


14 In the late 1950s, there was approximately one malpractice claim per hundred physicians in a year. Weiler, supra note 10, at 912. By the early 1990s, there were more than ten claims per hundred physicians per year. Id.


16 See Weiler, supra note 10, at 916–17.

17 See Inst. of Med., supra note 13, at 1; Pegalis & Wachsman, supra note 11, § 2:7 (defining defensive medicine as “the alternation of modes of medical malpractice, induced by the threat of liability, for the principal purpose of forestalling the possibility of lawsuits by patients as well as providing a good and legal defense in the event such lawsuits are instituted”); Weiler, supra note 10, at 942 (estimating the cost of defensive medicine at $20 billion per year).

18 See Hyman, supra note 10, at 1643; Weiler, supra note 10, at 913.
These problems demonstrate the need for reform.\textsuperscript{19} The public should not have to bear the cost of an arguably ineffective and inefficient medical-malpractice system that inadequately distributes costs and hinders patient safety.\textsuperscript{20}

Medical-malpractice reforms have been enacted over the past thirty years, but complaints continue about the medical-malpractice system.\textsuperscript{21} The most popular reforms include damage caps, attorney contingency fees caps, installment payments for damages, screening boards, and shorter statutes of limitation for malpractice actions.\textsuperscript{22} Where enacted, the damage caps have reduced the number of suits, the amount of damages, and, therefore, the overall cost of medical malpractice.\textsuperscript{23} Unfortunately, malpractice-insurance rates remain high.\textsuperscript{24} The savings, although real, are small in comparison to the overall cost of healthcare in the United States.\textsuperscript{25} As a negative consequence of these reforms, patients with legitimate claims and severe injuries face greater difficulties when bringing medical-malpractice suits.\textsuperscript{26} Thus, despite these tort reforms, problems persist.\textsuperscript{27}

In light of the futility of popular tort reform efforts, and with an eye toward improving patient safety, some reformers have suggested

\begin{footnotesize}
\begin{enumerate}
\item<sup>19</sup> See Randall R. Bovbjerg et al., \textit{U.S. Health-Care Coverage and Costs: Historical Development and Choices for the Future}, 21 J.L. MED. & ETHICS 141, 142 (1993); Weiler, \textit{supra} note 10, at 916–17.
\item<sup>20</sup> See Barry R. Furrow, \textit{Enterprise Liability and Health-Care Reform: Managing Care and Managing Risk}, 39 St. Louis U. L.J. 79, 100–08 (1994); Weiler, \textit{supra} note 10, at 908–09.
\item<sup>23</sup> See Am. Med. Ass’n., \textit{supra} note 21, at 23; Bovbjerg, \textit{supra} note 22, at 546–53.
\item<sup>24</sup> See Republican Nat’l Comm., \textit{supra} note 15, at 59.
\item<sup>26</sup> See Ass’n of Trial Lawyers of Am., Ten Reasons to Oppose Medical Malpractice “Reform”, http://www.atlanet.org/ConsumerMediaResources/Tier3/press_room/FACTS/medmal/tenreasons0504.aspx (last visited Apr. 7, 2005).
\item<sup>27</sup> See Am. Med. Ass’n., \textit{supra} note 21, at 2–22 (describing the recurring problem of medical-malpractice liability and the reforms in states).
\end{enumerate}
\end{footnotesize}
changing the medical-malpractice system.\textsuperscript{28} One such proposed reform, enterprise liability, would shift liability entirely away from individual healthcare providers to hospitals or similar institutions.\textsuperscript{29} Even more aggressive reformers have suggested replacing the fault based negligence system with a no-fault strict liability system that would compensate injured patients even if the provider was not negligent in causing the injury.\textsuperscript{30} Acknowledging the hurdles related to enactment, both enterprise liability and a no-fault strict liability system offer the possibility of improving the affordability and quality of healthcare delivered in the United States.\textsuperscript{31}

To better understand the utility and feasibility of these systemic medical-malpractice reforms, this Note compares enterprise liability and no-fault.\textsuperscript{32} Enterprise liability retains the fault requirement of the negligence based medical-malpractice system, but limits the liability to the hospital or equivalent enterprise.\textsuperscript{33} The no-fault reform takes enterprise liability a step further by eliminating the fault requirement.\textsuperscript{34} Part I describes the history of medical-malpractice law and the development of enterprise liability and no-fault as potential reforms.\textsuperscript{35} Part II addresses the question of which is the better reform model.\textsuperscript{36} Because medical-malpractice reform is a frequent topic of public debate, the theoretical models of reform should be weighed against one another before investing limited resources in implementation.\textsuperscript{37} Part II

\textsuperscript{29} See Abraham & Weiler, supra note 28, at 398–400; John V. Jacobi & Nicole Huberfeld, Quality Control, Enterprise Liability, and Disintermediation in Managed Care, 29 J.L. MED. & ETHICS 305, 305 (2001).
\textsuperscript{31} See Abraham & Weiler, supra note 28, at 382–84; Jacobi & Huberfeld, supra note 29, at 305; Studdert & Brennan, supra note 30, at 227–29; David M. Studdert et al., Can the United States Afford a “No-Fault” System of Compensation for Medical Injury?, 60 LAW & CONTEMP. PROBS., Spring 1997, at 1, 33; Weiler, supra note 10, at 919–20.
\textsuperscript{32} See infra notes 178–386 and accompanying text.
\textsuperscript{33} Exclusivity determines whether the enterprise would be the only party a patient/plaintiff could sue. See HEALTH-CARE LAW AND ETHICS 458 (Mark A. Hall et al. eds., 6th ed. 2003). Under an exclusive enterprise liability system, a patient may only sue the hospital or other comparable enterprise for an injury. Id. Under a nonexclusive enterprise liability system, a patient may sue the hospital as well as individual providers. Id.
\textsuperscript{34} Studdert & Brennan, supra note 30, at 227–29; Weiler, supra note 10, at 919–21.
\textsuperscript{35} See infra notes 39–173 and accompanying text.
\textsuperscript{36} See infra notes 174–386 and accompanying text.
\textsuperscript{37} See Lohr, supra note 21, at BU1.
concludes that no-fault better serves the goals of compensation and deterrence, but enterprise liability is a better model for reform in the United States because enterprise liability is more economically, socially, and politically appropriate.\textsuperscript{38}

I. Background

A. History of Medical Malpractice

To understand how either no-fault or enterprise liability might improve the healthcare system, it is helpful to understand the history and evolution of the U.S.’s current negligence-based medical-malpractice system.\textsuperscript{39} In the early part of the 19th century, actions against physicians for medical malpractice were rare.\textsuperscript{40} As the century progressed, people began to demand more from their healthcare providers.\textsuperscript{41} Additionally, the medical profession became populated with doctors toting a wide range of treatments, not all of which were scientifically sound or effective.\textsuperscript{42} As a result, patients increasingly held doctors accountable for inadequate care.\textsuperscript{43} By 1850, medical-malpractice suits had become a fixture of the American healthcare system.\textsuperscript{44}

In the late 1950s, there was approximately one malpractice claim per 100 physicians in a year.\textsuperscript{45} By the early 1990s, there were more than ten claims per 100 physicians in a year.\textsuperscript{46} The increased number of malpractice claims has caused financial and emotional strain on the healthcare system.\textsuperscript{47} These strains reached their height in the malpractice “crises” of the mid-1970s and mid-1980s.\textsuperscript{48} These crises were char-

\begin{itemize}
\item \textsuperscript{38} See infra notes 376–86 and accompanying text.
\item \textsuperscript{39} See Furrow, supra note 20, at 80–100 (discussing the evolution of the health system and liability).
\item \textsuperscript{40} James C. Mohr, American Medical Malpractice Litigation in Historical Perspective, 283 JAMA 1731, 1731 (2000).
\item \textsuperscript{41} See id. at 1732.
\item \textsuperscript{42} Id.
\item \textsuperscript{43} Id.
\item \textsuperscript{44} Id. at 1731–32.
\item \textsuperscript{45} Weiler, supra note 10, at 912.
\item \textsuperscript{46} Id.
\item \textsuperscript{47} See id. at 908–10. Although the healthcare system has been strained by medical-malpractice claims, it is also important to note that medical-malpractice law has improved healthcare and patient safety. See Paul A. Sarlo, Shining a Light on Malpractice, N.Y. Times, Feb. 15, 2004, at 14NJ. Although this Note focuses on the problems of the medical-malpractice system, it does not discount the positive effect medical-malpractice law has had on healthcare. See id.
\item \textsuperscript{48} See Studdert & Brennan, supra note 30, at 225.
\end{itemize}
acterized by a surge in malpractice claims with corresponding increases in malpractice insurance rates. In 1974 and 1975, major insurers refused to continue writing coverage for medical-malpractice liability. In the 1980s, businesses of all kinds were struggling to find affordable liability insurance. These insurance crises resulted in frustration with the tort system, which prompted legislative interest in general tort reform, and in particular in medical-malpractice reform. The 1970s and 1980s reforms included insurance reforms, limiting attorney contingency fees, damage caps, and instituting pretrial screening panels.

In the 1990s, no similar insurance crisis materialized, but pressure from medical practitioners fueled continued interest in reform. More states enacted damage caps and similar federal legislation came close to passage. Despite intense debate over healthcare reform during the Clinton Administration, the Administration’s comprehensive reforms were not enacted. The Clinton plan considered moving the nation toward an enterprise liability system, but the American Medical Association (the “AMA”) strongly opposed giving hospitals more control over physicians.

