ISSN 1936-5349 (print)
ISSN 1936-5357 (online)

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS

MULTISTAGE ADJUDICATION

Louis Kaplow

Discussion Paper No. 732

09/2012

Harvard Law School
Cambridge, MA 02138

This paper can be downloaded without charge from:

The Harvard John M. Olin Discussion Paper Series:
http://www.law.harvard.edu/programs/olin_center/

The Social Science Research Network Electronic Paper Collection:
http://ssrn.com/

Multistage Adjudication

Louis Kaplow*

Abstract

Legal proceedings often involve multiple stages: U.S. civil litigation allows motions to dismiss and for summary judgment before reaching a trial; government agencies as well as prosecutors employ investigative and screening processes before initiating formal adjudication; and many Continental tribunals move forward sequentially. Decisionmaking criteria have proved controversial, as indicated by reactions to the Supreme Court's recent decisions in *Twombly* and *Iqbal* and its 1986 summary judgment trilogy, which together implicate the four Supreme Court cases most cited by federal courts. Neither jurists nor commentators have articulated coherent, noncircular legal standards, and no attempt has been made to examine systematically how decisions at different procedural stages should ideally be made in light of the legal system's objectives. This Article presents a foundational analysis of the subject. The investigation illuminates central elements of legal system design, recasts existing debates about decision standards, identifies pathways for reform, and provides new perspectives on the nature of facts and evidence and on the relationship between substantive and procedural law.

JEL Classes: D81, K14, K41, K42

Keywords: litigation, law enforcement, adjudication, courts, motion to dismiss, summary judgment, screening, burden of proof

Forthcoming, Harvard Law Review (2013)

© Louis Kaplow. All rights reserved.

---

*Finn M.W. Caspersen and Household International Professor of Law and Economics, Harvard Law School, and Research Associate, National Bureau of Economic Research. I am grateful to Gary Born, Christopher Drahozal, Jesse Fried, Daniel Klerman, David Rosenberg, William Rubenstein, Steven Shavell, Holger Spamann, Kathryn Spier, Thomas Stipanowich, Stephen Ware, and participants in workshops at Harvard Law School for helpful discussions and comments; Ronnie Anguas, Daniel Marcet, Andrew Meiser, Stephen Pezzi, Silviu Pitis, Houston Shaner, and Dorothy Shapiro for research assistance; and Harvard Law School's John M. Olin Center for Law, Economics, and Business for financial support. Disclaimer: subsequent to completing most work on this Article, I served as a consultant to a plaintiff opposing a motion to dismiss, and my wife is in the legal department of a financial services firm.

TABLE OF CONTENTS

I. INTRODUCTION

In widely ranging settings and across jurisdictions, legal proceedings do not commence at the beginning of a trial. Instead, they involve multiple stages, both informal and formal. At various points along the path, a decision is made whether to terminate a case or allow it to continue to the next stage, the latter outcome ordinarily involving additional expenditures that generate further information. If the case proceeds at each decision node, it enters a final stage at which there is a judgment on liability. In some systems, liability may also be assigned earlier, short-circuiting subsequent proceedings.

This sort of process is familiar in U.S. civil litigation, where a motion to dismiss or for summary judgment may be granted pretrial. In criminal cases, a grand jury indictment or some substitute comes before trial. But multistage proceedings are more widespread, particularly when informal steps are included. Investigations, whether undertaken by the police or government agencies, often incorporate initial screening and interim assessments to decide whether to cease or undertake further efforts. These determinations may be made by those performing the work, or they may be hierarchical. Multistage decisionmaking is likewise employed in myriad nonlegal settings, from business (whether to launch new product lines, reorganize operations, or close factories) to medicine (whether to undertake a course of treatment) to everyday life (whether to change jobs, purchase a home, or switch cell phone providers).

Multistage legal procedures' importance is evident from U.S. federal court citation practice. The three most-cited Supreme Court decisions are the 1986 trilogy on summary judgment: *Anderson v. Liberty Lobby, Inc.*,[1] *Celotex Corp. v. Catrett*,[2] and *Matsushita Elec. Indus. Co. v. Zenith Radio Corp.*[3] Closely behind in fourth place is *Conley v. Gibson*,[4] which was the leading case on motions to dismiss until its oft-quoted language was "retire[d]" in the Supreme Court's 2007 decision in *Bell Atlantic Corp. v. Twombly*.[5] Moreover, *Twombly* and *Ashcroft v. Iqbal*,[6] decided in 2009, generated intense reactions, including proposals for congressional override.[7] These two decisions are viewed as among the more important of the Roberts Court, with the prospect of greatly changing federal litigation in important areas of law.

---

[1] 477 U.S. 242 (1986).

[2] 477 U.S. 317 (1986).

[3] 475 U.S. 574 (1986). *See* Adam N. Steinman, *The Irrepressible Myth of* Celotex*: Reconsidering Summary Judgment Burdens Twenty Years After the Trilogy*, 63 WASH. & LEE L. REV. 81, 143 tbl. 1 (2006) (showing that *Anderson* and *Celotex* each have over 70,000 citations by federal courts and tribunals through June 29, 2005, and *Matsushita* has over 30,000). Additionally, an earlier summary judgment decision, *Adickes v. S. H. Kress & Co.*, 398 U.S. 144 (1970), which is largely superseded by the trilogy, holds the tenth position.

[4] 355 U.S. 41 (1957). *See* Steinman, *supra* note 3, at 143 tbl. 1.

[5] 550 U.S. 544, 563 (2007).

[6] 129 S. Ct. 1937 (2009).

[7] *See, e.g.*, Robert G. Bone, *Plausibility Pleading Revisited and Revised: A Comment on* Ashcroft v. Iqbal, 85 NOTRE DAME L. REV. 849, 850 (2010) (noting proposal of The Notice Pleading Restoration Act, S. 1504, 111th Cong. (2009), and The Open Access to Courts Act, H.R. 4115, 111th Cong. (2009)); Michael R. Huston, Note, *Pleading With Congress to Resist the Urge to Overrule* Twombly *and* Iqbal, 109 MICH. L. REV. 415 (2010) .

For criminal cases in the United States, the right to a grand jury for capital and infamous crimes is protected by the Fifth Amendment[8] and has its origin in Magna Carta.

These procedures pose great challenges, both doctrinally and normatively. *Twombly* and *Iqbal* establish a "plausibility" standard for motions to dismiss.[9] Yet controversy surrounds what this criterion involves,[10] particularly because the Supreme Court seemed to reject interpretations grounded either in logic or in probabilities.[11] The cases can be viewed as addressing a dilemma.[12] On one hand, if conclusory claims are sufficient to survive a motion to dismiss, there may be a flood of groundless suits that threaten to impose high costs — monetary and otherwise — on blameless defendants.[13] On the other hand, if plaintiffs must already possess information

---

[8]U.S. CONST., amend. V.

[9]*Twombly*, 550 U.S. at 556, 559, 566, 569–70; *Iqbal*, 129 S. Ct. at 1949. It is unclear the extent to which this newly established standard will influence outcomes of motions to dismiss, and one reason there may be limited impact is that previous practice may already have reflected the standard promulgated in *Twombly* and *Iqbal*. *See* sources cited *infra* note 176. For preliminary evidence, see Kendall W. Hannon, *Much Ado About* Twombly*? A Study on the Impact of* Bell Atlantic Corp. v. Twombly *on 12(b)(6) Motions*, 83 NOTRE DAME L. REV. 1811, 1836–37 (2008) (finding that a comparison of outcomes before and after *Twombly* reveals essentially no measurable impact except on civil rights litigation; with those cases removed, there was almost no effect in the large sample — for example, motions to dismiss were granted in 37.4% of cases after, compared to 36.9% before *Twombly*); Patricia W. Hatamyar, *The Tao of Pleading: Do* Twombly *and* Iqbal *Matter Empirically?*, 59 AM. U. L. REV. 553, 616–24 (2010) (reporting regression results that are mixed and statistically insignificant, including a fall in grants of motions to dismiss without leave to amend after *Twombly* and *Iqbal*, compared to before *Twombly*); Jonah B. Gelbach, Note, *Locking the Doors to Discovery? Assessing the Effects of* Twombly *and* Iqbal *on Access to Discovery*, 121 YALE L.J. 2270 (2012) (finding that *Twombly* and *Iqbal* adversely affected plaintiffs in 15% to 21% of cases facing motions to dismiss); U.S. COURTS, MOTIONS TO DISMISS: INFORMATION ON COLLECTION OF DATA, *available at* http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/Motions_to_Dismiss_060110.pdf (last visited Mar. 27, 2012) (displaying monthly data on motions to dismiss filed, granted, and denied from four months before *Twombly* to twelve months after *Iqbal*, with no evident overall trend or breakpoints at the time of either decision).

[10]*See, e.g.*, Bone, *supra* note 7; Kevin M. Clermont & Stephen C. Yeazell, *Inventing Tests, Destabilizing Systems*, 95 IOWA L. REV. 821 (2010); Adam N. Steinman, *The Pleading Problem*, 62 STAN. L. REV. 1293 (2010); Daniel W. Robertson, Note, *In Defense of Plausibility:* Ashcroft v. Iqbal *and What the Plausibility Standard Really Means*, 38 PEPP. L. REV. 111 (2010); Nicholas Tymoczko, Note, *Between the Possible and the Probable: Defining the Plausibility Standard After* Bell Atlantic Corp. v. Twombly *and* Ashcroft v. Iqbal, 94 MINN. L. REV. 505 (2009).

[11]The Court granted the logical sensibility of the plaintiffs' claims in those cases; if plausibility had a purely logical interpretation, both outcomes would have been different. And the Court expressly repudiated a probability standard. *See Twombly*, 550 U.S. at 556 ("Asking for plausible grounds to infer an agreement does not impose a probability requirement at the pleading stage . . . ."); *Iqbal*, 129 S. Ct. at 1949 ("The plausibility standard is not akin to a 'probability requirement' . . . ."). The plausibility standard is examined in detail in section IV.A.

[12]*See, e.g.*, Robert G. Bone, *A Proceduralist's Perspective on Court Access After* Twombly, GCP: ONLINE MAG. FOR GLOBAL COMPETITION POL'Y, July 2009, Release Two, at 1, 5–6, *available at* https://www.competitionpolicyinternational.com/file/view/6069. As will be developed throughout the Article, this statement of the problem is actually incomplete, in some respects can be misleading, and fails to reveal the true magnitude of the difficulty.

[13]*See, e.g.*, *Twombly*, 550 U.S. at 560 n.6 ("But determining whether some illegal agreement may have taken place between unspecified persons at different ILECs (each a multibillion dollar corporation with legions of management level employees) at some point over seven years is a sprawling, costly, and hugely time-consuming undertaking not easily susceptible to the kind of line drawing and case management that the dissent envisions."); *id*. at 558–60 & n.6 (mentioning the "in terrorem increment of the settlement value," the "inevitably costly and protracted discovery phase," the problem of "allowing a potentially massive factual controversy to proceed," that "discovery accounts for as much as 90 percent of litigation costs when discovery is actively employed," and that the "threat of discovery expense will push cost-conscious defendants to settle even anemic cases"); RICHARD EPSTEIN, AEI-BROOKINGS JOINT CENTER FOR REGULATORY STUDIES, MOTIONS TO DISMISS ANTITRUST CASES: SEPARATING FACT FROM FANTASY (2006).

that adequately inculpates defendants as a prerequisite to discovery — a main purpose of which is go give plaintiffs access to information solely in defendants' possession, posing a classic catch-22 — important groups of meritorious cases may be eliminated from the system.[14]

Ready answers, even ways of thinking cogently about some of the tradeoffs, have been lacking. A reader of these two Supreme Court opinions might be excused from believing that the majority and dissent in both cases engaged in wishful thinking in downplaying the competing half of the dilemma.[15] Some commentators have proposed glosses on or substitutes for the Court's plausibility test,[16] but it is hard to see how recasting the test's language can avoid or even diminish the real-world conflict. Additionally, much of the discussion seems aimed at how plaintiffs' allegations must be worded, a focus that is appropriate for providing notice to defendants but one that seems inapt regarding the substantive standard because the impossibility of knowing the unknown cannot be surmounted by artful drafting.[17]

The test for summary judgment, although less controversial at the moment, is itself quite murky.[18] Rule 56 of the Federal Rules of Civil Procedure asks whether there is a "genuine

---

[14]*See, e.g.*, Arthur R. Miller, *From* Conley *to* Twombly *to* Iqbal*: A Double Play on the Federal Rules of Civil Procedure*, 60 DUKE L.J. 1 (2010); Rakesh N. Kilaru, Comment, *The New Rule 12(B)(6):* Twombly*,* Iqbal*, and the Paradox of Pleading*, 62 STAN. L. REV. 905 (2010); *cf.* Arthur R. Miller, *The Pretrial Rush to Judgment: Are the 'Litigation Explosion,' 'Liability Crisis,' and Efficiency Clichés Eroding Our Day in Court and Jury Trial Commitments?*, 78 N.Y.U. L. REV. 982 (2003).