Without any reforms enacted in the 1990s, medical-malpractice reform continues to be a topic of national debate. The 2004 Republican Party Platform specifically advocated non-economic damage caps as a reform, and it attacked Democrats for curtailing reform efforts and siding with trial lawyers instead of doctors and patients. During the 2004 presidential debates, President George W. Bush highlighted medical-malpractice reform as a means to alleviate healthcare costs and improve the quality of healthcare in America. Democratic presidential candidate Senator John Kerry conceded that

---

49 AM. MED. ASS’N., supra note 21, at 2–3; Studdert & Brennan, supra note 30, at 225.
50 Bovbjerg, supra note 22, at 502–03.
51 See id. at 503.
52 See id. at 503–04.
53 See id. at 513.
55 Id.
57 See Abraham & Weiler, supra note 28, at 383.
58 See Lohr, supra note 21, at BU1.
the medical-malpractice system needed reform, but he considered tort reforms, like damage caps, to be an inadequate approach to reducing healthcare costs or improving care.\textsuperscript{61} Their limited exchange exemplifies the heated debate being waged in Congress and in state legislatures throughout the country.\textsuperscript{62}

The debate is largely waged in state legislatures by two factions, one typically led by physicians, and the other by plaintiffs’ attorneys.\textsuperscript{63} The physicians and their supporters lobby for reforms that limit patients’ ability to bring suits and recover damages.\textsuperscript{64} This school of thought believes that by reducing the number of suits and the amounts plaintiffs recover, insurance rates will fall, doctors will be happier and more effective, thereby curbing healthcare costs.\textsuperscript{65} Reforms advocated by physicians include damage caps, limiting attorney contingency fees, installment payments for damages, screening boards to filter claims, and shorter statutes of limitation to reduce the time a patient would have to file a claim.\textsuperscript{66} These reforms have been selectively adopted in many states.\textsuperscript{67}

On the other side of the debate, plaintiffs’ attorneys, patients’ rights advocates, and like-minded supporters oppose these reforms and argue that such reforms unfairly limit plaintiffs’ rights and do little to improve healthcare.\textsuperscript{68} They stress that by making it more difficult for plaintiffs to bring suit, these reforms limit patients’ ability to demand quality care.\textsuperscript{69} For example, this school of thought alleges that damage caps disproportionately harm the most severely and grossly injured victims of medical malpractice.\textsuperscript{70} Furthermore, plaintiffs’ attorneys and patients’ rights advocates argue that tort reforms are futile efforts to

\textsuperscript{61} See id.

\textsuperscript{62} See Am. Med. Ass’n., supra note 21, at 24–45; Lohr, supra note 21, at BU1.


\textsuperscript{64} See Am. Med. Ass’n., supra note 21, at 23–45; Lohr, supra note 21, at BU1.

\textsuperscript{65} See Am. Med. Ass’n., supra note 21, at 23.

\textsuperscript{66} See id. at 23–24.

\textsuperscript{67} See id.


\textsuperscript{69} See Lohr, supra note 21, at BU1.

\textsuperscript{70} See Washoe County Registrar of Voters, Washoe County 2004 General Election Sample Ballot Rebuttal to Argument in Support of Question No. 3 and Argument Against Question No. 3, S6-E to S7-E (2004).
contain costs.\textsuperscript{71} They contend that limitations on patients’ rights are not justified by an imperceptible reduction in healthcare costs.\textsuperscript{72} As long as healthcare providers continue to feel unfairly burdened by the medical-malpractice system, debate over reform proposals will probably remain on the state and national agendas.\textsuperscript{73} The AMA is a powerful lobbying force, and the stories of doctors being forced out of practice by prohibitively high costs of medical-malpractice-liability insurance has potent political force.\textsuperscript{74} Voters and legislators, while sympathetic to doctors, are quick to see plaintiffs and attorneys as greedy and litigious.\textsuperscript{75} Until this changes, medical malpractice promises to remain a salient issue of public debate.\textsuperscript{76}

Any reforms that are implemented should thus improve the healthcare system.\textsuperscript{77} Some scholars argue that typical tort reform has not significantly improved the healthcare system over the past three decades and it is difficult to see how more of these same reforms will make a difference in the future.\textsuperscript{78} These scholars have proposed more

\textsuperscript{71} See Comm’n on Presidential Debates, \textit{supra} note 60 (statement of Senator John F. Kerry) (“Now, ladies and gentlemen, important to understand, the president and his friends try to make a big deal out of it. Is it a problem? Yes, it’s a problem. Do we need to fix it, particularly for OB/GYNs and for brain surgeons and others? Yes. But it’s less than 1 percent of the total cost of health care.”). Empirical evidence shows that medical malpractice only accounts for a small proportion of total healthcare costs. Weiler, \textit{supra} note 10, at 909. In 1992, the total cost of healthcare in the United States was $840 billion. \textit{Id.} A generous estimate of the cost of medical malpractice at that time was $9 billion. \textit{Id.} This figure may have underestimated the full cost of the malpractice system, however, because it does not include the cost of defensive medicine. \textit{See id.} at 909, 916, 942. Medical malpractice is blamed for causing doctors to order unnecessary tests and procedures to protect themselves from liability. \textit{See id.} at 909, 916–17. Such “defensive medicine” was estimated to cost the healthcare system another $18 billion in 1992, bringing the total cost of medical malpractice to $27 billion. \textit{See id.} Even this $27 billion figure is small relative to the total cost of $840 billion. \textit{See id.} These numbers have only continued to grow. \textit{See Bovbjerg, supra} note 15, at 45. In 2001, the cost of healthcare in the United States was 14.1\% of the country’s gross domestic product, and it has been steadily rising. \textit{See id.} In 2002, the total cost of healthcare was nearly $1.6 trillion. KAISER FAMILY FOUND., \textit{supra} note 25. Although the relatively small cost of medical malpractice tempers the urgency with which the nation needs to tackle malpractice reform, the medical-malpractice system does cause meaningful problems for healthcare providers and patients alike. W. John Thomas, \textit{The Medical Malpractice “Crisis”: A Critical Examination of a Public Debate}, 65 TEMP. L. REV. 459, 464 (1992); Weiler, \textit{supra} note 10, at 916–17.

\textsuperscript{72} See Comm’n on Presidential Debates, \textit{supra} note 60.

\textsuperscript{73} See e.g., Abraham & Weiler, \textit{supra} note 28, at 383; Gibeaut, \textit{supra} note 68, at 40; Lohr, \textit{supra} note 21, at BU1.

\textsuperscript{74} See Abraham & Weiler, \textit{supra} note 28, at 383.

\textsuperscript{75} See Weiler, \textit{supra} note 10, at 910.

\textsuperscript{76} See Lohr, \textit{supra} note 21, at BU1.

\textsuperscript{77} See INST. OF MED., \textit{supra} note 13, at 3.

novel and systemic reforms including enterprise liability and no-fault systems. The following two sections describe these reforms and their history in greater detail.

B. Enterprise Liability

Advocates first proposed enterprise liability as a systemic reform for the medical-malpractice system in the 1970s. The reform proposal evolved out of the steady development of caselaw holding hospitals liable for patient injuries. Not long ago, the idea of holding a hospital liable for the torts of its physicians was far-fetched. Early American hospitals were charitable institutions where the sick and poor would go to die. Those who could afford healthcare treatment stayed home and had doctors come to them. Courts dismissed the idea of holding a charitable hospital vicariously liable for the negligent acts of a physician. At that time, hospitals simply provided a venue for doctors to perform their duties. Hospitals exercised little or no control over doctors, leading courts to distinguish easily the hospital-doctor relationship from the typical employer-employee relationship that gives rise to vicarious liability. Additionally, these early hospitals may not have been able to keep their doors open to the sick and poor if they had been shouldered with vicarious liability. Therefore, courts established a doctrine of charitable immunity for hospitals, which clearly exempted hospitals from vicarious liability for physicians’ acts and from most direct liability for errors in treatment.

79 See Abraham & Weiler, supra note 28, at 436; David M. Studdert & Troyen A. Brennan, No-Fault Compensation for Medical Injuries, 286 JAMA 217, 222 (2001); Weiler, supra note 10, at 947–948;
80 See infra notes 81–173 and accompanying text.
81 Abraham & Weiler, supra note 28, at 384.
82 Id. at 385–94.
84 See id.
85 See id.
87 See Health-Care Law and Ethics, supra note 33, at 421–23.
89 See Schloendorff, 105 N.E. at 94; Health-Care Law and Ethics, supra note 33, at 421–23.
90 See McDonald, 120 Mass. at 434–36; Schloendorff, 105 N.E. at 94; Abraham & Weiler, supra note 28, at 385–86.
The doctrine of charitable immunity lasted until the 1950s when the law began to adjust to the changing world of healthcare. By the 1950s, hospitals were less like charitable institutions and more like sophisticated centers of healthcare delivery. Although physicians were still independent contractors of hospitals, their relationship with hospitals became increasingly like an employer-employee relationship. As a result, charitable immunity dissolved, causing hospitals to be held vicariously liable for the negligent acts of physicians when there is apparent authority or when the hospital negligently hired or supervised the physician.

Currently, hospitals can be held liable for physician negligence when physicians are held out as hospital employees. If the relationship appears sufficiently similar to that of an employer and employee, the tort doctrine respondeat superior applies, exposing a hospital to liability for a physician’s negligent acts. As hospitals have become more involved in the delivery of healthcare, including doctor-patient relationship and treatment decisions, courts have become more willing to hold them liable for treatment-related injuries, regardless of the hospital’s fault. Now, hospitals can be liable for healthcare related injuries through vicarious liability and corporate liability.

With hospitals assuming more responsibility for both the provision and liability of healthcare, reformers and academics began to consider the state of the healthcare system if hospitals were to assume all malpractice liability. Under an enterprise liability system, a hospital is liable for all malpractice regardless of whether the culpable individual healthcare provider is an employee, independent contractor, or

---

91 See Bing, 143 N.E.2d at 8; Abraham & Weiler, supra note 28, at 385–86.
92 See Bing, 143 N.E.2d at 8; Abraham & Weiler, supra note 28, at 385–86.
94 See President & Dirs. of Georgetown Coll. v. Hughes, 130 F.2d 810, 823–24 (D.C. Cir. 1942); Abraham & Weiler, supra note 28, at 386–92. Some states still maintain a diluted system of charitable immunity where the amount a patient can recover from a nonprofit hospital is capped. See Note, The Quality of Mercy: “Charitable Torts” and Their Continuing Immunity, 100 Harv. L. Rev. 1382, 1398 (1987) (arguing charitable immunity should be eliminated where it remains).
95 See Adamski, 579 P.2d at 978–79; Abraham & Weiler, supra note 28, at 386–89.
96 See Adamski, 579 P.2d at 978–79; Abraham & Weiler, supra note 28, at 386–89.
98 Abraham & Weiler, supra note 28, at 393. A hospital could be found liable under a theory of corporate liability when it, for example, negligently grants staff privileges to an inadequate physician. See id. at 381, 389, 393.
99 See id. at 393–94.
holder of admitting privileges. An exclusive enterprise liability system would allow patients to sue only hospitals, thus insulating individual providers from liability. By limiting the liability to hospitals, medical-malpractice suits would appreciably simplify, as there would be only one defendant available to potential plaintiffs.

Enterprise liability had a brief moment in the spotlight in the spring of 1993 when the Clinton Administration began vetting its proposals for healthcare reform. The original Clinton plan included a national system of enterprise liability. The AMA and the Physician Insurer Association of America (“PIAA”) aggressively opposed this proposal. Although one might expect a physicians’ lobbying group to support legislation that would immunize physicians from medical-malpractice liability, the AMA believed that by making hospitals exclusively liable for malpractice, physician autonomy would be curtailed by hospitals and administrators. The lobbying effort of the AMA and PIAA succeeded, and the final Clinton proposal only suggested pilot programs to test enterprise liability. The Clinton reforms ultimately died in Congress, but enterprise liability has remained a tantalizing idea for systemic reform to many health-policy experts.

An enterprise liability approach to medical malpractice in the United States would maintain most of the infrastructure and legal norms of the current medical-malpractice system. Liability would still be based on a finding of fault. The key change would make

---

100 See id. at 393.
101 See id. at 393–94. A non-exclusive system would still allow patients to sue individual providers. See id. Individual providers could continue to insure against personal liability as they do under the current system. See id.
102 See Jacobi & Huberfeld, supra note 29, at 307.
105 See Abraham & Weiler, supra note 28, at 383; Sage, supra note 104, at 437 n.44; Sage, supra note 103, at 410.
106 See Abraham & Weiler, supra note 28, at 383; Sage, supra note 104, at 437 n.44; Sage, supra note 103, at 410.
107 Abraham & Weiler, supra note 28, at 384; Weiler, supra note 104, at 223.
108 See Abraham & Weiler, supra note 28, at 394–95; Weiler, supra note 104, at 223.
109 See Abraham & Weiler, supra note 28, at 393.
110 See id. at 434.
hospitals assume the legal liability of individual providers, even those with off-site facilities.\textsuperscript{111} Doctors would need to submit themselves to additional control and supervision from hospitals in exchange for immunity from personal liability for malpractice.\textsuperscript{112} An enterprise liability system would funnel all malpractice claims to hospitals without changing the applicable law of negligence.\textsuperscript{113}

C. No-Fault Liability

No-fault also emerged as a possible systemic reform for medical-malpractice law in the 1970s.\textsuperscript{114} During the malpractice crisis of the 1970s, reformers suggested creating a no-fault system of liability for medical malpractice similar to the workers’ compensation scheme being developed at the time.\textsuperscript{115} In general terms, a no-fault system would eliminate the need for courts to find doctors or other healthcare providers negligent for an injured patient to receive compensation.\textsuperscript{116} Instead, a patient would automatically recover for injuries caused by medical care through an administrative system in which an injured patient would fill out a form and then a review board would process the claim.\textsuperscript{117} With the medical-malpractice crisis subsiding and scholars predicting that a no-fault system would be more costly due to higher compensation rates, interest dissipated.\textsuperscript{118}

In the last ten years, interest in creating a no-fault medical-malpractice system has resurfaced because of its potential to reduce medical error and improve patient safety.\textsuperscript{119} In 2000, the Institute of Medicine issued a report citing that between 44,000 and 98,000 deaths are caused every year by medical error.\textsuperscript{120} This report highlighted a pressing need to improve the way the U.S. healthcare system

\textsuperscript{111} See id. at 393–94; Furrow, supra note 20, at 109; Weiler, supra note 104, at 224.
\textsuperscript{112} See Furrow, supra note 20, at 109, 112.
\textsuperscript{113} See Abraham & Weiler, supra note 28, at 434; Weiler, supra note 104, at 224.
\textsuperscript{114} Weiler, supra note 10, at 910.
\textsuperscript{116} Studdert et al., supra note 31, at 6; Weiler, supra note 104, at 227; Weiler, supra note 10, at 919–20.
\textsuperscript{117} See Studdert et al., supra note 31, at 6; Weiler, supra note 10, at 919–20.
\textsuperscript{118} See Weiler, supra note 10, at 910–11.
\textsuperscript{119} See Mark A. Hall, Can You Trust a Doctor You Can’t Sue?, 54 DePaul L. Rev. 303, 309 (2005); Studdert & Brennan, supra note 79, at 217; Weiler, supra note 104, at 227; Weiler, supra note 10, at 912.
\textsuperscript{120} Inst. of Med., supra note 13, at 1.
identifies and manages error.\textsuperscript{121} Currently, there is not enough incentive for healthcare providers to report errors.\textsuperscript{122} Consequently, valuable information is being lost.\textsuperscript{123} A system that encourages providers to report errors would provide opportunities to identify recurring errors and ways to correct them.\textsuperscript{124} The fault-based medical-malpractice system provides the opposite incentive because admitting errors is an invitation to a lawsuit.\textsuperscript{125} Recognizing this tension between fault-based liability and the need to gather information to limit medical error, reformers have a renewed interest in creating a no-fault system.\textsuperscript{126}

Under a no-fault system, hospitals would have a strong financial incentive to gather information about errors and reduce them because hospitals would be compensating patients for their injuries.\textsuperscript{127} Regardless of whether the error was an act of negligence or a mistake that did not breach the standard of care, the hospital would have to compensate patients for injuries caused by the error.\textsuperscript{128} Physicians would be more willing to discuss cases candidly, including errors, because they would not be subject to individual liability, and their fault would not affect the hospital’s liability.\textsuperscript{129} Thus, a no-fault system could be a catalyst for significant improvements in patient safety and care.\textsuperscript{130}

Although the medical-malpractice system is largely based on a negligence theory of liability, the tort system has pockets of strict liability that could serve as models for a no-fault medical-malpractice system.\textsuperscript{131} Strict liability has deep roots, dating to English common law.\textsuperscript{132} Before negligence became the norm in tort law, tortfeasors


\textsuperscript{123} See Gostin, supra note 122, at 1743.

\textsuperscript{124} See Hall, supra note 119, at 309; Studdert & Brennan, supra note 79, at 219.

\textsuperscript{125} See Studdert & Brennan, supra note 79, at 218–19.


\textsuperscript{127} See Studdert & Brennan, supra note 79, at 221.

\textsuperscript{128} See id.

\textsuperscript{129} See id.

\textsuperscript{130} See id.

\textsuperscript{131} See DAN B. DOBBS & PAUL T. HAYDEN, TORTS AND COMPENSATION 556 (4th ed. 2001); Weiler, supra note 10, at 910–11 (discussing worker’s compensation as a strict liability model for no-fault).

\textsuperscript{132} DOBBS & HAYDEN, supra note 131, at 590–92.
were held strictly liable for injuries they caused regardless of their fault.\textsuperscript{133} As long as tortfeasors caused an injury, they were liable for damages.\textsuperscript{134} Over time, negligence emerged as a fairer approach to distributing the costs of risks and injuries, but strict liability did not entirely disappear.\textsuperscript{135}

Workers’ compensation is a contemporary example of no-fault liability in the U.S. legal system.\textsuperscript{136} Under workers’ compensation statutes, when an employee is injured in the course of employment, the employer compensates the employee for the injury regardless of the employer’s fault in causing the injury.\textsuperscript{137} All states and the federal government enacted workers’ compensation statutes, most by 1920.\textsuperscript{138} These programs, administered by states or private insurers, hold employers strictly liable for injuries arising in the course of employment.\textsuperscript{139} An employer is liable for fixed amounts that are set out for specific injuries in the statutes.\textsuperscript{140} The employee can recover medical expenses and lost wages, but cannot recover for pain and suffering.\textsuperscript{141} Injured employees are entitled to immediate and periodic payments as long as the disability exists.\textsuperscript{142} Disputes are settled by an administrative agency.\textsuperscript{143} The system is funded through insurance that employers are required, or strongly encouraged, to purchase through the workers’ compensation statutes.\textsuperscript{144} A no-fault medical-malpractice system might resemble the administrative system of workers’ compensation, though a no-fault medical-malpractice system would probably have a narrower range of compensable injuries.\textsuperscript{145}

Products liability is another area of strict liability in tort law.\textsuperscript{146} A seller of a product may be strictly liable for injuries caused by the product if: (1) the seller is in the business of selling the product, (2) the product is expected to and does reach the user without substantial change, and (3) the product was sold in a defective condition unrea-

\textsuperscript{133} Id.
\textsuperscript{134} Id.
\textsuperscript{135} See id.
\textsuperscript{136} Id at 822.
\textsuperscript{137} DOBBS & HAYDEN, supra note 131, at 822.
\textsuperscript{138} Id.
\textsuperscript{139} Id. at 822, 827.
\textsuperscript{140} Id. at 822.
\textsuperscript{141} Id.
\textsuperscript{142} DOBBS & HAYDEN, supra note 131, at 822.
\textsuperscript{143} Id. at 822–23.
\textsuperscript{144} Id. at 823.
\textsuperscript{145} See id. at 822–23; Studdert et al., supra note 31, at 5–9.
\textsuperscript{146} DOBBS & HAYDEN, supra note 131, at 626.
sonably dangerous to the consumer.\textsuperscript{147} Products liability is already a part of health law.\textsuperscript{148} Pharmaceutical and medical device companies are subject to products liability suits, and the experiences of these industries provide insight into how to design a no-fault medical-malpractice system in light of the special needs and issues of healthcare.\textsuperscript{149}

There even exists a small pocket of no-fault liability in current medical-malpractice law.\textsuperscript{150} Virginia and Florida implemented no-fault systems to provide compensation for babies who suffer neurological injuries during delivery regardless of fault.\textsuperscript{151} The programs provide compensation for these limited injuries regardless of the fault of healthcare providers.\textsuperscript{152} The no-fault models in Virginia and Florida may offer additional guidance for developing a comprehensive no-fault system.\textsuperscript{153}

A comprehensive no-fault system liability could use a wide range of recovery schemes.\textsuperscript{154} The most liberal models allow for recovery for all kinds of injuries with no regard for causation, but this would be expensive and impractical.\textsuperscript{155} Therefore, a no-fault system should only allow compensation for limited injuries.\textsuperscript{156} Furthermore, any feasible no-fault system would also offer compensation based on the level of causation.\textsuperscript{157}

\textsuperscript{147} See Restatement (Second) of Torts § 402(A) (1965); see also Restatement (Third) of Torts: Products Liability § 1 (1998) (“One engaged in the business of selling or otherwise distributing products who sells or distributes a defective product is subject to liability for harm to persons or property caused by the defect.”).