[15]For example, in neither case did the majority indicate how, if the plaintiff's claim were valid, is could have obtained the sort of information needed to make the requisite allegations. Nor did the dissenters in *Twombly* credibly explain how plaintiffs with an entirely groundless conspiracy allegation could be prevented, once into discovery, from costing defendants tens of millions of dollars and taking numerous depositions of key executives, thereby disrupting their businesses. Nor did the dissenters in *Iqbal* credibly show how myriad individuals who might purport to suffer discrimination would be precluded from taking depositions at the highest levels of government concerning alleged practices asserted to have been dictated or condoned from the top.

[16]*See, e.g.*, Robert G. Bone, Twombly*, Pleading Rules, and the Regulation of Court Access*, 94 IOWA L. REV. 873 (2009); Stephen R. Brown, *Correlation Plausibility: A Framework for Fairness and Predictability in Pleading Practice After* Twombly *and* Iqbal, 44 CREIGHTON L. REV. 141 (2010); Steinman, *supra* note 10. Additional writing addresses the implications of these recent cases in particular substantive legal fields, especially antitrust. *See, e.g.*, Richard A. Epstein, *Of Pleading and Discovery: Reflections on* Twombly *and* Iqbal *with Special Reference to Antitrust*, 2011 U. ILL. L. REV. 187; Herbert Hovenkamp, *The Pleading Problem in Antitrust Cases and Beyond*, 95 IOWA L. REV. BULL. 55 (2010); *see also* Brook Detterman, *Rumors of* Conley's *Demise Have Been Greatly Exaggerated: The Impact of* Bell Atlantic Corp. v. Twombly *on Pleading Standards in Environmental Litigation*, 40 ENVTL. L. 295 (2010) (claiming, surprisingly, especially since the article is published after *Iqbal*, that *Twombly* is inapplicable in the environmental context).

[17]Implicitly, it seems to be assumed that, if cases have merit, plaintiffs indeed will have sufficiently detailed information that can be described in their complaint, whereas if cases are meritless, plaintiffs neither will have sufficient (though misleading, perhaps taken out of context) fragments to describe nor will they be able to purport to have them. Note that, if this state of affairs prevailed, *Twombly* and *Iqbal* would really be about notice, which no one seems to assume. *See infra* notes 177 & 197.

[18]*See, e.g.*, 10A CHARLES ALAN WRIGHT, ARTHUR R. MILLER & MARY KAY KANE, FEDERAL PRACTICE AND PROCEDURE § 2725 (2011) ("[T]here is no established standard governing the question of what constitutes a material fact. . . . To determine the existence of a genuine issue, the court must engage in a rather sophisticated and careful inquiry, especially since there is no precise formula for determining when this is the case."); Edward H. Cooper, *Directions for Directed Verdicts: A Compass for Federal Courts*, 55 MINN. L. REV. 903, 903 (1971) (describing the standard for directed verdict as follows: "Exercise of this control in an individual case inherently involves a large measure of careful judgment, tailoring general principles of deference to jury freedom to the unique facts before the court. Many of the

dispute,"[19] which *Anderson* held to be the same as the standard under Rule 50 for judgment as a matter of law.[20] But what does that test require? The question-begging answer contained in the rule itself is whether "a reasonable jury would . . . have a legally sufficient evidentiary basis to find for the party on that issue."[21] Setting aside the considerable ambiguity created by the phrase "reasonable jury,"[22] this test demands that a case should be given to the factfinder (or that the factfinder's decision should be sustained) if and only if there is a legally sufficient basis for it to reach the judgment in question.[23] In all, it seems that dispositive motions in U.S. civil courts must be decided using rather ambiguous, open-ended criteria,[24] where little guidance has been offered regarding what either test's actual content is — or what that content should be.

These familiar formal rules governing civil litigation in U.S. courts may not be the most significant overall. In many areas of federal law, from antitrust to environmental regulation to OSHA, and with countless counterparts in state and local government, critical decisions are routinely made at early and interim stages within administrative agencies. For example, whether the Department of Justice or Federal Trade Commission, when reviewing a merger, chooses to make a second request for information and ultimately decides to file a challenge is often decisive, for the parties may well give up the fight before even a preliminary decision by a court.[25] In addition, interim decisions by police and prosecutors whether to pursue or drop

---

various and frequently confusing phrases used in an attempt to establish guiding standards provide no more help than this general statement. For many of the problems, nothing more can be said."). Despite this recognition by some commentators, there has been surprisingly little effort, particularly in recent years, devoted to explicating the test. Instead, scholarship has tended to focus on other questions, such as what defendants must show when moving for summary judgment in order to require some response from the plaintiff, an issue presented in *Celotex*. *See, e.g.*, David L. Shapiro, *The Story of* Celotex*: The Role of Summary Judgment in the Administration of Civil Justice*, *in* CIVIL PROCEDURE STORIES 359 (Kevin M. Clermont ed., 2nd ed. 2008); Steinman, *supra* note 3.

[19]FED. R. CIV. P. 56(a). This phrase replaced the familiar term "genuine issue" in the 2010 amendments. The Advisory Committee indicates that no change in meaning was intended. *See id*., advisory committee's note.

[20]Anderson v. Liberty Lobby, Inc., 477 U.S. 242, 249–50 (1986). At the time of *Anderson*, the disposition covered by Rule 50 was referred to as a directed verdict, which language was changed in the 1991 amendments to the Federal Rules of Civil Procedure without any intended change in the standard. *See* FED. R. CIV. P. 50(a)(1), advisory committee's note.

[21]FED. R. CIV. P. 50(a)(1).

[22]An illustration of the potential scope offered by this term for judges to grant summary judgment despite potential disputes about the facts is offered by the Supreme Court's decision in *Scott v. Harris*, 550 U.S. 372 (2007), where the members of the Court viewed a video and concluded, with one dissent, that summary judgment should be granted; others' reactions to the video were more heterogeneous. *See* 11 JEFFREY W. STEMPEL & STEVEN S. GENSLER, MOORE'S FEDERAL PRACTICE § 56.06 (3d ed. 2012) [hereinafter MOORE'S FEDERAL PRACTICE]; Dan M. Kahan, David A. Hoffman & Donald Braman, *Whose Eyes Are You Going to Believe?* Scott v. Harris *and the Perils of Cognitive Illiberalism*, 122 HARV. L. REV. 837 (2009).

[23]The discussion of the standard in MOORE'S FEDERAL PRACTICE, *supra* note 22, § 56.22, further illustrates the problem. On one hand, the treatise observes: "Courts tend to state that a factual dispute is 'genuine' if the dispute is one that requires a trial for resolution. Without more, that formulation is tautological." They proceed to clarify: "What the courts really mean is that a fact dispute is 'genuine' when the record includes evidence that would permit a reasonable jury to find for the nonmoving party." But their attempted clarification merely restates the problem of how much evidence is sufficient to permit the finding in question. For further elaboration, see section IV.E.

[24]*See also infra* note 282 (quoting Cooper's views on the standard under Rule 50).

[25]*See, e.g.*, Malcolm B. Coate, Andrew N. Kleit & Rene Bustamante, *Fight, Fold or Settle? Modelling the Reaction to FTC Merger Challenges*, 33 ECON. INQUIRY 537 (1995). More broadly, a central question in antitrust writing is how agencies should perform initial screening and prioritize industries and activities for investigation and subsequent

investigations are momentous: continuation can ruin lives and destroy entities' value even when cases are meritless, and termination of valid cases undermines deterrence.

It is also notable that, with regard to staging, administrative and court procedures vary greatly within and across jurisdictions. Different agencies adopt different approaches; civil litigation differs from criminal litigation, both of which differ from similar activities within agencies and investigative entities; and U.S. states often differ from each other and from the federal government. Common law jurisdictions are hardly homogeneous (for example, the United Kingdom has replaced the grand jury,[26] and the United States employs a distinctive discovery process). To a varying extent, Continental legal systems[27] in civil cases follow a more sequential process for developing evidence, although they tend to lack formal interim termination points like motions to dismiss and for summary judgment.[28] This great diversity in approaches provides further impetus to the present investigation.[29]

In sum, multistage legal procedures are ubiquitous, vary tremendously across legal systems, and constitute one of the most important institutional features of adjudication. In this light, it is surprising that scholars have almost completely ignored the question of how preliminary or interim legal decisions ought to be made.[30] Nor has there been much attention to the optimal structure of multistage adjudication: when to have distinct stages, how many, what issues and evidence to consider at each, and in what order. This gap exists despite massive legal

---

action. *See, e.g.*, Rosa Abrantes-Metz & Patrick Bajari, *Screens for Conspiracies and Their Multiple Applications*, ANTITRUST, Fall 2009, at 66; Rosa M. Abrantes-Metz & Luke M. Froeb, *Competition Agencies Are Screening for Conspiracies: What Are They Likely to Find?*, ECONOMICS COMMITTEE NEWSL. (ABA Section of Antitrust Law), Spring 2008, at 10; Rosa M. Abrantes-Metz, Luke M. Froeb, John F. Geweke & Christopher T. Taylor, *A Variance Screen for Collusion*, 24 INT'L J. INDUS. ORG. 467 (2006); Joseph E. Harrington, Jr., *Behavioral Screening and the Detection of Cartels*, *in* EUROPEAN COMPETITION LAW ANNUAL: 2006 — ENFORCEMENT OF PROHIBITION OF CARTELS 51 (Claus-Dieter Ehlermann & Isabela Atanasiu eds., 2007). Nevertheless, that literature does not systematically consider the questions examined in the present Article.

[26]Administration of Justice (Miscellaneous Provisions) Act 1933; *see* Nathan T. Elliff, *Notes on the Abolition of the English Grand Jury*, 29 J. CRIM. L. & CRIMINOLOGY 3 (1938).

[27]Throughout, I use the terminology of Continental legal systems rather than Civil Law systems because often (including here), the term "civil" is being used to distinguish between civil and criminal cases.

[28]*See, e.g.*, KONRAD ZWEIGERT & HEIN KÖTZ, INTRODUCTION TO COMPARATIVE LAW 271–75 (Tony Weir trans., 3d ed. 1998); Hein Kötz, *Civil Justice Systems in Europe and the United States*, 13 DUKE J. COMP. & INT'L L. 61 (2003); Rolf Stürner, *Inaugural Speech: Procedural Law and Legal Cultures — Introduction to the Overarching Topic of the Conference*, *in* PROZEßRECHT UND RECHTSKULTUREN: PROCEDURAL LAW AND LEGAL CULTURES 9 (2004); Rolf Stürner, *The Principles of Transnational Civil Procedure: An Introduction to Their Basic Conceptions*, *in* RABELS ZEITSCHRIFT FÜR AUSLÄNDISCHES UND INTERNATIONALES PRIVATRECHT 201 (Rolf Stürner, Stephan R. Göthel & Herbert Küpper eds., 2005) [hereinafter Stürner, *Transnational Civil Procedure*].

[29]To this long list, one should add alternative dispute resolution, which is of particular interest because parties and providers have great flexibility in system design. *See infra* note 140.

[30]Even regarding final adjudication, the question of how best to make decisions has also been largely neglected until quite recently. *See, e.g.,* Louis Kaplow, *Burden of Proof*, 121 YALE L.J. 738 (2012).

writing on procedure,[31] including recent decades' work in law and economics.[32]

This Article analyzes these questions, with an emphasis on the first: how decisions are optimally made at each stage, taking the structure of the rest of the legal system as given. In some respects, the results are quite general and have broad application, for they pertain to any sort of system with any number of stages. Hence, the implications are relevant not only to motions to dismiss and summary judgment (as well as the burden of proof at trial) in U.S. civil litigation — which will be highlighted in Part IV — but also to indictments in criminal settings, all manner of agency proceedings, interim decisionmaking by police or other investigative bodies, some features of Continental legal systems, and alternative dispute resolution. In other respects, this investigation has important limitations. Preliminary assessments of complex regimes are inevitably incomplete, the analysis focuses on certain important but specific legal settings, and optimal system design and operation in any context depend on many empirical matters that are heretofore unexplored and would be difficult to assess.