\textsuperscript{148} See Laura Pleicones, Note, Passing the Essence Test: Health Care Providers Escape Strict Liability for Medical Devices, 50 S.C. L. Rev. 463, 464–65 (1999) (discussing a decision of the South Carolina Supreme Court to hold healthcare providers not strictly liable for defective medical devices).


\textsuperscript{150} See Studdert et al., supra note 31, at 1–2; Weiler, supra note 10, at 936 n.87.


\textsuperscript{152} See Weiler, supra note 10, at 936 n.88.

\textsuperscript{153} See id.

\textsuperscript{154} See Studdert et al., supra note 31, at 10–11.

\textsuperscript{155} See id. at 6–10.

\textsuperscript{156} See id.

\textsuperscript{157} See id.; Weiler, supra note 104, at 227 (suggesting compensation only be provided to injuries that last at least two months).
The Swedish system provides a working example of what a no-fault system could look like in the United States.\textsuperscript{158} Sweden uses a no-fault model that compensates injuries caused by treatment that could have been avoided.\textsuperscript{159} In Sweden, if a patient believes an injury was a result of medical care, the patient fills out an application for compensation, usually with the assistance of a physician.\textsuperscript{160} After the patient files the claim, the treating physician fills out a report.\textsuperscript{161} Then, adjustors and physicians working for a national central claims office decide whether the injury was caused by treatment and whether the injuries could have been avoided.\textsuperscript{162} Only patients with injuries caused by treatment that could have been avoided receive compensation.\textsuperscript{163} Although the causation and avoidability test does not clearly indicate exactly which treatment injuries deserve compensation, the structured judgments of the claims office seem to be more predictable and objective than judgments of negligence.\textsuperscript{164} In addition to requiring avoidability to merit compensation, Sweden also instituted a disability threshold to control costs.\textsuperscript{165} To be eligible for no-fault compensation, the patient must have spent at least ten days in the hospital or have used more than thirty sick days.\textsuperscript{166} The Swedish process usually takes six months from filing a claim to receiving a decision.\textsuperscript{167}

A no-fault system in the United States would not be exactly like Sweden’s system, but three characteristics of their system can be used to compare no-fault with enterprise liability as reforms for the United States.\textsuperscript{168} First, a no-fault system would require hospitals to pay for medical errors.\textsuperscript{169} In this respect, a no-fault system would be similar to an enterprise-liability model.\textsuperscript{170} Second, a no-fault system would not inquire whether treatment was negligently provided.\textsuperscript{171} Rather, an inquiry would be made as to whether the injury was avoidable to deter-

\begin{thebibliography}{99}
\bibitem{158} Studdert et al., \textit{supra} note 31, at 8.
\bibitem{159} \textit{See id.}
\bibitem{160} \textit{See id.} at 6.
\bibitem{161} \textit{See id.}
\bibitem{162} \textit{See id.} at 7.
\bibitem{163} Studdert et al., \textit{supra} note 31, at 7.
\bibitem{164} \textit{See id.}
\bibitem{165} \textit{See id.} at 8.
\bibitem{166} \textit{Id.}
\bibitem{167} \textit{Id.} at 6.
\bibitem{168} \textit{See, e.g.,} Abraham & Weiler, \textit{supra} note 28, at 434; Studdert et al., \textit{supra} note 31, at 5–9; Weiler, \textit{supra} note 10, at 919–20.
\bibitem{169} \textit{See Weiler, }\textit{supra} note 10, at 919–20.
\bibitem{170} \textit{See Abraham & Weiler, }\textit{supra} note 28, at 434.
\bibitem{171} \textit{See Weiler, }\textit{supra} note 10, at 919–20.
\end{thebibliography}
mine liability.\textsuperscript{172} Third, a no-fault system would use an administrative body to process claims and make compensation decisions rather than a court of law.\textsuperscript{173}

II. Analysis

A. Framework for Analysis

Enterprise liability and no-fault are both sweeping models of reform for medical-malpractice law.\textsuperscript{174} The two share many characteristics; in fact, enterprise liability is part of the foundation of no-fault.\textsuperscript{175} Although enterprise liability and no-fault have each been discussed in existing literature, the two have not been explicitly compared to determine which would be a better reform.\textsuperscript{176} This Part compares enterprise liability with no-fault to determine which would be the better reform model.\textsuperscript{177}

This Note compares the two models on the basis of four factors.\textsuperscript{178} The four factors are: (1) the goal of compensation, (2) the goal of deterrence, (3) the requirement of economic feasibility, and (4) the requirement of sociopolitical feasibility.\textsuperscript{179} Each of these factors are discussed individually in turn in greater detail, leading to the conclusion that enterprise liability is the better model for reform because it is significantly more feasible.\textsuperscript{180}

The first two factors, deterrence and compensation, are often described as the twin goals of tort law.\textsuperscript{181} Medical-malpractice law similarly serves to prevent injuries and compensate injured patients.\textsuperscript{182} Medical-malpractice law deters healthcare providers from harming patients, and compensates patients for injuries.\textsuperscript{183} Thus, any evalua-
tion must consider how the alternatives serve deterrence and compensation.\textsuperscript{184}

In addition to evaluating how an alternative medical-malpractice system serves deterrence and compensation, the system must also be feasible.\textsuperscript{185} A system’s feasibility should be evaluated by its cost and economic efficiency.\textsuperscript{186} Alternative medical-malpractice systems must be cost-sensitive because the already expensive U.S. healthcare system cannot afford a more costly malpractice system.\textsuperscript{187} A malpractice system also should be efficient.\textsuperscript{188} Efficiency in this context measures what percentage of medical-malpractice costs actually compensate patients and how quickly malpractice claims are resolved.\textsuperscript{189}

Feasibility also should be measured by how well the system fits the prospective cultural and political environment.\textsuperscript{190} A systemic reform of medical-malpractice must rally support and gain political momentum to be enacted.\textsuperscript{191} In addition to political demands, an alternative medical-malpractice system must meet social and cultural needs.\textsuperscript{192} Healthcare is a unique realm of public policy where questions about life, mortality, and human frailty are ever-present.\textsuperscript{193} Health problems can be intensely personal, complicated, and consuming for patients.\textsuperscript{194} Systemic changes in medical-malpractice law will have ramifications that extend far beyond the courtroom.\textsuperscript{195}

\textsuperscript{184} See id.
\textsuperscript{185} See Studdert et al., supra note 31, at 3.
\textsuperscript{186} See id.
\textsuperscript{187} See Nat’l Ctr. for Health Statistics, supra note 25, at 14, 326 (stating that healthcare spending in the United States was $1.5 trillion in 2002, 14.9% of the GDP).
\textsuperscript{188} See Weiler, supra note 10, at 926 (stating the cost of malpractice litigation consumes fifty-five to sixty percent of every claims dollar).
\textsuperscript{189} See id.
\textsuperscript{190} See Skocpol, supra note 56, at 173–78 (discussing the failure of the Clinton Administration’s Health Security proposal and the important role social and political forces played in bringing about this failure); Weiler, supra note 10, at 947–48.
\textsuperscript{191} See Skocpol, supra note 56, at 173–78; Weiler, supra note 10, at 947–48.
\textsuperscript{192} See Skocpol, supra note 56, at 173–78; Weiler, supra note 10, at 947–48.
\textsuperscript{194} See id. at 8; Gibeaut, supra note 68, at 44.
\textsuperscript{195} See Abraham & Weiler, supra note 28, at 436; Studdert et al., supra note 31, at 32–34; Weiler, supra note 10, at 947–48.
B. Goal of Compensation

1. Enterprise Liability

A medical-malpractice system should compensate patients for injuries caused by malpractice. An enterprise-liability system would make compensation more frequent, consistent, and predictable. The current medical-malpractice system compensates few patients relative to the amount of injuries sustained. Additionally, compensation varies significantly from case to case. Enterprise liability would improve patient compensation by streamlining medical-malpractice claims and by making outcomes more predictable. More injured patients would be able to receive compensation because malpractice actions would be simpler and less expensive in so far as plaintiffs would only bring a claim against one defendant, the hospital.

Enterprise liability also promises to improve compensation by making claims more predictable. Hospitals would become more experienced with medical-malpractice claims and would be able to distinguish legitimate claims from frivolous claims more easily. Hospitals might settle more valid claims because their reputations would not be as vulnerable as the reputations of individual physicians who are motivated to fight every claim. Hospitals might focus their resources on fighting frivolous and weak claims and settling stronger claims. Thus, claims would become more predictable and compensation would reflect actual injuries more closely.

196 See Pegalis & Wachsman, supra note 11, § 2:10; Weiler, supra note 104, at 227.
197 See Abraham & Weiler, supra note 28, at 401-06; Edward P. Richards & Thomas R. McLean, Administrative Compensation for Medical Malpractice Injuries: Reconciling the Brave New World of Patient Safety and the Torts System, 49 St. Louis U. L.J. 73, 89 (2004) (stating that an enterprise liability system would increase money available to patients and would compensate patients more often because corporate defendants are less sympathetic to juries); Sage, supra note 104, at 476.
198 See Hyman, supra note 10, at 1643.
199 See id. at 1643-44.
200 See Abraham & Weiler, supra note 28, at 406.
201 See id. at 403, 406; Weiler, supra note 104, at 224–25.
202 See Abraham & Weiler, supra note 28, at 403; Furrow, supra note 20, at 101, 109.
203 See Abraham & Weiler, supra note 28, at 403.
204 See id. at 404; Weiler, supra note 10, at 916.
205 See Abraham & Weiler, supra note 28, at 404, 406 (discussing how damages might be standardized by an enterprise liability system).
206 See id. at 403, 406; Mark Geistfeld, Malpractice Insurance and the (Il)legitimate Interests of the Medical Profession in Tort Reform, 54 DePaul L. Rev. 436, 459–60 (2005) (comparing damage caps with enterprise liability and finding that enterprise liability would make malpractice premiums more representative and fairer).
Enterprise liability would improve compensation, but maintaining the fault requirement should limit compensation to the same kind of injuries and claims that are successful under the current system.\textsuperscript{207} Enterprise liability would improve compensation largely through making the litigation process more efficient and cost-effective.\textsuperscript{208}

2. No-fault

An enterprise liability system would improve compensation through gains in efficiency.\textsuperscript{209} A no-fault system would couple these improvements with an expansion of compensation to more injured patients.\textsuperscript{210} A study in 1997 tested the economic feasibility of a no-fault system like Sweden’s in Colorado and Utah.\textsuperscript{211} The study found that, even by conservative estimates, two to three times the number of patients would be compensated in Utah and Colorado respectively, while only modestly increasing the cost relative to the negligence-based malpractice system.\textsuperscript{212} A dramatic increase in compensation like this would alleviate the problem of deserving patients not receiving compensation.\textsuperscript{213}

Potential economic problems temper the advantages of the no-fault system.\textsuperscript{214} A larger number of claimants might spread compensation resources too thin.\textsuperscript{215} Although more patients would be eligible for compensation, there may not be adequate funds to compensate these patients meaningfully.\textsuperscript{216} Because of limited resources, it may be better to limit compensation to cases where there has been a judicial finding of negligence.\textsuperscript{217} The goal of compensation would not be well

\begin{flushleft}
\begin{tabular}{l}
\textsuperscript{207} See Abraham & Weiler, supra note 28, at 432–33.  \\
\textsuperscript{208} See id. at 406; Weiler, supra note 104, at 224–25.  \\
\textsuperscript{209} See Abraham & Weiler, supra note 28, at 406; Bovbjerg & Sloan, supra note 126, at 71; Weiler, supra note 104, at 224–25.  \\
\textsuperscript{210} See Abraham & Weiler, supra note 28, at 406; Bovbjerg & Sloan, supra note 126, at 70; Weiler, supra note 10, at 919–20.  \\
\textsuperscript{211} See generally Studdert et al., supra note 31 (describing the study and presenting its results).  \\
\textsuperscript{212} Id. at 29–30.  \\
\textsuperscript{213} See Barbara Brill, An Experiment in Patient Injury Compensation: Is Utah the Place?, 1996 Utah L. Rev. 987, 1003; Studdert et al., supra note 33, at 29–30.  \\
\textsuperscript{214} See Mary Ann Kupeli, Survey, Tort Law = No Fault Compensation: An Unrealistic Elixir to the Medical Malpractice Ailment, 19 Suffolk Transnat’l L. Rev. 559, 568–71 (1996).  \\
\textsuperscript{215} See Bovbjerg & Sloan, supra note 126, at 73; Studdert et al., supra note 31, at 29 (estimating that no-fault liability in Colorado and Utah would have a higher direct cost than the existing negligence system).  \\
\textsuperscript{216} See Bovbjerg & Sloan, supra note 126, at 73; Studdert et al., supra note 31, at 29.  \\
\textsuperscript{217} See Kupeli, supra note 214, at 560. But see Studdert et al., supra note 31, at 33.
\end{tabular}
\end{flushleft}
served if a no-fault system opened the door for recovery so wide that all patients, especially the most severely injured and deserving, would not be adequately compensated for their losses.\textsuperscript{218}

3. Which System Is Better?

A no-fault system would be a more effective way to compensate injured patients than enterprise liability.\textsuperscript{219} No-fault promises to compensate a greater number of injured patients more consistently and appropriately.\textsuperscript{220} A no-fault system runs the risk of casting the net too wide and compensating undeserving patients, but because there are limited resources available to compensate patients, it is likely that only legitimately injured patients would be compensated.\textsuperscript{221} The social advantages of assisting people with injuries caused by medical treatment, even if the injuries were caused by mistakes rather than negligence, are high.\textsuperscript{222}

In addition to increasing the number of compensable injuries and parties, a no-fault system would also serve to standardize compensation.\textsuperscript{223} Currently, the fault-based medical-malpractice system creates the possibility that one injured patient could hit the metaphorical jackpot with their medical-malpractice claim and be awarded huge punitive damages in addition to compensatory damages.\textsuperscript{224} This would not change in a fault-based enterprise liability system.\textsuperscript{225} These large awards are designed to serve deterrence goals rather than compensation.\textsuperscript{226} By standardizing compensation under a no-fault system, the punitive damages would be spread across many injured parties, thus providing more people with compensation and having compensation better reflect the nature of the injuries sustained.\textsuperscript{227} Compensation is not well-served when so few patients recover for their injuries, nor is it served by having like patients receive different amounts.\textsuperscript{228}

\textsuperscript{218} See Studdert et al., \textit{supra} note 31, at 29.
\textsuperscript{219} See \textit{id.} at 33; Weiler, \textit{supra} note 104, at 227; Weiler, \textit{supra} note 10, at 921–22.
\textsuperscript{220} See Studdert et al., \textit{supra} note 31, at 29; Weiler, \textit{supra} note 10, at 921–22.
\textsuperscript{221} See Bovbjerg & Sloan, \textit{supra} note 126, at 70, 73; Studdert et al., \textit{supra} note 31, at 29–31; Weiler, \textit{supra} note 10, at 921–22.
\textsuperscript{222} See Bovbjerg et al., \textit{supra} note 31, at 29–31; Weiler, \textit{supra} note 10, at 921–22.
\textsuperscript{223} See Bovbjerg & Sloan, \textit{supra} note 126, at 70–71; Weiler, \textit{supra} note 10, at 922.
\textsuperscript{224} See Weiler, \textit{supra} note 10, at 922.
\textsuperscript{225} See \textit{id}.
\textsuperscript{226} See \textit{id}.
\textsuperscript{227} See Bovbjerg & Sloan, \textit{supra} note 126, at 70–71; Weiler, \textit{supra} note 10, at 922.
\textsuperscript{228} See Randall R. Bovbjerg et al., \textit{Administrative Performance of “No-Fault” Compensation for Medical Injury}, 60 LAW & CONTEMP. PROBS., Spring 1997, at 71, 71–72.
Enterprise liability would not improve compensation nearly as much as a no-fault system because enterprise liability would continue to compensate roughly the same inadequate number of patients.\(^{229}\) An enterprise liability system would only improve compensation relative to the current system by giving hospitals more experience with malpractice claims, thereby helping them more easily and consistently settle legitimate claims and oppose weak claims.\(^{230}\) Although these improvements should not be overlooked, relative to a no-fault system, enterprise liability does not serve compensation goals as well.\(^{231}\)

C. Goal of Deterrence

1. Enterprise Liability

A medical-malpractice system should deter physicians from committing malpractice and encourage them to provide the best possible care.\(^{232}\) Enterprise liability would deter physicians from committing malpractice more effectively than the current system, and improve patient safety.\(^{233}\) An enterprise liability system would give hospitals added incentive to gather more data about medical errors, which might reveal patterns, revealing potential ways to eliminate errors.\(^{234}\) Additionally, the hospital would have additional incentive to correct or remove inadequate physicians.\(^{235}\) Currently, peer review boards are not effective because there is too much professional courtesy.\(^{236}\) By imposing liability on hospitals for all medical-malpractice, hospitals would have a heightened financial incentive to identify physicians providing substandard care, correct their treatments, or stop the physicians from practicing.\(^{237}\) Defensive medicine also would be curbed since physicians would no longer fear being held personally

\(^{229}\) Compare Abraham & Weiler, supra note 28, at 401–06 (discussing the ability of enterprise liability to provide compensation), with Weiler, supra note 10, at 922–24 (discussing how a no-fault system could compensate more patients more equitably than the current system).

\(^{230}\) See Abraham & Weiler, supra note 28, at 403, 406.

\(^{231}\) Compare Abraham & Weiler, supra note 28, at 401–06 (discussing the ability of enterprise liability to provide compensation), with Weiler, supra note 10, at 922–24 (discussing the ability of no-fault to provide compensation).

\(^{232}\) See Pegalis & Wachsman, supra note 11, § 2:10.

\(^{233}\) See Abraham & Weiler, supra note 28, at 407–14; Furrow, supra note 20, at 101.

\(^{234}\) See Abraham & Weiler, supra note 28, at 413; Furrow, supra note 20, at 110.

\(^{235}\) See Abraham & Weiler, supra note 28, at 413–14; Furrow, supra note 20, at 110.

\(^{236}\) See Pegalis & Wachsman, supra note 11, § 2:10.

\(^{237}\) See Abraham & Weiler, supra note 28, at 413–14; Furrow, supra note 20, at 110.
liable for malpractice. Because defensive medicine can result in unnecessary care, which puts patients at higher risk of iatrogenic injury, a reduction would improve patient safety.