The Article's method is to ask what procedural rules best advance social welfare, wherein the two central considerations are the legal system's effects on behavior — deterrence of harmful conduct and the chilling of desirable activity — and total system costs.[33] Until Part IV, the analysis proceeds in pure form, setting aside its applicability under prevailing rules and institutional constraints, which obviously vary greatly across legal systems and also could potentially be reformed. Nevertheless, the approach seems, prima facie, to have relevance in many existing settings.

---

[31]For example, CHARLES ALAN WRIGHT & ARTHUR R. MILLER, FEDERAL PRACTICE AND PROCEDURE (2012), runs eighty-three volumes, and MOORE'S FEDERAL PRACTICE (3d ed. 2012) is thirty-three volumes, yet neither considers these questions nor directs the reader to significant treatments of the issues.

[32]The most likely suspect would be Richard Posner, who has written most broadly, including on the economic analysis of procedure. The latest edition of his famous one-thousand-page text includes, for the first time, a single page on motions to dismiss and summary judgment, and, quite uncharacteristically, confines itself to a statement of the law, eschewing any analysis. RICHARD A. POSNER, ECONOMIC ANALYSIS OF LAW 762 (8th ed. 2011). His book on the federal courts contains a somewhat lengthier discussion of dispositive motions, but limits its attention to lower-court practice, with emphasis on the granting of such motions as a means of workload reduction. RICHARD A. POSNER, THE FEDERAL COURTS: CHALLENGE AND REFORM 175–83 (1996) [hereinafter POSNER, FEDERAL COURTS]. Most on-point for present purposes is a student primer on procedure by Robert Bone, although this intentionally elementary presentation focuses on the use of decision analysis to determine expected outcomes and sets aside central considerations, such as how the legal system influences behavior. ROBERT G. BONE, CIVIL PROCEDURE: THE ECONOMICS OF CIVIL PROCEDURE 125–57 (2003). *See also* Keith N. Hylton, *When Should a Case Be Dismissed? The Economics of Pleading and Summary Judgment Standards*, 16 SUP. CT. ECON. REV. 39 (2008) (exploring some of the same questions but not incorporating many of the key elements examined here or performing a complete analysis and hence having little overlap in either the actual substance or the conclusions reached). More broadly, the questions considered in this Article are related to those examined in the literature on accuracy in adjudication, which has analyzed them in simpler models that mainly address different issues. *See* sources cited *infra* note 146. For a survey of law and economics literature on litigation generally, a few elements of which (cited at pertinent points below) touch on pieces of the questions examined here, see Kathryn E. Spier, *Litigation*, *in* 1 HANDBOOK OF LAW AND ECONOMICS 259, 305-07 (A. Mitchell Polinsky & Steven Shavell eds., 2007).

[33]That the objective should be the advancement of social welfare — which refers to the well-being of all members of society — is articulated and defended in general terms in Louis Kaplow & Steven Shavell, *Fairness Versus Welfare*, 114 HARV. L. REV. 961 (2002), and specifically with regard to legal procedure in *id*. at 1164–225 and Louis Kaplow, *The Value of Accuracy in Adjudication: An Economic Analysis*, 23 J. LEGAL STUD. 307, 382–99 (1994).

For example, Rule 1 of the Federal Rules of Civil Procedure commands that its rules "should be construed and administered to secure the just, speedy, and inexpensive determination of every action and proceeding."[34] Just determinations involve assigning liability in meritorious cases, those in which defendants actually committed harmful acts, and no liability in unmeritorious cases — objectives that loosely correspond to providing deterrence while minimizing chilling effects — and the concern for expense matches the welfare component involving system costs.[35] Additionally, the result derived here that decision rules optimally depend on many facts and circumstances that vary across cases is in rough accord with *Iqbal*'s pronouncement that "[d]etermining whether a complaint states a plausible claim [is] a context-specific task that requires the reviewing court to draw on its judicial experience and common sense."[36] To be clear, no strong claims are made that this Article's welfare-based conclusions are the best interpretation of any particular procedural rule in U.S. civil litigation or anywhere else. Rather, these brief suggestions indicate that the analysis is likely to have implications in important realms along dimensions that prior work has not attempted to illuminate.

Part II contains the core analysis. It begins by examining the first stage in a two-stage system. The choice at the first stage is between termination (an immediate judgment of no liability) and continuation. Continuation is taken to entail additional costs but to generate additional information that is used to reach a final determination of liability at the second stage (which process is taken as given at this point in the analysis).

In any particular scenario, continuation, relative to termination, has one benefit and two costs. The benefit is that, with regard to actually harmful acts, greater deterrence will be provided by the prospect that such cases will proceed to final adjudication, where they will have some chance of giving rise to liability; some deterrence also results because, regardless of the

---

[34]FED. R. CIV. P. 1; *see also* Celotex Corp. v. Catrett, 477 U.S. 317, 327 (1986) ("Summary judgment procedure is properly regarded not as a disfavored procedural shortcut, but rather as an integral part of the Federal Rules as a whole, which are designed to secure the just, speedy and inexpensive determination of every action..."). For similar expressions of a welfare-based view of the purposes of legal procedure, advanced in the context of addressing the strength of the burden of proof, see *In re Winship*, 397 U.S. 358, 370 (1970) (Harlan, J., concurring) ("[T]he choice of the standard for a particular variety of adjudication does . . . reflect a very fundamental assessment of the comparative social costs of erroneous factual determinations."), and *id*. at 371 ("[T]he choice of the standard to be applied in a particular kind of litigation should, in a rational world, reflect an assessment of the comparative social disutility of each [type of error].").

[35]Speed per se is not examined in this Article; nor does it receive much attention from the courts or legal scholars in the present context (or in most others). Focusing solely on outcome objectives — the benefit of finding liability in meritorious cases and the cost of doing so in unmeritorious cases — and costs, Rule 1's formulation does not indicate either how to analyze any of these components or how they should be traded off when they conflict, as they do directly regarding the rules under investigation here. Likewise, courts and commentators have neither elaborated this statement of objectives as a conceptual matter nor explored how it might be operationalized when there are tradeoffs. (It would be wonderful if the rules could be interpreted to achieve perfectly accurate, instantaneous, and cost-free resolutions in every case; short of that, the statement offers limited guidance.) Instead, most normative analysis of procedural rules makes only vague reference to these objectives — treating them as platitudes — or fails to offer any statement of purposes at all.

[36]Ashcroft v. Iqbal, 129 S. Ct. 1937, 1950 (2009); *see* 5B ARTHUR R. MILLER & MARY KAY KANE, FEDERAL PRACTICE AND PROCEDURE § 1357 (3d ed. 2011) ("In the wake of the 2007 decision in *Twombly* and the 2009 decision in *Iqbal*, district judges are now permitted to consider 'judicial experience' and 'common sense' when deciding a Rule 12(b)(6) motion to dismiss."). *See infra* section IV.A.

final outcome, defendants whose cases are continued bear additional litigation costs. The first, obvious social cost of continuation is the adjudication expense that is borne. The second cost, a chilling effect, is the counterpoint to deterrence: for cases in which defendants actually engaged in benign acts, the prospect of continuation faces potential actors with both litigation costs and some chance of being found liable.[37]

The analysis then decomposes each of these three considerations and explores how their underlying determinants influence the optimal first-stage decision. Although many results are expected, some of the findings are complex and counterintuitive. For example, the most direct effect of higher system costs favors termination, but higher costs also influence the magnitude of deterrence and chilling (as explained, individuals contemplating acts expect to face these greater expenses if their cases are continued), and the welfare consequences of deterrence and chilling become more favorable (for chilling, less unfavorable) because each act that is discouraged involves a greater reduction in expected adjudication costs when litigation is more expensive. Another implication of the analysis is that, although information being solely in defendants' possession may favor continuation when such is typically true, it tends to favor termination when it is true in the scenario at hand but not in most other scenarios. More broadly, because there are so many pertinent factors that vary tremendously not only across fields of law but also between individual cases, implementation of an optimal approach requires challenging contextual judgments by legal decisionmakers.

Part II then extends the analysis by considering how best to make initial or interim decisions in a system with three or more stages and how these rules differ from how decisions at the final stage are optimally made.[38] Finally, this Part examines the relationship among optimal decisions across stages. This analysis is not only important for wholesale system design but also is illuminating when, for various institutional or other reasons, the decision rules at some stages are fixed, and perhaps in ways that may not be optimal. For example, if a later-stage rule is too lenient — cases are too readily continued or, at the final stage, liability is too readily imposed — an earlier stage rule may optimally be set more stringently. Interestingly, as will be explained, the opposite need not be true; indeed, too strict a later rule can actually raise the optimal stringency of earlier rules. The analysis also explores whether the optimal toughness of decision rules rises as one moves to later stages (in a system in which each rule may be set freely), as is commonly supposed. For example, with regard to U.S. civil litigation, most assume that the threshold for a plaintiff to survive summary judgment is and should be higher than that to survive a motion to dismiss but lower than that required to prevail at trial. Yet this structure may not be optimal. Among other reasons, more information does not imply that the optimal standard

---

[37]Some readers may notice that the elements of this cost-benefit framework differ qualitatively from those in more familiar problems concerned with the value of information, such as medical decisionmaking. There, treatment outcomes are primarily valued in themselves rather than for their effects on ex ante behavior. *See, e.g.,* HOWARD RAIFFA, DECISION ANALYSIS: INTRODUCTORY LECTURES ON CHOICES UNDER UNCERTAINTY 27–33, 157–87 (1968). As will be explained in section III.E, this more familiar decision setting does correspond to an important subset of legal contexts, those regarding the regulation of future conduct. In addition, subsection III.C.3 will explain how the extension of the analysis to costly sanctions can be taken to incorporate concerns with outcomes per se, such for the mistaken assignment of liability.

[38]The latter question pertains to the burden of proof, and the discussion in section II.C on the final stage relates the present analysis to that in Kaplow, *supra* note 30.

for continuation or for liability is higher. (By comparison, is it obvious that one should, on average, be more — or would it be less — inclined to prescribe surgery after additional tests or a second opinion?) In addition, the more stages that have been completed, the more adjudication costs have been sunk, which reduces the expected system costs of further continuation.[39]

Part III examines a number of variations and extensions. It first considers optimal staging: the series of questions noted previously concerning when it is optimal to combine or separate adjacent stages and the optimal ordering of stages. Second, it investigates whether the tradeoffs among deterrence, chilling, and system costs, which vary greatly across areas of law, are best addressed by tailored substantive rules applied through uniform procedures or by customizing legal procedure, and it also comments on whether there is a meaningful conceptual distinction between the two tactics. Next, it examines optimal system design when there are additional enforcement tools that may be adjusted along with the decision threshold at each stage of adjudication. Specifically, one might not only have the options of termination and continuation at nonfinal stages but also that of immediately assigning liability, a choice available at the summary judgment stage in U.S. civil litigation. In addition, with regard to deterrence (and chilling effects and system costs), it is possible to adjust the degree of enforcement effort and the level of sanctions, raising the question of what combinations improve or worsen the tradeoff between deterrence benefits and chilling plus adjudication costs. Also considered is how socially costly sanctions (like imprisonment) may affect the analysis, as well as concerns per se with the mistaken imposition of sanctions. This Part then considers the interaction between decision rules in multistage legal proceedings and the initiation of cases by considering government enforcers' and prospective private plaintiffs' incentives, as well as the possibility of settlement. Finally, this Part explores how the analysis differs when the legal setting involves the regulation of future conduct, where, instead of deterrence and chilling, the concern is with actors' subsequent behavior, such as with merger approvals, licensing, and zoning.

Part IV applies the Article's analysis to legal rules, with an emphasis on those for litigation in U.S. civil cases. As already suggested, no strong position is advanced regarding the best interpretation of existing rules or specific reforms. Nevertheless, there are numerous powerful lessons that, at a minimum, significantly change the way these matters should be understood and alter the appropriate path for research and policy work. Section IV.A examines the rule for granting a motion to dismiss, with an emphasis on *Twombly* and *Iqbal*. It focuses on the "plausibility" test and attempts to make sense of the mysterious suggestion that it is not a probability requirement, drawing on the analysis of Part II. Section IV.B explores a number of foundational matters concerning the nature of facts and evidence, the inattention to which seems partly responsible for courts and commentators' difficulties in intelligibly addressing some of the key issues, especially with regard to the standard for motions to dismiss. Included are the distinction between facts and evidence, the interaction of background facts and case-specific particulars in forming well-grounded beliefs, and the meaning and relevance of the notion that all information may be in defendants' possession.

---

[39]This point is trivially true at the final stage (setting aside appeals, *see infra* note 83, which are fairly low cost relative to discovery and trial), for there are no subsequent costs to be incurred.

Section IV.C addresses the substantial informational challenges posed if decisions at each stage of adjudication are to be made in a manner that advances social welfare rather than on formalistic grounds. Various structural methods of dealing with this demand are briefly considered, as well as the manner in which the existing legal system is likely to adapt — and undoubtedly already is reacting. Section IV.D examines the exercise of judicial discretion, which is inevitably great in a legal regime that permits the sorts of legal decisions considered here. Of interest are what judges can effectively do, how their incentives and proclivities bear on what they are likely to do, and problems of accountability that are raised.

Section IV.E turns to the rule for summary judgment, which, as noted earlier, refers to the highly ambiguous and question-begging formulation of the legal test for judgments as a matter of law. It considers how the analysis in Part II bears on the optimal rule and relates optimal decisionmaking at the summary judgment and motion to dismiss stages, focusing on section II.D's more general analysis of how optimal rules at one stage relate to those employed at another and, specifically, when, why, and to what extent it is optimal for the later decision rule to be tougher.

This Article's aim is conceptual: to derive from first principles how multistage legal proceedings are optimally designed. The method is to focus on basic features of standard settings and to determine the consequences of decisions to continue rather than terminate litigation at various stages. The formulations for optimal rules are complex, subtle, and in some respects surprising, but on reflection can readily be understood in terms of effects on the deterrence of harmful acts, the chilling of benign behavior, and the costs of operating the legal system. Unfortunately, case-specific informational challenges, institutional constraints, and limited systemic empirical knowledge, among other considerations, make it difficult to reason directly and simply from the analytical conclusions to rule interpretations or particular reforms.[40] The present goal is to be informative and provocative, not definitive and prescriptive. It is impossible to make progress without first undertaking the sort of investigation attempted here, the results of which reveal numerous insights and new perspectives on central features of legal system design as well as on current rules and practice that are absent in prior literature if for no other reason than that many of the relevant questions have not been asked.

II. ANALYSIS

A. First Stage in a Two-Stage System

Examination of the first stage in a two-stage system of adjudication — taking as given how decisions are made in the second, final stage, when it is reached — reveals many of this Article's insights, serves as a foundation for analysis in subsequent sections, and keeps the number of moving parts to a minimum, which facilitates understanding. Nevertheless, the formulation for an optimal first-stage decision rule proves to be much more complex than one

[40]In addition, a complete analysis of systemic reform would encompass additional policy instruments, such as heightened sanctions for misrepresentations of facts, fee-shifting, enhanced judicial staffing, and much more.

may have anticipated and, in some respects, embodies counterintuitive results.

1. Setting. — For concreteness, the analysis is conducted in a stylized setting that is limited to some central features of the problem at hand.[41] To begin, consider the activities that the legal system confronts. Acts are taken to be of two types: harmful and benign.[42] Both generate private benefits, the magnitude of which varies among those individuals who might commit them. The former entail harm to others — such as with contract breach, torts, fraud, price-fixing, and patent infringement — whereas the latter do not. An individual with the opportunity to commit one of the types of acts will do so if the private benefit from the act exceeds the resulting expected cost generated by the legal system (on which more in a moment). For simplicity, it is assumed that, in any given situation, an individual has the opportunity to commit only one of these two types of acts.[43]

For those individuals who commit each type of act, a fraction (one hopes, higher for the harmful type) enters the legal system; variations are explored later.[44] For each case that enters the system, some information is initially available. This information concerns two sorts of matters: whether the case involves a harmful or benign act, and also the continuation costs associated with the case at hand (for example, the preliminary information may indicate that the expected cost of engaging in discovery or conducting a trial is unusually high or low).

Different cases will be associated with different information. The term scenario is employed throughout to denote clusters of cases that look the same at the point of the pertinent decision. When individuals contemplate whether to commit an act, they are aware that, should they be brought into the legal system, their cases may present themselves as being in one or another of these possible scenarios, each with some probability. Moreover, because many scenarios will contain both harmful and benign acts — generally with different likelihoods for each — the challenge facing a tribunal will be to make a decision under uncertainty.

Specifically, for each of these scenarios, the decision to be made at stage one is whether to terminate or to continue — here, to final adjudication.[45] Termination of a case removes it

---

[41]For a formal statement, see Kaplow, Optimal Multistage Adjudication (July 2012) (unpublished manuscript).

[42]Regarding the latter, we are interested in those benign acts that might, at least initially, appear similar to harmful acts (or may be thus represented by private plaintiffs or other enforcers). *See also infra* note 60 (on the grouping of acts).

[43]A number of the assumptions employed here have little qualitative effect on the conclusions. For example, one could allow the benign act to impose harm, but at a different level than that caused by the other act, and individuals could be permitted to choose among the two types of acts and inaction (which would introduce a choice between the acts, thereby making deterrence more beneficial and chilling more harmful).

[44]*See infra* note 55 (exploring the difference between enforcement by investigation and enforcement by monitoring or auditing); subsection III.C.2 (assessing the impact of allowing enforcement effort to vary and also attending explicitly to the enforcement costs incurred before the stage-one decision, which are ignored here for simplicity); subsection III.D.1 (examining the incentives of private parties and government enforcers to file cases).

[45]Settlements, considered briefly in subsection III.D.2, are otherwise ignored. For the most part, the language of termination and continuation is employed because these terms are unambiguous and can be used in a variety of settings, including ones in which no motions are made but instead, say, an agency makes an internal decision.

from the system, with no further consequences.[46] Continuation involves additional costs being incurred by the defendant, by the decisionmaking body, and by the enforcer — which may be understood as a private plaintiff, government prosecutor, or an agency. (For convenience, costs incurred by anyone but the defendant will often be referred to collectively as legal system costs.[47]) These costs are assumed to generate additional information, which is employed (along with the initial information) to reach a final judgment at stage two.

A finding of no liability, which also might be described as a stage-two termination, removes the case from the legal system. A finding of liability, which can also be called a stage-two continuation, results in the assignment of a sanction, taken here to be a costless monetary payment.[48] Because information is taken to be imperfect even at the final stage, some who committed harmful acts will be mistakenly exonerated and others who committed benign acts will be mistakenly sanctioned.[49] For present purposes, the decisionmaking algorithm at stage two — whether to assign liability in a given scenario, characterized by the information then available — will be taken as given; analysis of optimal final-stage decisionmaking is deferred to section C.

Return now to the question of actors' ex ante decisions. An individual contemplating either type of act will, as stated, commit the act if and only if the private benefit exceeds the expected cost generated by the legal system. It is now possible to be more explicit about this expected cost. When deciding whether to act, the individual will be unsure about many pertinent factors, conditional on committing the act. First, there is some probability that the act will enter the legal system. If it does, then there will be different possible scenarios (associated with different information available to the decisionmaker), some of which result in decisions to terminate and others in decisions to continue. In each scenario in which the decision rule dictates continuation, a cost will be incurred — and note that the prospect of this adjudication cost, just like the prospect of a sanction, will make commission of the act less attractive. Moreover, continuation generates further information and thus presents a variety of possible second-stage scenarios, some associated with findings of no liability and others with liability and, accordingly, the imposition of the sanction.

Combining these components, each type of act will, at the outset, be associated with an expected cost: the sum of an expected adjudication cost and an expected sanction. We would

---

[46]For simplicity, appeals (*see infra* note 83) are ignored; alternatively, one can interpret terminations as those that survive appeal.

[47]Distinctions among components, or whether costs of going forward are borne by private or public enforcers, will be irrelevant for most purposes, the exception being the initiation of cases and settlement, examined in subsections III.D.1 and III.D.2 respectively.

[48]For variations, including how the analysis is affected when the level of the sanction may be adjusted and when the sanction is not costless (including the possibility that individuals are risk averse and hence the prospect of uncertain sanctions imposes behavior- and welfare-relevant risk), see subsection III.C.3.

[49]The discussion avoids reference to false positives and false negatives — and of type I and type II errors — because these pairs of terms are ambiguous (depending on which outcome is taken to be the default; for example, continuation can be viewed as a positive outcome because the case proceeds or as a negative outcome when that results from the denial of a motion to terminate), authors use them inconsistently (sometimes in the same article), and readers often need additional time to digest them and occasionally draw the wrong implication.

hope that this sum is greater for harmful acts, because expected adjudication costs are greater (since continuations are more likely) and expected sanctions are greater (since both first-stage continuation is more likely and, conditional on continuation, second-stage liability is more likely). We can see that stage-one continuation decisions influence both types of expected costs and, through both channels, the commission of each type of act. Continuation increases the deterrence of harmful acts and the chilling of benign acts.[50]

2. Optimal Decision Rule. — Consider how best to make the stage-one termination/ continuation decision in any given scenario, taking as given the decisions for other possible scenarios. This caveat is quite important because the optimal decision in any scenario depends on how decisions in others are understood to be made, as will be explained. In principle, one can use this method to consider various permutations in order to determine the optimal decision for every scenario.

In the scenario at hand at stage one, we can ask how the decision to continue rather than to terminate influences social welfare. Most obviously, continuation results in both the actor and the system incurring adjudication costs. In addition, relative to termination, continuation will augment the expected sanction associated with each type of act. The magnitude of each elevation will depend on the scenario, that is, on the information then available. If it is very favorable to liability, the contribution to the deterrence of harmful acts will be relatively great and to the chilling of benign acts rather small; conversely if the information is unfavorable. Moreover, as previously noted, the information will indicate the magnitude of actors' (of both types) expected continuation costs, which may vary across scenarios. Moving back to the beginning, the point in time at which actors decide whether to act, a decision rule dictating continuation rather than termination in the scenario under consideration will increase the expected cost of committing the harmful act by some increment and the expected cost of committing the benign act by some other increment.

Continuation rather than termination will be optimal at stage one, in a given scenario, if and only if the following inequality holds:[51]

Deterrence Gain > Chilling Cost + Continuation Costs

[50]One could employ a single term, such as deterrence or discouragement, for both types of acts. Here, the language of deterrence will be employed only with respect to harmful acts (because this phenomenon is what most have in mind when they refer to deterrence), and the term chilling will be used to refer to the discouragement of benign acts (which is also in accord with common usage).

[51]If scenarios are taken to be discrete, each with a positive probability mass, then switching the rule from termination to continuation in a particular scenario (taking the rules for other scenarios as given) will have a discrete influence on the three identified components, which could conceivably change the direction of the inequality. For example, as explained below, an element of the net deterrence gain and of the net chilling cost is the forgone benefit from harmful and benign acts, respectively, and these involve an average; it could be that the inequality is satisfied for lower values in the pertinent range but not for higher values. If that were so, it actually would be optimal to randomize the decision in the scenario in question (that is, to continue rather than terminate some intermediate percentage of the cases). The more finely the scenarios are defined, the smaller the probability mass of each, and the less important this possibility will be.

The benefit of continuation is that deterrence is enhanced, and there are two costs: the increase in chilling costs and in adjudication costs. Let us now decompose each of these components in turn.

Beginning with the first, we have:

Deterrence Gain = Deterrence Effect × Social Gain per Deterred Act

That is,
the benefit from enhanced deterrence is the product of the increase in deterrence — specifically, how many acts are deterred — and the net social gain per act that is deterred. Each of these two factors requires a further breakdown.

For the deterrence effect, it is first necessary to ascertain the rise in the expected cost associated with harmful acts. This increase will be the product of the fraction of harmful acts that flow into the legal system, the portion of those cases that are in the scenario in question, and the additional costs of continuation for harmful acts (the latter of which was explained previously). This additional deterrence punch,[52] note, will vary across scenarios. In some, there will be a large portion of harmful acts; in others, few. Moreover, if there are also very few benign acts in the scenario, then it is likely that, if there is continuation, it will result in liability; if there are many benign acts in the scenario, that likelihood will be lower.[53]

For a given increase in the expected costs associated with the commission of a harmful act, it remains to determine how many acts will be deterred as a consequence. This quantity will depend on the degree of deterrence already achieved[54] and on the distribution of prospective actors' benefits from committing the harmful act. To illustrate, suppose that the expected cost of committing a harmful act is 50 if the decision in the scenario under consideration is to terminate, but that the expected cost rises to 60 if it is to continue. In that instance, individuals whose benefits from committing the harmful act fall between 50 and 60 would be deterred as a consequence of continuation. Those whose benefits fall below 50 would have been deterred in any event, and those with benefits above 60 will remain undeterred.

To complete this component, there is the further empirical question — the answer to which, like everything else considered here, will vary across contexts — of how many

---

[52]The deterrence punch (and, analogously below, the chilling punch) refers, as stated, to the increased cost associated with committing the act. The term deterrence effect, which appears in the preceding box, was previously stated to indicate the number of acts deterred, which depends on the deterrence punch and also, as described in the paragraphs to follow, on the concentration of individuals' benefits from acts in the pertinent range.