Another advantage of an enterprise-liability system would be that, unlike individual physicians under the current system, hospitals’ medical-malpractice insurance would become experience-rated. This would mean that hospitals with better safety records would have access to less expensive insurance. This would give hospitals a financial incentive to improve patient safety by looking for systemic improvements and by pressuring their individual providers to provide better, safer care.

The major drawback to an enterprise liability system with respect to deterrence is that it would lessen the responsibility of individual physicians to give their patients the best possible care. Deterrence is sacrificed at the individual level in hopes of making gains in patient safety by providing greater incentives to gather and analyze data about medical error.

2. No-fault

A no-fault system of liability should also facilitate efforts to improve patient safety. A no-fault system would encourage hospitals to educate individual providers more aggressively about patient safety and to discipline them for providing inadequate care. A no-fault system would magnify the incentives for hospitals to improve patient safety.

A no-fault system would hold hospitals liable for a broader range of injuries caused by medical care than a negligence system. This increased liability would encourage hospitals to find more ways to re-

239 See id.; Sage, supra note 104, at 475–76.
241 See Abraham & Weiler, supra note 28, at 403–04, 409–11; see also Geistfeld, supra note 206, at 460 (stating that enterprise liability would replace individual premiums with enterprise premiums, which would be fairer).
242 See Abraham & Weiler, supra note 28, at 403–04, 409–11; Furrow, supra note 20, at 110.
243 See Kupeli, supra note 214, at 570.
244 See Abraham & Weiler, supra note 28, at 407–14; Furrow, supra note 20, at 110.
248 See Weiler, supra note 104, at 227; Weiler, supra note 10, at 919–20.
duce medical error, including error not caused by negligence.\textsuperscript{249} Physicians might have less reason to be hesitant about sharing their concerns about potential errors because their fault would not affect the liability of the hospital.\textsuperscript{250} The advantages of an enterprise liability system would be heightened by a no-fault system, which would further relieve individual providers from their unproductive fear of liability and give hospitals greater incentive to reduce medical error.\textsuperscript{251}

Like enterprise liability, no-fault also has the problem of potentially reducing the accountability of individual providers by shifting liability to hospitals.\textsuperscript{252} An enterprise liability system would remove the accountability from individuals and shift it to a hospital, but hospitals would still only lose cases when a provider is found to be at fault.\textsuperscript{253} Under an enterprise liability system, the law would still focus on the individual actions of providers.\textsuperscript{254} This focus pressures physicians to provide perfect care.\textsuperscript{255} A shift to a no-fault system would reduce this pressure and might allow physicians to be careless without fear of retribution from the legal system.\textsuperscript{256}

Alternatively, a no-fault system might make doctors more careful because they would know that their errors, even errors that were not the result of negligence, would still merit compensation.\textsuperscript{257} Individual providers would have incentive to keep their hospitals happy by doing everything they could to keep their error rates as low as possible.\textsuperscript{258} In a profession steeped with moral and ethical obligations, this may be the more likely effect of a no-fault system.\textsuperscript{259}

\textsuperscript{249} See Hall, supra note 119, at 309–10; Weiler, supra note 10, at 939.
\textsuperscript{250} See Weiler, supra note 10, at 939.
\textsuperscript{251} See Weiler, supra note 104, at 228–30; Weiler, supra note 10, at 939.
\textsuperscript{252} See Kupeli, supra note 214, at 570.
\textsuperscript{253} See Abraham & Weiler, supra note 28, at 383.
\textsuperscript{254} See id.; Furrow, supra note 20, at 109.
\textsuperscript{255} See Abraham & Weiler, supra note 28, at 383.
\textsuperscript{256} See Kupeli, supra note 214, at 570. But see Weiler, supra note 104, at 223 (noting that individual providers are already distanced from personal liability because insurance almost always pays the costs of malpractice suits).
\textsuperscript{257} See Weiler, supra note 10, at 938.
\textsuperscript{258} See id.
\textsuperscript{259} See Pegalis & Wachsmann, supra note 11, § 2:8 (listing some of the ethical guidelines of the practice of medicine); Weiler, supra note 10, at 938. The AMA Principles of Medical Ethics are available at http://www.ama-assn.org/ama/pub/category/2512.html (last visited Aug. 14, 2005).
3. Which System Is Better?

Enterprise liability and no-fault would both facilitate improvements in patient safety, but a no-fault system has the potential to better serve the goal of deterring healthcare providers from injuring their patients.260 Under a no-fault system, efforts to improve patient safety could thrive because individual providers could disclose errors without fear of individual liability.261 Additionally, a no-fault system would provide hospitals with more incentive to improve care than under an enterprise liability system because hospitals would have to compensate patients even if there was no negligence and even if a patient did not file a lawsuit.262 By forcing hospitals to pay for more adverse medical outcomes, hospitals would have a larger incentive to improve safety.263

A no-fault system would do a better job of reducing the practice of defensive medicine than an enterprise-liability system.264 No-fault relieves the pressure of legal liability from individual physicians more than enterprise liability.265 The less fear physicians have of malpractice claims, the less likely they are to prescribe unnecessary care simply to shield themselves from liability.266 No-fault takes an additional step in reducing defensive medicine because medical errors that stem from defensive care will merit compensation, even if there was no fault.267 Thus, no-fault makes defensive medicine even more unnecessary and risky because defensive medicine would expose hospitals to greater liability by increasing the opportunity time for medical error to occur.268

D. Requirement of Economic Feasibility

For a medical-malpractice reform to compensate patients better and deter malpractice, it must be feasible.269 For purposes of this Note, economic efficiency considers the total cost of a system: the administra-

260 Compare Abraham & Weiler, supra note 28, at 407–14 (discussing how enterprise liability would work as an injury-prevention system), with Studdert & Brennan, supra note 79, at 220 (discussing how no-fault could prevent error).
261 See Weiler, supra note 10, at 916–17, 942.
264 See Bovbjerg & Sloan, supra note 126, at 72; Weiler, supra note 10, at 916–17, 942.
265 See Weiler, supra note 10, at 916–17, 942.
266 See Bovbjerg & Sloan, supra note 126, at 72; Weiler, supra note 10, at 916–17, 942.
267 See Weiler, supra note 10, at 938–39.
268 See id. at 938–39, 942.
tive costs relative to compensation and the amount of time necessary to resolve medical-malpractice claims.  

1. Enterprise Liability

An enterprise-liability system would save costs relative to the current medical-malpractice system.  

The range of compensable events would not necessarily increase. Lawsuits would be more straightforward because patients would simply sue hospitals, not individual providers involved in treatment. Increased efficiency would reduce the costs and time spent defending and bringing malpractice claims. An enterprise-liability system would also realize savings if improvements in patient safety materialize; thus, reducing the actual incidence of medical malpractice and compensable events. Finally, doctors would no longer need the costly individual malpractice insurance that is driving some physicians out of their practices.

Enterprise liability also has some financial risks. Implementing an enterprise-liability system might be difficult for hospitals, especially nonprofit hospitals or those hospitals in poor areas already operating on a tight budget. Some hospitals are struggling to pay for emergency medical care they must provide regardless of whether patients can pay for it. Giving hospitals the increased responsibility of physicians’ medical-malpractice liability may prove to be too much for some hospitals. For this reason, any attempt to institute enterprise liability should begin with hospitals that have the financial strength to test the economic feasibility of the system.

270 See Weiler, supra note 10, at 926.
272 See Abraham & Weiler, supra note 28, at 403–04.
273 See Abraham & Weiler, supra note 28, at 403–06; Weiler, supra note 104, at 224–25.
274 See Abraham & Weiler, supra note 28, at 403–04; Jacobi & Huberfeld, supra note 29, at 307.
276 See id. at 383; Geistfeld, supra note 206, at 459–60; Burages, supra note 6, at 8A.
277 See Abraham & Weiler, supra note 28, at 423, 426.
278 See id. at 423–27.
279 See id.
280 See id.
281 See id.
2. No-fault

A no-fault approach to medical-malpractice liability would probably be more expensive than the current negligence-based model.\(^ {282}\) With healthcare costs rising, any systemic change that increases the cost of healthcare should be carefully scrutinized.\(^ {283}\) A no-fault scheme may begin a slippery slope towards excessive compensation.\(^ {284}\) Currently, few of the potentially compensable claims are actually brought as medical-malpractice actions.\(^ {285}\) In the negligence-based medical-malpractice system, a large investment of time and money is required to bring a suit, and a favorable outcome for patients is far from guaranteed.\(^ {286}\) If barriers to compensation were lowered, a floodgate of claims could be opened.\(^ {287}\) If a no-fault system was not carefully designed to contain costs, the surge in claims could overwhelm the financial stability of the system.\(^ {288}\)

Despite valid concerns about the risk of increasing the cost of healthcare by creating a no-fault system, there may be some cost advantages relative to the current system.\(^ {289}\) A no-fault system might prove to be an easier system and fairer way to manage costs.\(^ {290}\) Although tort reform efforts such as capping punitive damages or limiting the contingency fees of attorneys have been successful in reducing the number of malpractice claims, the effects of the reforms are not equitably spread throughout society.\(^ {291}\) They tend to limit options for poor plaintiffs and for plaintiffs with difficult cases.\(^ {292}\)

A no-fault system might make the effects of cost control mechanisms more predictable and equitable.\(^ {293}\) For example, to reduce the

---

\(^ {282}\) See Bovbjerg & Sloan, supra note 126, at 73; Studdert et al., supra note 31, at 31–32.


\(^ {284}\) See Bovbjerg & Sloan, supra note 126, at 73; Kupeli, supra note 214, at 570 (expressing concern about how a no-fault system would be funded).

\(^ {285}\) See Hyman, supra note 10, at 1643; Weiler, supra note 10, at 913.

\(^ {286}\) See Weiler, supra note 10, at 913.

\(^ {287}\) See Kupeli, supra note 214, at 570; Studdert et al., supra note 31, at 2–3; Weiler, supra note 10, at 913.