[53]There is a subtle but important distinction being made in the text. For example, it is possible that a scenario as a whole is quite common — so that a good portion of harmful acts fall within it — but that it is likewise true that many (and perhaps many more) benign acts fall within it as well (making subsequent liability relatively unlikely if the decision is to continue). Or one could have a rather rare scenario, so few harmful acts fall within it, but there may be virtually no benign acts in the scenario (making subsequent liability highly likely if the decision is to continue).

[54]This is the point in the argument at which the dependence of the optimal decision rule in a given scenario on the decisions presumed to be made in other scenarios becomes apparent.

individuals with the opportunity to commit a harmful act have a private benefit in the range of 50 to 60. If most have lower benefits, or if most have higher benefits, few would be deterred specifically as a consequence of continuation in the scenario in question. However, if many have benefits between 50 and 60 — that is, if we are near the sweet spot of the distribution — many would be deterred.

Once the number of deterred acts is determined, we can turn to the social gain per deterred act, which involves two benefits and a cost.[55] The first, obvious benefit is that the harm associated with the act is avoided. The more harmful the act in question, the greater is this gain. Second is a reduction in expected aggregate adjudication costs. Had the act been committed, there is a probability it would have entered the legal system and, conditional on that, a further possibility that the case would have been continued, which would result in costs being incurred by both the actor and the legal system (including whoever is the enforcer).[56] Observe that, in this respect, a more costly legal system — say, more expensive discovery or trial — actually favors continuation because, through deterrence, these greater adjudication costs are borne less often.[57]

For each act that is deterred, there is a social cost associated with the private benefit that is forgone.[58] Depending on the context, this benefit may relate to an affirmative act (operating a factory) or not needing to undertake an action (installing pollution control equipment, wherein the harmful "act" would be interpreted as abstention from this costly installation). It is also possible to say something about the magnitude of this forgone benefit. In the preceding illustration, it would be between 50 and 60 — say, on average, about 55 — because acts that are deterred specifically as a consequence of the continuation decision under consideration are those

---

[55]Under different assumptions about the legal setting than those employed in subsection 1, there is a third type of benefit associated with deterrence: a reduction in chilling costs. To see this point, consider the operation of a different mode of enforcement, which is sometimes referred to as enforcement by investigation, wherein legal inquiries are triggered by the observation of a harmful act. *See, e.g.*, Kaplow, *supra* note 30, at 833–36; Kaplow, *On the Optimal Burden of Proof*, 119 J. POL. ECON. 1104, 1122–28 (2011). Consider murder, auto theft, or a visible discharge of a harmful substance, wherein all that remains is for the legal system to identify the perpetrator. By contrast, the setting considered thus far is more akin to enforcement by monitoring (for example, police patrols) or auditing (including random inspections), where it is supposed that enforcement generates some identification rate for each type of act. When investigations are only (or primarily) triggered by the actual commission of harmful acts, the social gain per deterred act is augmented in a key way: reducing the number of harmful acts decreases the number of investigations that will be launched (holding constant the rate at which harmful acts are investigated), which in turn reduces the chilling of benign acts because there are fewer opportunities for misidentification by the legal system. Whether viewed as a higher deterrence gain or a lower chilling cost from continuation, the effect is, all else equal, to make continuation more advantageous relative to termination. Of course, all else is not equal: we are comparing entirely different modes of enforcement applicable in different settings, so in general we would expect much to differ.

[56]Note that the expected adjudication costs depend, in general, on the stage-one decision rules in all scenarios (or at least many of them) because, at the time an act is committed, it is uncertain (even conditional on entering the legal system) what information will be available to the tribunal at stage one.

[57]*Cf.* A Mitchell Polinsky & Steven Shavell, *Enforcement Costs and the Optimal Magnitude and Probability of Fines*, 35 J.L. & ECON. 133 (1992) (showing that expected enforcement costs augment the optimal sanction).

[58]In some settings, the assumption that individuals' benefits from harmful acts count as social benefits may be controversial — notably, in the case of certain crimes; *see, e.g.*, Kaplow & Shavell, *supra* note 33, at 1251 n.705, 1338–50. If one wished, the analysis could readily be modified by excluding this component and in other respects proceeding as indicated. *See also infra* subsection III.C.3 (addressing possible benefits and costs associated per se with the assignment of liability).

with benefits in this range. If initial deterrence was much greater, this benefit range would be correspondingly higher, so the forgone benefit per act deterred would be larger, reducing the net deterrence benefit per act.[59] Likewise, if initial deterrence was very low, this cost of deterrence would be small, making the net deterrence benefit per act larger.

These observations make apparent that the deterrence gain associated with continuation in a given scenario depends importantly on the termination/continuation decisions in other scenarios.[60] For example, if cases in most other scenarios are terminated, deterrence will be relatively low, which implies, as just explained, that the social gain per deterred act will be large since the forgone private benefit is small. On the other hand, when most other scenarios involve termination, expected system costs associated with harmful acts will be relatively low, a factor that reduces the net social gain per deterred act. Yet another source of interdependence is that a given increment, say of 10, to the expected cost of committing harmful acts may deter more or fewer acts depending on preexisting deterrence and the distribution of actors' benefits from harmful acts. For example, raising the expected cost of harmful acts from 10 to 20 rather than from 50 to 60 might deter more or fewer harmful acts, depending on whether the sweet spot of the distribution was closer to the former or latter range.

To summarize, the deterrence gain from continuation, compared to termination, in a particular scenario is the product of the deterrence effect and the social gain per deterred act. The deterrence effect, in turn, had two components. First, we take the product of the likelihood of an act entering the legal system, its then being in the scenario in question, and the increment to an actor's costs (adjudication costs and sanctions) that results from continuation — which together determine the rise in the costs of committing a harmful act, viewed ex ante. Second, examination of the baseline expected costs (that which prevails under termination) and the heightened expected costs (under continuation) allows us to see how many acts will be deterred (they will be those that fall between these two levels). The social gain per deterred act is the sum of the harm and the aggregate expected adjudication costs that are both avoided when an act is deterred, but offset by the average forgone benefit per deterred act.

The optimality of continuation also depends on the chilling cost:

---

[59]The situation in which the forgone benefit exceeds the social harm of the act is referred to as involving overdeterrence (specifically, by reference to first-best, or ideal behavior). Here, overdeterrence from a social perspective would only exist if the forgone benefit exceeded the sum of the harm directly caused by the act and the aggregate expected adjudication cost. If such were present, termination would clearly be optimal because all three effects from continuation would be adverse.

[60]An important aspect of subsection 1's set-up that now can better be appreciated is that the different scenarios under discussion are assumed to pertain to the prospective acts of some given set of actors. By contrast, if some actors knew that their acts would present themselves to the tribunal in, to take a simple extreme, only a single scenario, then only the decision in that scenario would influence their behavior. The analysis throughout the Article takes the relevant cluster of scenarios to be those applicable to a defined set of actors. (Put another way, if there were two groups of individuals who may commit some act, and if the first group presented itself to tribunals in one cluster of scenarios and the second group in a different cluster of scenarios, then the framework developed here would simply treat those two groups as if their acts were different, for the differences among acts that matter concern not merely the harm that the act may or may not cause but also the information that may be generated by the commission of the act.)

Chilling Cost = Chilling Effect × Social Loss per Chilled Act

As this expression suggests, the determinants of the chilling cost resulting from continuation are qualitatively similar to those of the deterrence gain — although, importantly, the magnitudes could be quite different, and this relationship will tend to vary, often greatly, across scenarios. Because the analysis is now familiar, the explanation will be abbreviated.

The chilling effect is determined in precisely the same manner as was the deterrence effect. First, we have the chilling punch, which is the product of the fraction of benign acts that enter adjudication, the likelihood (conditional on that) of being in the scenario in question, and the augmentation of expected costs (both defendants' adjudication costs and possible sanctions) due to continuation. These factors determine the overall increase in the expected costs (viewed ex ante) that an actor associates with the commission of a benign act.

As before, we can use this result to determine the number of acts chilled as a consequence of a continuation decision in the given scenario. Suppose that expected costs would rise from 5 to 6. (Lower figures are chosen because, typically, benign acts would be less likely to enter the legal system and, even if they did, sanctions would be less likely to be imposed. However, there could easily be scenarios — ones with mostly benign acts — where the increment in expected costs from continuation would be larger for benign acts than for harmful acts.) The number of chilled acts will correspond to the quantity whose private benefits fall in the range from 5 to 6. Note further that, even though the range in this illustration is narrow, the number of chilled acts could be quite large, notably, if the sweet spot of the distribution of individuals' benefits from benign acts falls in or near that range (and if there are a large number of potential benign acts).

The social loss per chilled act has two components, one cost and one offsetting benefit. (Of course, unlike with harmful acts, there is no need to account for any harm per act that is discouraged.) The cost corresponds to the forgone private benefit of the act that is chilled, which is the usual cost most have in mind when they worry about adverse effects of laws with regard to the chilling of benign activity. As with deterrence, this component could be very large — which would be so if the preexisting degree of expected costs associated with benign acts was high — or quite small — in the opposite situation. Note that there is an offsetting benefit (one also present with the deterrence of harmful acts): each benign act that is discouraged no longer has the possibility of entering the legal system and thereby resulting in adjudication costs being incurred. The net social loss per chilled act is the difference between these two figures.[61]

[61]It is possible that this net figure would be negative: that is, chilling the marginal act would be a net social benefit. This would arise when, for example, the initial expected cost that prospective actors associate with benign acts was very small (so the forgone benefit was low), but aggregate adjudication costs were large. Specifically, this relationship could hold in a regime in which the expected sanction on benign acts was very low but the legal system component of total adjudication costs (which, note, is not borne by actors and thus does not contribute to chilling) was large. For example, consider an activity that is highly dangerous when conducted improperly, which makes it optimal for the government to employ costly inspections that usually need to be conducted because it is not readily apparent whether

To determine the overall chilling cost, we would multiply the number of chilled acts by the average social loss per chilled act. As with deterrence, this cost could be very high or quite low, depending on a number of factors that vary across scenarios. Also, as before, the magnitude of the chilling cost in a given scenario depends on the termination/continuation decisions in other scenarios for a number of reasons, including that they determine the preexisting expected costs borne by individuals who commit benign acts and thus the level of the forgone private benefit per act that is chilled.[62]

Continuation Costs are the third component of our formulation for the optimal rule:

Continuation Costs = Number of Cases in the Scenario
× Cost per Continued Case in that Scenario

As the box indicates, continuation costs are given by the product of the number of cases in the scenario in question and the cost per continued case. The former total generally includes both harmful and benign acts. Starting with this quantity, we would begin with the number of undeterred harmful acts[63] — which, note, depends on termination/continuation decisions in other scenarios because these influence the level of deterrence — and multiply by the fraction of harmful acts that enter the legal system and also, conditional on that, present themselves as being in the scenario in question. To this, we add the number of unchilled benign acts in the scenario, which is determined similarly.[64]

The cost per continued case in the scenario is the sum of expected defendant's costs and legal system costs going forward.[65] Note that this figure, like most of the others considered, may also depend on the scenario. That is, the information available at the time of the stage-one decision may in some scenarios indicate that these continuation costs will likely be large

---

there is a safety violation. Then, if the private benefit of the activity was very small, it would indeed be optimal to discourage it. This result may be accomplished by charging a license fee that covers inspection costs; firms with private benefits below the fee would choose not to operate.

[62]In addition, as explained with regard to deterrence, continuation decisions in other scenarios affect expected system costs per unchilled act, which also affects the net social cost per chilled act. Also paralleling the preceding discussion, a given chilling punch need not chill the same number of acts when decisions in other scenarios differ. Suppose, for example, that continuation in the scenario at hand raises the chilling punch by 1 unit. In the initial example, this increase was from 5 to 6, whereas if there was continuation in more other scenarios, it might be from 10 to 11. Depending on the distribution of individuals' benefits from benign acts, it may be that many more fall in the 5 to 6 range than in the 10 to 11 range. If this difference is sufficiently large, it might offset the point that the forgone benefit per act is higher in the second case (averaging, say, 10.5 rather than 5.5).

[63]Note that the anticipated decision rule in the scenario under consideration will influence this total. Because we wish to determination continuation costs — which necessarily are associated with a decision to continue — it is appropriate to determine this quantity under the assumption of continuation.

[64]Observe that this first component of continuation costs differs qualitatively from those involving the deterrence gain and chilling cost. As explained, those two factors are determined by the degree to which continuation rather than termination changes actors' behavior, whereas the present factor depends on the overall quantity of undeterred and unchilled acts that prevail.

[65]As mentioned in subsection 1, if we were considering private plaintiffs, who unlike agencies and government prosecutors are not formally part of the legal system, their costs would be included here as well.

(perhaps the scenario suggests that discovery will be highly intrusive or that the trial will be complex) and in others that these costs will probably be small.[66]

Having identified each of the determinants of the optimal termination/continuation decision rule and how they relate to each other, we can now return to the rule itself, which is that cases in a scenario should be continued if and only if:

Deterrence Gain > Chilling Cost + Continuation Costs

As is now apparent, this seemingly simple expression depends on a large number of determinants that interact in complex and subtle ways — some of the subtlety arising because the same factor may enter into the formulation in multiple places, and occasionally with opposing effects. Notably, higher continuation costs for actors raise the extent of deterrence, which is beneficial, but also raise the extent of chilling and continuation costs, which are detrimental. Even more, both deterrence and chilling, and thus continuation, are more valuable when continuation costs are generally higher because the concomitant reduction in the number of acts flowing into the system reduces expected aggregate adjudication costs by a greater amount.

Second, even though this criterion is applicable to a given scenario under consideration, taking the decisions to be made in all other scenarios as fixed, we have seen that there are many ways in which the decisions in those other scenarios influence the optimal decision in the scenario in question. Indeed, all three main components in the box are dependent on how other decisions are made.

Stepping back from the analytics, some additional reflections regarding the complexity of this decision algorithm are in order. Although the factors are many and their interactions in some respects intricate, all are at some level intuitive once the logic is appreciated. That is, inclusion of each factor is the direct consequence of methodically tracing through the impact on individuals' behavior and on the operation of the legal system of continuation, by contrast to termination, at the first stage of adjudication in a given scenario. It should also be apparent that, in the specified setting, these considerations are exhaustive of the pertinent effects on social welfare.

The problem of optimal system design is daunting, as are the challenges of making decisions in particular cases. There are great empirical hurdles regarding many of the pertinent factors, whether considering averages for certain areas of law or specific values in a particular case (scenario). The analysis in this Article, especially in this Part and the next, simply seeks to tell it like it is. One might have hoped that the decisionmaking problem would have been much more tractable, but we can now see that it is not. This reality needs to be confronted, not ignored. Direct attention to the practical implications, whether under existing law or for purposes of designing reforms, is deferred until Part IV.

---

[66] As discussed in section IV.A, it is apparent that the perception that expected discovery costs would be large influenced the Supreme Court's decisions in *Twombly* and *Iqbal*.

3. Discussion. — Although subsection 2 already describes the determinants of the optimal decision rule, one by one, it is useful to examine a number of considerations that refer to combinations of components. This excursion also helps map existing intuitions to this explicit analytical framework.

One useful concept is the diagnosticity/cost ratio. In deciding whether to order an additional medical test or take another core sample to detect the presence of oil, one is concerned not just with the information value or the expense in isolation. Roughly speaking, if the additional effort is expected to be twice as informative, one would be willing to spend twice as much. Because our problem is more complex than these more familiar ones — due to the centrality of ex ante incentives — this idea, while relevant, has a more intricate bearing.[67]

Diagnosticity is a subtle notion, for it is not obvious how one quantifies information. Clearly, we are not interested, say, in how many relevant documents might be discovered, but rather in some sense of what we are likely to learn and how that may affect our decision regarding liability. In general, the relevant measure depends on the decisionmaking environment. Here, we are motivated to continue rather than terminate mainly by the prospect that, given what we know now and what we may learn, it will ultimately be sensible to assign liability. Of course, we cannot know what we will learn, but we can use the existing information, limited as it may be, to estimate the probability of different outcomes in that regard.[68]

It is also necessary to assess the value of this additional information. This quantity is implicitly given by the factors that determine the deterrence gain and the chilling cost. For the deterrence gain, for example, we examined the calculation of the deterrence effect, one component of which was the likelihood that sanctions would be imposed conditional on continuation. But that was not all. We next used that increment to determine how many acts would be deterred, and finally multiplied that by the social benefit per deterred act. In sum, much of the analysis in subsection 2 can be understood as quantifying the value of the information expected to be gleaned if a case is continued. In addition, there is also the chilling cost due to the possibility that liability will mistakenly be imposed, and this too depends on the quality of the information: the higher that quality is, the less will be the increment to the expected sanction for benign acts. Also note that, as explained in subsection 2, the value of information with regard to deterrence and chilling depends not only on the scenario under consideration but also on how decisions are made in other scenarios: those other decisions influence the social consequences of deterring or chilling a marginal act, which in turn is a central component of the value of information, as just explained.

In thinking about the diagnosticity/cost ratio, we also need to consider the denominator. The cost of continuation may vary significantly across scenarios: what is known about a case

---

[67] An additional difference is that the analysis in this Part assumes that nonfinal decisions may terminate a case but cannot assign liability without further proceedings, an assumption that is relaxed in subsection III.C.1.

[68] *Cf.* Bell Atlantic Corp. v. Twombly, 550 U.S. 544, 556 (2007) (referring to "enough fact to raise a reasonable expectation that discovery will reveal evidence of" a violation).

that enters stage one may indicate how much remains to be learned and what efforts are required to do so. Note also that, while there may be some tendency for more information acquisition to be associated with greater cost, any such relationship is quite loose and will be context-specific. A classic illustration is the broad "fishing expedition" in which permitting free discovery may impose massive costs in search of what may be little useful information. At the opposite extreme, it may sometimes be apparent that just a few documents or a single deposition has a significant prospect of being highly probative.

Next we have to review how the cost of continuation enters our formulation for the optimal decision rule. As explained in subsection 2, there is both the expected, straightforward influence — the "plus continuation costs" term — and also additional subtle influences that arise because our problem involves effects on ex ante behavior. Continuation costs borne by defendants contribute to both deterrence and chilling. In addition, for any increment to deterrence or chilling (created by supplemental defendants' costs and by the increment to expected sanctions), the reduction in activity of both types has the added value (or, for chilling, reduced cost) of decreasing the frequency with which cases enter the legal system and thus result in adjudication costs being incurred. Here, an important distinction must be noted: the direct continuation costs incurred when a case continues and likewise the contribution to deterrence and chilling depend on the scenario-specific continuation costs, whereas the system cost savings from fewer acts being committed depends on average system costs over all scenarios.[69] Put another way, for a given level of ordinary continuation costs, it matters whether the continuation costs in the scenario under consideration are atypically high or low.

This final point and some of the others reinforce the idea that the optimality of continuation will tend to vary, often greatly, across scenarios. Depending on the nature of the information available in a given scenario — namely, what it tells us about the increment to deterrence and chilling, as well as the scenario-specific continuation costs — continuation may be very favorable, highly detrimental, or a close call. It may be a useful heuristic to contemplate ranking the desirability of continuation across scenarios and then deciding on a cutoff, above which cases would be continued and below which they would be terminated (although this thought experiment does not capture the complete decision problem).[70]

To summarize the preceding discussion, the diagnosticity/cost ratio is a useful, intuitive notion, but one that is complex to apply and also importantly incomplete in the setting of

---

[69]Regarding the latter, discouragement of an act reduces by one the number that might enter the legal system. Because the deterred or chilled act might have presented itself at stage one in any of a number of scenarios, the expected savings in adjudication costs (i.e., the average) will be the savings in each scenario weighted by the probability of the case being in that scenario.

[70]This heuristic is only helpful to frame thinking, for one must analyze each component of the optimal decision rule in each scenario to do the ranking. Moreover, because of the multidimensionality of the optimal test and the interdependence across scenarios, there does not exist a unique, correct ordering. For example, in some scenario, the deterrence punch may be very high, but the continuation costs may be substantial as well. If the optimal pattern of decisions involves termination in most other scenarios, continuation may be quite favorable, generating a high rank, but if the opposite is true, so that incremental deterrence is not very valuable, the scenario would rank low. More abstractly, as one moves the cutoff, both the absolute and relative desirability of continuation in other scenarios will change, and the latter can alter the ranking. For a formal treatment, see Kaplow, *supra* note 41.

multistage legal proceedings. This ratio points us in a good direction but does not reliably take us to our destination.

Another familiar idea, particularly with regard to decisions on motions to dismiss in U.S. civil litigation, where continuation is required in order to obtain discovery, is to focus on how much of the pertinent information is solely in the possession of the defendant.[71] The standard, intuitive view is that, the more this is true, the stronger the argument for continuation. Much more can be said, however, by making explicit use of our framework.[72]

The idea seems most relevant to assessing the deterrence gain. Suppose, for example, that in some category of settings, enforcers — whether private plaintiffs or government agencies — virtually never have and are not able to obtain much information without accessing that in defendants' possession. If one terminated all such cases, deterrence would be negligible. In that event, the marginal value of deterrence would be high, particularly when harm is great. As explained in subsection 2, a major offset is the cost of forgone private benefits, which, for acts just at the margin, have a value in the range of expected sanctions (plus actors' expected adjudication costs). Under the current assumption, these benefits are near zero. Hence, the harm per deterred act minus the forgone benefit per act is large. Note further that similar logic applies to chilling because the cost of chilling an act likewise depends importantly on the value of marginal chilled acts, which also would be near zero under the present assumptions.

This rationalization, however, is incomplete. For the total deterrence gain to be large, it is not sufficient that the benefit per deterred act is high. We also need to know how much continuation would contribute to deterrence. Taking an extreme, if the information in the scenario under consideration does not in any way suggest that the act was harmful, meaning that only a small fraction of cases in the scenario involve harmful rather than benign acts, then the

---

[71] *See, e.g.*, 2 PHILLIP E. AREEDA & HERBERT HOVENKAMP, ANTITRUST LAW 84–85 (3rd ed. 2007) ("Discovery is most clearly required when the key facts supporting a claim are peculiarly within the other party's knowledge. Because conspiracies, for instance, are usually concealed, conjecture may be inescapable until after the discovery process."); Louis Kaplow, *On the Meaning of Horizontal Agreements in Competition Law*, 99 CALIF. L. REV. 683, 767–68 n.199 (2011) (noting the irony that the *Twombly* Court cited the Areeda and Hovenkamp treatise on another matter, but ignored their direct treatment of the question presented, which appeared verbatim in the previous, pre-*Twombly* edition); *see also* AREEDA & HOVENKAMP, *supra*, at 118 ("The [*Twombly*] majority said little about the reality that facts pertaining to the conspiracy are within the hands of the defendants . . . ."). The present discussion, tracking most prior statements, oversimplifies in assuming that various information, by its very nature, either is or is not available to the enforcer in a given scenario. In fact, the extent of information available will to a degree depend on the enforcer's efforts. *See infra* subsection III.D.1.

[72] *See also infra* section IV.B (discussing how the nature of facts bears on how one should understand the notion that information may be primarily in the possession of the defendant). One could also piggyback on the discussion of the diagnosticity/cost ratio. Information being solely in the defendants' possession may suggest that diagnosticity is high since so little is known initially. By analogy, performing the first medical test or drilling the first core sample may seem particularly alluring. A moment's reflection, however, reveals that this idea is importantly incomplete (and in a manner that parallels some of the analysis in the text to follow). Although there are often diminishing returns to additional investigation, it is also true that in the vast majority of instances it is not sensible to perform the first test. In the medical setting, for tests that have nontrivial costs, we do not administer them to everyone, but only those exhibiting symptoms or belonging to a high-risk group. Oil companies do not take core samples everywhere on the globe, but only in areas where other information suggests some threshold likelihood that oil is present. In our present context, this caveat might be related to the plausibility requirement for motions to dismiss. *See infra* section IV.A.

added deterrence punch from continuation will be negligible, so the overall deterrence gain would be small.[73] If continuation costs are high, termination would be optimal. Therefore, subsection 2's lesson that the deterrence gain equals the product of the deterrence effect and the social gain per deterred act has multiple and, in this example, conflicting implications.

It is also important to apply the teaching that the optimal decision in a given scenario depends on decisions in other scenarios. Suppose that, for the type of harmful act in question, information is almost always solely in the defendant's possession, in which case we may have decisions to terminate in most scenarios. In that case, if in the present scenario there is even a modest — but much higher than average — initial indication that a harmful act occurred, the argument for continuation will be relatively powerful: to address the likely deterrence deficit, it makes sense to allow continuation in scenarios presenting the strongest likelihood and thus the greatest contribution to deterrence. (As mentioned in subsection 2, there is also the countervailing point that, if most other scenarios involve termination, then the expected system costs arising from each harmful act are small, which reduces the net social gain per deterred act. For purposes of the present discussion, it is assumed that this offset is of lesser weight.)