\(^ {288}\) See Kupeli, supra note 214, at 570; Studdert et al., supra note 31, at 2–3; Weiler, supra note 10, at 913.

\(^ {289}\) See Studdert et al., supra note 33, at 30–31.

\(^ {290}\) See id.; Weiler, supra note 10, at 927.

\(^ {291}\) See Studdert & Brennan, supra note 32, at 225; Weiler, supra note 10, at 910.

\(^ {292}\) See Weiler, supra note 10, at 910; Lohr, supra note 23, at B11.

\(^ {293}\) See Bovbjerg & Sloan, supra note 135, at 71 (stating that no-fault would make compensation more efficiently delivered, better tailored to individuals, and better managed); Weiler, supra note 10, at 921–23. But see Kupeli, supra note 229, at 569.
costs within a no-fault system, one might consider increasing the dis-
ability threshold necessary to receive compensation.\(^{294}\) This would
control costs by eliminating some of the more minor claims, and it
would favor people with the most severe injuries.\(^{295}\) Because damages
in a no-fault system would be dispersed in a regulated and controlled
fashion, it would be easier to anticipate the effect of changes in the
compensation requirements.\(^{296}\) In the current torts system, the un-
predictable nature of damage awards makes it difficult to anticipate
how much will be saved by cost-cutting measures and who will be af-
fected.\(^{297}\)

A no-fault system could be even less expensive than the existing
negligence system.\(^{298}\) If the system were carefully designed to control
costs and limit payments, it does have the potential to reduce the costs
of the healthcare system.\(^{299}\) A no-fault system could produce savings
by greatly reducing physician malpractice insurance, by reducing de-
fensive medicine, by increasing patient safety, and by more efficiently
compensating claimants.\(^{300}\) These savings combined with careful de-
sign might make the cost of a no-fault system an advantage instead of
a weakness.\(^{301}\)

The current system spends about fifty-five to sixty cents on the
dollar to compensate patients for medical malpractice.\(^{302}\) This is not
very efficient.\(^{303}\) More patients would be compensated for the same
amounts of money under an administrative system like no-fault.\(^{304}\)

A no-fault system of liability might run into problems of
efficiency to the extent that it would be a much more bureaucratic
system than the negligence model.\(^{305}\) Paper work, regulations, and
review boards could potentially choke the system with red tape.\(^{306}\) The

\(^{294}\) See Studdert et al., supra note 31, at 10–11 (discussing how the Swedish no-fault sys-
tem controls cost); Weiler, supra note 104, at 227 (suggesting a two month injury thresh-
old).

\(^{295}\) See Studdert et al., supra note 31, at 10–11; Weiler, supra note 104, at 227.

\(^{296}\) See Studdert et al., supra note 31, at 12–13 (explaining cost control in Sweden’s no-
fault system).

\(^{297}\) See Weiler, supra note 10, at 914.

\(^{298}\) See Bovbjerg & Sloan, supra note 126, at 70–73; Weiler, supra note 10, at 926–27.

\(^{299}\) See Studdert et al., supra note 31, at 33.

\(^{300}\) See Weiler, supra note 10, at 926–27.

\(^{301}\) See Studdert et al., supra note 31, at 33.

\(^{302}\) Weiler, supra note 10, at 926.

\(^{303}\) See id.

\(^{304}\) See Studdert et al., supra note 31, at 30; Weiler, supra note 10, at 926.

\(^{305}\) See Kupeli, supra note 214, at 570–71; Weiler, supra note 10, at 926.

\(^{306}\) See Studdert et al., supra note 31, at 9–10 (describing Sweden’s more administrative
no-fault system).
initial system, subsequent statutes, and regulations would have to be attentive to streamlining the administrative system in order to maximize funds for patient compensation.\footnote{See id. at 5–10.}

3. Which System Is Better?

Although no-fault serves the goals of compensation and deterrence well, an enterprise liability system is more economically feasible.\footnote{Compare Abraham & Weiler, \textit{supra} note 28, at 406 (discussing how enterprise liability would be less expensive than the current system), \textit{with} Studdert et al., \textit{supra} note 31, at 29–32 (discussing the affordability of no-fault).} An enterprise liability system has two key advantages over a no-fault system.\footnote{Compare Abraham & Weiler, \textit{supra} note 28, at 406 (discussing how enterprise liability would be less expensive than the current system), \textit{with} Studdert et al., \textit{supra} note 31, at 29–32 (discussing the affordability of no-fault).} First, an enterprise liability system would not drastically expand the number of patients or injuries meriting compensation like a no-fault system would.\footnote{Compare Abraham & Weiler, \textit{supra} note 28, at 393 (discussing how malpractice would still be a requirement for imposing liability), \textit{with} Studdert & Brennan, \textit{supra} note 79, at 219 (explaining how a no-fault system eliminates the need to find negligence to impose liability).} This characteristic may be a weakness of enterprise liability’s capacity to serve the goal of compensation, but it is an economic advantage.\footnote{Compare Abraham & Weiler, \textit{supra} note 28, at 406 (discussing how enterprise liability would be less expensive than the current system), \textit{with} Studdert et al., \textit{supra} note 31, at 29–32 (discussing the affordability of no-fault).} A no-fault system probably would increase the cost of the medical-malpractice system.\footnote{See Studdert et al., \textit{supra} note 31, at 29–32.} Enterprise liability, in contrast, should save money through increased efficiency.\footnote{See Abraham & Weiler, \textit{supra} note 28, at 406.} Malpractice claims should become more predictable and lawsuits should become streamlined and less expensive under an enterprise-liability system.\footnote{See id. at 432.}

The second economic advantage enterprise liability has over no-fault is that it should have fewer initial start-up costs.\footnote{See id. at 432.} A no-fault system would require all initial costs of an enterprise liability system plus the cost of establishing an administrative body to review and process
claims. Since enterprise liability is a less radical change, it should be easier and less expensive to implement enterprise liability.

E. Requirement of Sociopolitical Feasibility

1. Enterprise Liability

Enterprise liability has sociopolitical appeal because it would relieve physicians from individual liability. Physicians are a sympathetic group to the electorate, and the success of recent tort reforms at the state level indicates the political will to relieve the burden of malpractice from physicians. Also, more attention is being dedicated to the problem of medical error and the startling number of injuries and deaths it causes. As enterprise liability promises to identify and reduce medical error more effectively, it could gain political momentum.

At the same time, enterprise liability may face substantial sociopolitical challenges. Enterprise liability fell flat as a proposed reform in 1996, and if anything, the interest groups that opposed the reforms are even more powerful now. Enterprise liability may be viewed as a step too close to socialized medicine to be viable in the United States. Enterprise liability focuses the medical-malpractice system on hospitals. Although this focus would streamline the malpractice system, it would also make hospitals more involved in the provision of care. The American people seem averse to socialized medicine, and attempts to institute an enterprise liability system may, to some, look like an attempt to enact socialized medicine.

See id.; Studdert et al., supra note 31, at 30.

Compare Abraham & Weiler, supra note 28, at 406, 432 (discussing how enterprise liability could be less expensive than the current system and a bridge to later reforms), with Studdert et al., supra note 31, at 29–32 (discussing the costs of a no-fault system).

See Abraham & Weiler, supra note 28, at 382.

See Am. Med. Ass’n, supra note 21, at 24–41; Gawande, supra note 193, at 11.


See id. at 383.

See id.

See id. at 383–84; Keith Myers, Medical Errors: Causes, Cures, and Capitalism, 16 J.L. & Health 255, 288 (2002) (arguing that socialized medicine would reduce incentives to reduce medical error).

See Abraham & Weiler, supra note 28, at 382–83.

See id. at 383.

See Skocpol, supra note 56, at 6–8, 174; Myers, supra note 324, at 288.
cusing malpractice on hospitals, doctors would become less autonomous and more like employees of the hospitals.\textsuperscript{328} Patients seem to want their doctors to direct their care and treatment, rather than an administrative unit such as a hospital.\textsuperscript{329}

Involving the hospital more deeply in the provision of care seems to compromise the intimacy and trust of the doctor-patient relationship.\textsuperscript{330} These effects and perceptions of an enterprise liability system are important because if they are not addressed, the system may never be instituted.\textsuperscript{331} American culture focuses a great deal on individualism, and enterprise liability is more concerned with systems than personal relationships.\textsuperscript{332} The existing fault-based medical-malpractice system is focused on the individual patient and individual providers.\textsuperscript{333} A move toward enterprise liability would sacrifice some of this individualism.\textsuperscript{334}

An enterprise liability system also faces design challenges.\textsuperscript{335} As an example, how would the system deal with a malpractice claim against a rural solo practitioner where the injury happened in the doctor’s office?\textsuperscript{336} Is it fair to assign liability to a far-off hospital that essentially has no control over the rural physician?\textsuperscript{337} How would you choose which hospital should pay for an injury at a doctors’ office when the doctor has staff privileges at multiple hospitals?\textsuperscript{338} These design questions are not easy to answer.\textsuperscript{339} The fact that enterprise liability will be challenging to design does make it less politically and culturally feasible.\textsuperscript{340}

\textsuperscript{328} See Abraham & Weiler, \textit{supra} note 28, at 383.
\textsuperscript{329} See Gawande, \textit{supra} note 193, at 11–12; Abraham & Weiler, \textit{supra} note 28, at 383.
\textsuperscript{330} See Gawande, \textit{supra} note 207, at 11–12; Abraham & Weiler, \textit{supra} note 28, at 383. \textit{But see} Hall, \textit{supra} note 119, at 309–10 (suggesting that a system that encourages doctors to disclose errors to patients might improve trust).
\textsuperscript{331} See Skocpol, \textit{supra} note 56, at 178–83; Abraham & Weiler, \textit{supra} note 28, at 383.
\textsuperscript{333} See Pegalis & Wachsman, \textit{supra} note 11, § 2:10.
\textsuperscript{334} See Abraham & Weiler, \textit{supra} note 28, at 382–83; Henderson, \textit{supra} note 332, at 404–05.
\textsuperscript{335} See Abraham & Weiler, \textit{supra} note 28, at 415; Weiler, \textit{supra} note 104, at 226 (urging states to experiment with enterprise liability).
\textsuperscript{336} See Furrow, \textit{supra} note 20, at 112.
\textsuperscript{337} See \textit{id.}
\textsuperscript{338} See \textit{id.}
\textsuperscript{339} See \textit{id.}
\textsuperscript{340} See Skocpol, \textit{supra} note 56, at 178–79 (referring to the political volatility of a complex, expensive, and bureaucratic system).
2. No-fault

A no-fault system may have significant cultural appeal because it would be more accepting of the fact that doctors make mistakes, and the system would nurture forthrightness in the doctor-patient relationship. Americans care a great deal about their healthcare. This is reflected by how much Americans spend on healthcare. The doctor-patient relationship is the cornerstone of healthcare provision, and doctors and patients might both be well-served if the relationship did not become so adversarial when invoking the law.