By contrast, suppose that, in many scenarios, harmful acts will generate nontrivial information that can be obtained by enforcers, public or private, but that, in the scenario under consideration, essentially all key information is in the defendant's sole possession. In this instance, termination would tend to be optimal. The other scenarios, where there are strong indications that a harmful act is before the tribunal, will optimally involve continuation, so the deterrence deficit will be much smaller than in the preceding example.[74] Accordingly, since we will have a much smaller deterrence gain, but the same chilling cost (let us say[75]) and continuation cost, termination is relatively more favorable. In addition, a proper assessment of the portion of cases in the present scenario that involve harmful rather than benign acts will naturally be influenced by an understanding of all scenarios. Specifically, that assessment may be lower if a type of harmful act usually generates certain information available to enforcers, but such information is absent in the scenario under consideration. This point implies that continuation will generate a small increment to the deterrence punch, which also suggests that the deterrence gain is smaller.[76]

---

[73]As explained in subsection 2, yet another reason the deterrence gain could be small is that few individuals in positions to commit harmful acts have private benefits near the pertinent level of expected costs.

[74]This point is subject to the important caveat, explained in subsection 2, that a given contribution to the deterrence punch need not have the same deterrence effect because of differences in the concentration of actors' private benefits from committing harmful acts.

[75]Even if the preliminary evidence in the other scenarios that a harmful act occurred is notably stronger, it will usually be imperfect — that is, continuation will expose some benign actors both to adjudication costs and prospect of liability. Because there is continuation rather than termination in the other scenarios (compared to the preceding example), there will be greater chilling and thus a tendency for marginal chilling costs to be greater, although there is the additional factor (that might cut in either direction) that the concentration of individuals' benefits from benign acts may differ in the two relevant ranges for expected costs. *See supra* note 62.

[76]The analysis of this scenario in which information is atypically in defendants' possession assumes, in accord with the exposition throughout, that those contemplating the commission of a harmful act do not know in advance that their act will (for certain or with high probability) present itself in this unusual manner. If they did, their harmful acts should, for present purposes, be analyzed as distinct from others that usually did generate information, and the preceding analysis would accordingly be applicable.

This conclusion is reinforced by strategic considerations. One concerns negative inferences. That is, if, conditional on there actually having been a harmful act, there usually is notably more information indicative of liability, then the absence of such information may lead us to believe that the particular enforcer has a low-merit case, in the extreme, that it is entirely fabricating its claim (an inference that is notably weaker when harmful acts usually do not generate information available to the enforcer). Another point concerns incentives to initiate claims, the subject of subsection III.D.1. If it is known that cases will be continued even when an enforcer has negligible information (claiming that it is all in the defendants' hands), then meritless claims will be encouraged. By contrast, if harmful acts rarely give rise to much incriminating information that can credibly be conveyed and all such cases will be terminated at an early stage for a lack of information, most meritorious suits may be discouraged.

A related idea concerns another factor formally outside the present set-up: the relative importance of external forces that may generate deterrence (and chilling).[77] Most obviously, if we are considering, say, private lawsuits, it will be relevant if there is also effective public enforcement. If so, preexisting deterrence will be greater and thus deterrence less valuable. Similarly, in some settings, market forces or other reputational channels may create substantial deterrence, lessening the incremental deterrence gain from legal proceedings.

The analysis of this subsection indicates that some familiar intuitions have value, but their validity is approximate and incomplete. We can see that they depend on a number of factors that may not initially have been evident but, on reflection, do make sense. These enhancements also enable these familiar notions to be made more operational. Specifically, once one understands more precisely their power and limitations, and the factors that determine their weight, we can better see how to shape rules and apply them to particular cases.

B. First or Intermediate Stages in a Multistage System

This section extends the analysis of section A to settings in which there are multiple stages before final adjudication. This generalization is of interest for a number of reasons. Most obviously, some legal systems, including U.S. civil litigation, have this feature, so it is important to analyze the optimal decision rule at each stage, which also serves as a platform for section D's exploration of the relationship among decision rules at different stages. In addition, the present analysis is a necessary predicate for section III.A's investigation of optimal staging: whether stages should be combined or separated and the optimal ordering of stages are topics that presume the possibility of additional stages and that are illuminated by an understanding of how decisions are optimally made under different structural permutations.

Although incorporating additional stages multiplies the number of factors, the core logic behind the optimal decision rule at any stage (except the final one; see section C) is largely the

[77] *Cf.* A. Mitchell Polinsky & Steven Shavell, *The Uneasy Case for Product Liability*, 123 HARV. L. REV. 1437 (2010) (suggesting that private suits for product liability may be detrimental when market forces and product regulation are strong).

same as that depicted in section A for the first stage in a two-stage system.[78] Consider a legal system with three or more stages. As before, individuals initially decide whether to commit acts of the two types, and fractions of those committing each type of act enter the legal system. In this setting, a decision to terminate at the first stage ends the process, whereas a decision to continue means that some adjudication costs are incurred by the actor and the legal system, some additional information is obtained, and the case enters the next stage.

In stage two, the description of the scenario will include all information learned to date, that is, both the information from when the case first entered the legal system and also that gleaned in moving forward to the current stage. Although more has been learned, uncertainty will remain; that is, in general the decisionmaker will be unable to tell for sure which type of act, harmful or benign, is under scrutiny at that point (if it could, its optimal decision would be obvious). Nevertheless, as more information accumulates, the ability to distinguish the acts will tend to improve: the initial signal may be quite faint, but at later points there may well be a sharper indication of whether the case involves a harmful or benign act.[79]

This process repeats until a case either terminates or ultimately reaches the final stage. At that point, a decision is made whether to find no liability or to assign liability and thus apply the sanction.

We are now in a position to analyze how decisions are optimally made in any given (nonfinal) stage.[80] The analysis follows the approach of section A in considering a particular scenario, taking as given decisions in other scenarios — and, now, how decisions will be made in all scenarios at other stages, before or after.[81] Most components of each of our main three factors — deterrence gain, chilling cost, and continuation costs — are similar or the same as they were in section A's analysis. There are, however, some important differences.

For the deterrence gain, a major factor is the deterrence punch: the extent to which the prospect of continuation raises the expected costs for those contemplating the commission of harmful acts. If a case is at the penultimate stage, the only adjustment to section A's analysis is

---

[78]One way to think about this observation is to collapse all prior stages into what was there taken to be the initial scenario and all subsequent stages into a single, final stage. Nevertheless, given the importance of the problem, the fact that it has not previously been analyzed, and the interest in the questions outlined in the preceding paragraph in the text, it is worth some effort to articulate this notion more fully.

[79]There is, as stated, only a tendency: for example, an initial scenario (at stage one) may include predominantly harmful acts, but at the next stage, one may be in a sub-scenario in which it is less clear that the case involves a harmful act. For example, suppose that initially it is highly likely that the act is a harmful one. At the next stage, there is a 90% chance that we will learn that the act is almost certainly harmful and a 10% chance that we will learn that it is only 50-50. Even though most cases in the initial scenario will fall in the former sub-scenario, in which we have a sharper identification of harmful acts, some will fall in the latter, where we are more uncertain. (By analogy, tomorrow morning's weather report may be a better indicator of tomorrow's weather than is the previous night's forecast, but it is entirely possible that the advance forecast indicates a 90% chance of rain whereas, in some instances, the subsequent one involves a revision to a 50% chance.)

[80]For more formal analysis, see Kaplow, *supra* note 41.

[81]As in section A, one can, in principle, combine the analysis of different scenarios and different stages to determine the optimal decision rule for each scenario at each stage. This process is obviously complicated, all the more so because optimal decisions in different scenarios and at different stages are interdependent.

that, instead of considering the fraction of harmful acts that enter the legal system, we need to consider the fraction that enter the legal system and survive to the present stage (and fall in the scenario under consideration). The rise in expected costs due to actors bearing adjudication costs and the prospect that liability will be found and thus the sanction applied at the final stage is determined as before.

At prior stages, on the input side there is the same adjustment: we need to know not only the fraction of harmful acts that enters the legal system but also what portion reaches the stage under consideration. On the output side, the analysis is a bit more complicated, but the idea is largely the same. With continuation, we do not simply assess the expected sanction that will be imposed at the next stage, for the next stage is no longer the final stage. Instead, we need to assess the probability that the case will be continued at the next stage (and the next and the next, if applicable) as well as the conditional probability that, if the case does reach the final stage, liability will be found. Thus, we again have an expected sanction, but the construction of this expectation is more involved.

Regarding the deterrence gain, we see that the main difference concerns determination of the increase in the deterrence punch attributable to continuation. Note that, as the final stage is approached, the magnitude of the increment to the deterrence punch associated with the expected sanction will tend to be rising because cases will have survived longer and surviving cases will be stronger on average than those that were terminated along the way. This effect will not hold in all scenarios, but this general tendency will prevail.[82] In addition, we learned in section A that actors' expected adjudication costs also contribute to deterrence, and these too will be incrementally incurred upon continuation through each stage. However, as one progresses to later stages, more of the costs will be sunk (as elaborated in a moment), so this contribution to the deterrence punch will tend to be falling.

For the chilling cost, we need to make the same modifications: that is, for benign acts, we need to take into account both the probability of reaching the stage in question (not just the fraction of acts entering the legal system) and, if the case is continued, the probability that it will be subsequently continued and ultimately result in liability. Accordingly, the aforementioned reasons that the deterrence punch may tend to rise or fall over time will be applicable to the chilling punch as well.

The final difference concerns continuation costs. When short of the penultimate stage, continuation not only generates the costs of moving to the next stage but also gives rise to a probability of moving to the stage after that (and so on). Hence, the relevant notion is expected continuation costs, some (those from moving to the subsequent stage) borne with certainty and others (those thereafter) borne with a probability. Observe that expected continuation costs tend to be falling as one moves to later stages because the costs of prior stages will be sunk. All else

---

[82]For example, at an early stage, a particular scenario may strongly indicate liability, which implies a high probability that sanctions will ultimately be applied. But, in some subsequent scenarios, the new information will be adverse to liability. *See supra* note 79. Therefore, even though fewer hurdles remain, it may be less likely that all will be overcome; indeed, if the additional information is sufficiently negative, the case will be terminated promptly, making this probability zero.

equal, this phenomenon will make continuation more attractive at later stages.

Each of these modifications makes the optimal decision rule at any stage more complex, but we can see that the fundamental determinants are qualitatively the same and interact much as they did in section A's analysis. It should also be apparent that much of what was said there in comparing optimal decisions across scenarios is also applicable to the comparing decisions across stages. For example, in some systems and at some stages, continuation to the next stage may be especially costly or unusually cheap. Similarly for diagnosticity (and the many subtle features that are relevant when this basic idea is decomposed). Anticipating some of the analysis in section III.A, one might wish to design a legal system's stages in order to examine first that information which has the most favorable diagnosticity/cost ratio — for example, the few documents or witnesses that may be particularly illuminating. If that were done, the benefit of continuation at later stages, for a given likelihood that the scenario involves harmful rather than benign acts, would tend to be lower than otherwise. The caveat is crucial, for when early information indicates that the case almost surely involves a harmful act, continuation would be favored and, as subsection III.C.1 suggests, prompt closure with assignment of liability would tend to be optimal. Similarly, when early information is strongly negative, there would be an additional reason to terminate. When these considerations are set to the side, the fact that, in a system thus designed, continuation at later stages entails learning less while spending more will increasingly favor termination.

C. Final Stage

Although the optimal decision rule for the final stage is fairly complex, it is simpler than that for preceding stages. Entering the final stage is qualitatively the same as entering any nonfinal stage. As explained in section B, to determine both the deterrence and chilling punches, we will wish to know not just the fraction of harmful and benign acts, respectively, that enter the legal system but also how many of each survive to the final stage and present themselves as being in the scenario at hand. Again, the understanding of this scenario will reflect all the information that has accumulated along the way, up to and including, say, the trial.

If a case is terminated at the final stage, there is nothing further to examine, just as at all earlier stages. By contrast, a decision to continue is now straightforward: it simply amounts to the assignment of liability. Since there is no further continuation,[83] there are no continuation costs to be borne by the actor or the legal system.[84] Also, there is no further need to estimate the expected sanction that might later be imposed. The optimal decision rule for the final stage —

---

[83]The present analysis abstracts from appeals, the consideration of which raises a number of additional issues. At the broadest level, each appellate level could be viewed as an additional stage, subject to the same analysis as developed in this article, in which case trial would be an interim rather than final stage. In some legal systems, this interpretation seems fitting because additional evidence may be introduced and the analytical framework here focuses on facts. In others — including most appellate settings in the United States — appeals are largely confined to legal questions applied to an existing record. The function of appeals often differs, including for example, enhancing uniformity, employing a more expert decisionmaker, reducing the potential for shirking or corruption by trial court judges, and inducing parties to reveal information about the strength of their cases through the decision to appeal.