Despite a no-fault system’s cultural appeal as a more accepting and generous way of dealing with medical error, a no-fault system would face serious political and cultural roadblocks. A no-fault system would be more bureaucratic and administrative—adding another layer of paperwork to an already complex system. This added bureaucracy may also prompt cries of socialization, which may make a no-fault system politically unrealistic.

Perhaps more critical than its bureaucratic and socialistic characteristics, a no-fault system may not have an organized interest group to champion it. Doctors seem convinced that they want damage caps and similar medical-malpractice reforms. Plaintiffs’ attorneys would probably oppose a no-fault system because they would be largely cut out of the claims process. Like the AMA, the Association of Trial Lawyers of America (“ATLA”) is a powerful lobbying force. Attorneys who bring medical-malpractice claims probably stand to lose the

341 See Furrow, supra note 20, at 122–23 (making the same argument with respect to enterprise liability); Studdert & Brennan, supra note 30, at 227–28.
342 See generally Gawande, supra note 193 (illustrating the human side of medicine).
343 See Nat’l Ctr. for Health Statistics, supra note 25, at 14.
345 See Kinney, supra note 269, at 123–25; Kupeli, supra note 214, at 569–70; Chandler Gregg, Comment, The Medical Malpractice Crisis: A Problem with No Answer?, 70 Mo. L. Rev. 307, 311 (2005) (stating that reforms like no-fault have received little support).
346 See Weiler, supra note 10, at 931–32.
347 See Skocpol, supra note 56, at 174 (discussing “Reagan’s Revenge” and Americans’ distaste for big government and bureaucracy).
348 See Abraham & Weiler, supra note 28, at 383; Kinney, supra note 269, at 123–25 (stating that physicians are skeptical of no-fault, consumer groups are silent, and the trial bar prefers the status quo); Gregg, supra note 345, at 311.
349 See Am. Med. Ass’n, supra note 21, at 23–24.
351 See Abraham & Weiler, supra note 28, at 383.
most if a no-fault system were implemented. Attorneys would no longer be needed for patients to receive compensation for medical errors. Thus, some attorney groups probably would fight a no-fault system aggressively. Given how potent their current opposition is to medical-malpractice reforms like damage caps, a system that largely cuts out attorneys could be expected to generate even more opposition. Without strong political momentum, it is difficult to imagine an extensive reform like no-fault being enacted.

3. Which System is Better?

Because an enterprise liability model more closely resembles the current medical-malpractice system, it would alienate fewer interest groups if implemented and be more politically feasible. A no-fault system has the cultural misfortune of being an inherently more bureaucratic and administrative approach to compensating injured patients. It would thus be vulnerable to claims that it would further clog the healthcare system with paperwork and complicated rules. The lawsuit-based enterprise liability system would not be vulnerable to these kinds of attacks.

No-fault’s greatest political weakness is that it would not have a powerful interest group to champion it. An enterprise liability system would not alienate attorneys nearly as much. In fact, attorneys may even profit from an enterprise liability system. Hospitals have

352 See Brill, supra note 213, at 1005.
353 See Bovbjerg & Sloan, supra note 126, at 73; Brill, supra note 213, at 1005;
354 See Bovbjerg & Sloan, supra note 126, at 73; Brill, supra note 213, at 1005; Ass’n of Trial Lawyers of Am., Health Care Resource Center, http://www.atlanet.org/ConsumerMediaResources/Tier3/press_room/FACTS/health/index.aspx (including links to articles such as The Truth About Medical Malpractice in America, Debunking the Top 5 Myths About Medical Malpractice, and Ten Reasons To Oppose Medical Malpractice ‘Reform’).
355 See Kinney, supra note 269, at 124–25; Ass’n of Trial Lawyers of Am., supra note 354.
356 See Skocpol, supra note 56, at 6–8, 174; Kinney, supra note 269, at 225; Kupeli, supra note 214, at 569–70.
357 See Skocpol, supra note 56, at 6–8, 174; Abraham & Weiler, supra note 28, at 432; Kupeli, supra note 214, at 569–70; Ass’n of Trial Lawyers of Am., supra note 354.
358 See Skocpol, supra note 56, at 174; Studdert et al., supra note 31, at 29–31.
360 See generally Ass’n of Trial Lawyers of Am., supra note 26.
362 See generally Ass’n of Trial Lawyers of Am., supra note 354.
363 See Abraham & Weiler, supra note 28, at 403–06 (discussing improvements in patient compensation).
deep pockets and would probably settle legitimate claims more often than doctors would.\footnote{364}{See id.}

Although it would probably be more difficult to enact a no-fault system, it is worth noting that enterprise liability already fizzled as a reform during the Clinton administration.\footnote{365}{See id.} Enterprise liability faced opposition from the AMA and the PIAA.\footnote{366}{Id.} Doctors may prefer a no-fault system to enterprise liability because it eliminates the fault requirement and its culture of blame.\footnote{367}{See Weiler, supra note 10, at 913.} If this difference is sufficiently enticing to doctors, the AMA may support no-fault reforms.\footnote{368}{See Am. Med. Ass’n, supra note 21, at 4–8; Abraham & Weiler, supra note 28, at 383; Weiler, supra note 10, at 913.}

It is probably more likely that the AMA would oppose both enterprise liability and no-fault because doctors are unwilling to submit themselves to additional control from hospitals.\footnote{369}{See Am. Med. Ass’n, supra note 21, at 23–45.} The AMA will continue to lobby for damage caps and similar reforms.\footnote{370}{See Lohr, supra note 21, at BU1.} The AMA will not feel pressured to compromise and support alternative reforms like enterprise liability or no-fault because the Bush administration and Republican Congressional majorities support damage caps and similar reforms promoted by the AMA.\footnote{371}{See e.g., Skocpol, supra note 56, at 174; Kupeli, supra note 214, at 569–70.}

Enterprise liability and no-fault both face daunting political challenges.\footnote{372}{See Brill, supra note 213, at 1005; Ass’n of Trial Lawyers of Am., supra note 26.} Enterprise liability is probably more feasible because it does not alienate attorneys.\footnote{373}{See Brill, supra note 213, at 1005; Kinney, supra note 269, at 123 (stating that physicians feel enormous personal responsibility for their patients, contributing to their discomfort with no-fault).} No-fault may have no champions if both physician and plaintiff attorneys’ lobbies oppose it.\footnote{374}{See Brill, supra note 213, at 1005; Ass’n of Trial Lawyers of Am., supra note 26.} Without a determined and well-financed lobby, it is hard to imagine a dramatic change like no-fault overcoming political opposition and realizing the improvements it was designed to create.\footnote{375}{See Brill, supra note 213, at 1005; Kinney, supra note 269, at 225; Ass’n of Trial Lawyers of Am., supra note 26.}
F. Aggregate Weighing: Which Model Is Better?

Having evaluated enterprise liability and no-fault in terms of how well each satisfies the goals of compensation and deterrence and meets the requirements of economic and sociopolitical feasibility, the question becomes which is the better reform.\footnote{See supra notes 196–375 and accompanying text.} No-fault offers greater promise that it will serve the goals of compensation and deterrence, but enterprise liability is more feasible both economically and in the current sociopolitical context.\footnote{See supra notes 196–375 and accompanying text.}

Although no-fault is a promising reform because it has so much potential to improve compensation and deterrence, enterprise liability is a better reform because it is more feasible than no-fault.\footnote{See supra notes 196–375 and accompanying text.} The theoretical virtues of no-fault cannot do good if the system cannot be enacted.\footnote{See Studdert et al., supra note 31, at 3.} Because enterprise liability does promise to streamline compensation and significantly improve patient safety, it is a valuable reform that could be enacted.\footnote{See id. at 434.}

Enterprise liability also has the advantage of being a potential step toward no-fault liability.\footnote{See Abraham & Weiler, supra note 28, at 434.} Enterprise liability could be an end in itself or a stepping-stone for a no-fault system.\footnote{See id.} Because enterprise liability is a more feasible reform, it should be implemented first.\footnote{See id.} Enterprise liability is a more moderate reform that leaves open the possibility for a more radical change like no-fault.\footnote{See id.} At this time, no-fault has many theoretical virtues, but its implementation remains unlikely.\footnote{See Skocpol, supra note 56, at 6–8, 174; Abraham & Weiler, supra note 28, at 434; Kupeli, supra note 214, at 569–70.} Enterprise liability also has great promise for improving healthcare and is a more realistic option for reform.\footnote{See Abraham & Weiler, supra note 28, at 434–36.}

**Conclusion**

Enterprise liability is the better reform for the medical-malpractice system because it is more economically, socially, and po-
politically feasible than a no-fault system. An enterprise liability system also promises to serve the goals of compensation and deterrence better than the current fault based system that targets individual providers. Although a no-fault system may better compensate patients and deter malpractice, an enterprise liability system would be less costly and easier to implement. An enterprise liability system could be a step toward a no-fault system if experience indicates no-fault would further improve medical-malpractice law and the healthcare system.

Kristie Tappan