[84]There would be an analogue to continuation costs if sanctions were socially costly, as explored in subsection III.C.3.

that is, the optimal burden of proof — is, therefore, a truncated version of what we had previously.[85]

Deterrence Gain > Chilling Cost

As in subsection A.2, each of these two components can be analyzed further. And, as in section B, the main difference concerns the calculation of the deterrence and chilling punches. Before, we considered everything that determined the likelihood that, say, a harmful act would end up as a case in the given scenario (here, at the final stage, reflecting survival at all prior stages as well as presenting the information associated with the scenario in question) and multiplied by the expected costs borne by an actor as a consequence of continuation: the defendant's expected continuation costs plus the prospect that liability would be imposed times the magnitude of the sanction. In the final stage, the former component is zero (there are no further continuation costs because adjudication is at an end) and the probability element of the latter component equals one because we are assuming that the final stage has been reached and contemplating that we are imposing liability. The rest of the analysis of the deterrence gain is as before: the number of acts deterred is determined by considering the difference between prospective actors' expected costs with and without continuation (here, imposition of liability) and asking how many harmful acts have private benefits in that range, and the social gain per deterred act is qualitatively the same. The chilling cost is determined analogously.

Note that, although there are no continuation costs after the final stage, total legal system costs are still relevant to this calculus. As explained in subsection A.2, for each act that is deterred or chilled, there is a reduction in expected aggregate adjudication costs because one less act has the prospect of entering the legal system and then, with some further probability, proceeding to the next stage, and so forth. Accordingly, at the final stage, higher adjudication costs unambiguously favor liability. Recall that, in this respect, even chilling effects are

---

[85]One may wish to compare the present analysis to that in Kaplow, *supra* note 30, at 752–72, and Kaplow, *supra* note 55, on the determinants of the optimal burden of proof. If one matches the corresponding elements, point by point, the formulations obviously have to be equivalent. The main differences are that the setting here introduces two additional complexities. First, one or more stages precede the final stage, whereas in those articles, which focus on the burden of proof, it was assumed for simplicity that cases entering the legal system proceeded directly to final adjudication. As a consequence, there was no need to focus on conditional probabilities concerning whether a case overcame prior hurdles. Second, the core case in those articles abstracted from the costs of adjudication, whereas they play an important role here. In any event, analysis of the final stage is included here for completeness and to facilitate the comparison of optimal decision rules across stages, including the final stage, not to offer a significant advance in our understanding of the burden of proof.

It is also useful to keep in mind one of the major themes of that prior work: optimal decision rules in final adjudication do not have the form of a Bayesian posterior probability threshold — that is, a target minimum likelihood that the individual before the tribunal committed a harmful act rather than a benign act. (Rather, they can be expressed as a threshold value for the likelihood ratio, which is quite different.) *See* Kaplow, *supra* note 30, at 772–805, 812–13; Kaplow, *supra* note 55, at 1117–21. This disjunction is even greater with regard to nonfinal stages because of continuation costs. For example, as discussed in subsection A.3, it is not even possible to form a simple ranking of scenarios in terms of the relative desirability of continuation due to the multidimensional nature of the problem as well as the fact that the values along some dimensions depend on how decisions are made in other scenarios (and at other stages), which can change the ordering. *See supra* note 70. From this perspective, the statements in *Twombly* and *Iqbal* that the plausibility test is not just a probability requirement are normatively correct.

favorable. That is, even when chilling is undesirable, the extent to which this is so is less when aggregate adjudication costs are higher, so the net disadvantage of chilling, the right side of our inequality, is smaller.

D. Relationship Among Decision Rules at Different Stages

The analysis throughout this Part of the optimal decision rule for a given stage of legal proceedings (including the final stage) takes as given the rules for other stages (whether they are set optimally or not). This section addresses two questions pertaining to the relationship among decision rules across stages. First, how do tougher or more lenient rules for continuation at some stages influence the optimal decision rules at other stages? Second, is it optimal for the stringency of the rule for continuation (and, at the final stage, for liability[86]) to become tougher at later stages, as is commonly supposed?[87]

Regarding the first question, a major theme of the analysis thus far emphasizes how the formulation of the optimal decision rule at every stage depends on what happened before and what is expected to happen subsequently. In section A, a key element of both the deterrence and chilling punch was the manner in which cases that continued to trial would be decided, and this point was generalized in section B, where there may be multiple subsequent stages. Likewise, in all stages, from the first through final adjudication, a key input was the likelihood that an act, whether of the harmful or benign type, would enter the legal system and survive to the stage under examination. Hence, cross-stage interdependencies are central to the determination of the optimal decision rule at each stage. To explore our question further, it is helpful to consider four settings.

Suppose initially that the first-stage decision rule is strict, which is to say that, in most scenarios, cases are terminated. This result may be thought to arise because it is optimal —

---

[86]Note that since the analysis covers decision rules at all stages, the criterion at the final stage — the burden of proof — is also encompassed. Therefore, the discussion is applicable to determining how the toughness of termination/continuation decisions bears on the optimal height of the burden of proof, and vice versa.

[87]Despite this general understanding, if one views in a vacuum the two tests for motions to dismiss and summary judgment, as described in the Introduction and elaborated in sections IV.A and IV.E, it is hardly clear which imposes a higher hurdle for a case to proceed. If no "genuine dispute" means that even a scintilla of contrary evidence is sufficient to reach the factfinder, and if "plausibility" requires something notably more substantial, which seems to be envisioned, then the motion to dismiss standard would actually be tougher, which contradicts what one supposes is the unanimous view of courts and commentators. *See, e.g.*, Clermont & Yeazell, *supra* note 10, at 833 n.47 ("As to facts, the Court's articulation and application of the new test in *Twombly* and *Iqbal* may appear to require a stronger claim than does summary judgment, but that relationship would be nonsensical. It would instead make policy sense to require a weaker claim at the pleading stage, but (1) there is a limited number of choices among decisional standards, (2) any standard less demanding than summary judgment's reasonable possibility test would equate to the old scintilla of slightest-possibility standard, and (3) nothing in *Twombly* or *Iqbal* suggests that the Court meant such a low standard."); *see also* 2 AREEDA & HOVENKAMP, *supra* note 71, at 118 (stating that, to survive a motion to dismiss, "a plaintiff's allegations must 'plausibly suggest[]' conspiracy," but to survive a motion for summary judgment, "the evidence must 'tend to rule out the possibility that the defendants were acting independently,'" there being a distinction because "[t]he 'plausibly suggesting' threshold for a conspiracy complaint remains considerably less than the 'tends to rule out the possibility' standard for summary judgment"; yet failing to offer definitions or argument supporting the asserted obvious ranking in light of the fact that merely "tending" to rule something out can easily be understood as weaker than "plausibly" ruling something out).

perhaps costs of continuing to the second stage are particularly high — or on account of an institutional constraint. Compared to a situation in which more cases are continued at stage one, this strictness tends to favor looser rules at later stages, that is, ones more inclined toward continuation and, ultimately, liability. Because few cases survive stage one, there is likely to be a large deterrence deficit, making the gain per deterred act larger, and also a smaller chilling cost because marginal chilled acts will be ones with lower forgone benefits. (Both points are subject to the caveats presented in subsection A.2, which will not be repeated for the settings that follow.[88]) In addition, because of the tough screening at stage one, the mix of surviving cases will tend to be stronger: that is, ones involving a greater share of harmful acts rather than benign acts. Note finally that, if a strict approach was taken at stage one because of the high costs of continuing to stage two in particular (rather than the high expected costs of subsequent continuations), these costs will be sunk for those cases under consideration at later stages. (Anticipating the second question addressed in this section, note that we have just described a setting where decreasing stringency may be optimal.)

Next, suppose instead that the first-stage decision rule is lenient. Then, at the second stage, for example, there will be more cases, which implies that a given decision rule would result in greater deterrence and chilling, making the benefit of deterring the marginal harmful act lower and the cost of chilling the marginal benign act higher, both leaning toward termination, that is, a tougher rule. Likewise, when little screening is done at the first stage, the remaining mix of cases is weaker, which is to say, a higher portion involve benign rather than harmful acts, relative to the prior illustration. Finally, anticipating section III.A, if the reason for the lenient first-stage screening is that the legal system places high diagnosticity/low cost steps early in the process, then at later stages the diagnosticity/cost ratio will be less favorable, also favoring termination. In each of these two settings, we see that optimal subsequent rules have a tendency to lean in the opposite direction of how stringently early rules are set.

Consider now the reverse interaction: how later rules affect the optimal toughness of early rules. First, suppose that later rules will be very generous toward continuation or, ultimately, liability. For a given flow of cases into those stages, both deterrence and chilling will tend to be high. Accordingly, the optimal first-stage rule (or that at other early stages) will tend to be more stringent. (Again foreshadowing the second topic in this section, note that we have another setting in which stringency would be falling as we move to later stages.)

Second, suppose that later rules are quite strict toward continuation or liability. In this instance, it is not obvious that earlier decision rules should attempt to offset the large deterrence deficit and take advantage of the low cost of chilling the marginal act through leniency. If it is costly to continue cases, and if most cases in a scenario will ultimately be terminated at a later

[88]First, and directly offsetting the point in the text, is that a high rate of stage-one terminations implies low expected system costs for undeterred and unchilled acts, which reduces the value of subsequent continuation and liability. Second, and possibly cutting in either direction, the number of acts deterred or chilled per unit of increase in the expected cost of an act could be larger or smaller, depending on the distributions of individuals' benefits from the two types of acts.

point or result in no liability, then the continuation costs may largely be wasted.[89]

Some of these examples align with familiar ideas about legal rules, particularly those that favor a more stringent test for summary judgment. A lenient approach at the motion-to-dismiss stage is seen as justifying a tougher approach toward summary judgment (a view that seems to be motivated by the point about the relative weakness of the mix of cases that remain rather than by the other arguments).[90] And a concern that factfinders, juries in particular, may excessively favor plaintiffs is thought to explain some of the more stringent approaches at the summary judgment stage, wherein defendants' motions sometimes seem to be granted even when nontrivial disputes over facts exist.

The foregoing analysis is incomplete in that it does not fully address the reasons for the existence of the posited stringency or looseness at stages other than the one under consideration. One set of factors that may be at play — particularly if those other rules are being set optimally — are different values for key determinants of the optimal decision rules at all stages. To illustrate, suppose in our first example, where it was posited that the first-stage decision rule was strict, that the reason for this is that harm is particularly low and, moreover, there are an unusually large number of benign acts that the legal system has difficulty distinguishing from harmful ones. Then, the rule is optimally strict because the deterrence gain is low and the chilling cost high. In that event, despite the stage-one strictness, it may well be true at later stages that a strict approach is likewise optimal, for the low harm means that deterrence is of low value, and the high incidence of benign acts may remain, even if to a lesser extent, at later stages. In other words, shifts in many key parameters may well have similar effects on the optimal stringency of decision rules at all stages.

Turn now to our second question, concerning how the optimal stringency of decision rules changes as one moves to later stages, and in particular the common intuition that optimal stringency rises.[91] As a preliminary matter, it is not clear in theory what it means to compare

---

[89]The analysis in section A indicates that this point is not entirely true because those continuation costs borne by defendants, like the prospect of formal sanctions, contribute to deterrence and also to chilling. Therefore, when the social gain per deterred act is large and the social cost per chilled act is low, and defendants also bear a significant portion of aggregate continuation costs, it is possible that a more lenient approach toward continuation would be optimal precisely because of the continuation costs that would be incurred.

[90]*See* Celotex Corp. v. Catrett, 477 U.S. 317, 327 (1986) (advancing a nontrivial summary judgment hurdle in light of the Federal Rules's adoption of notice pleading, which no longer removed factually weak cases from the legal system at the outset).

[91]*See supra* note 87. Although this belief seems widely held, it does not appear to be fully taken advantage of by litigants and by lower courts in writing opinions on motions to dismiss and for summary judgment. After *Twombly* and *Iqbal* in particular, one might expect a plaintiff seeking to survive a motion to dismiss and worried about how the judge will interpret the vague plausibility test (*see infra* section IV.A) to argue, where possible, that its case exceeds the (clearer?) standard for summary judgment and hence, a fortiori, should not be dismissed. Similarly, a defendant seeking to succeed on summary judgment might wish to argue that the case fails to meet the *Twombly*/*Iqbal* standard and hence, a fortiori, it should be granted summary judgment (the latter argument perhaps assuming greatest importance in the time period covering cases that were filed and got past the motion-to-dismiss stage before *Twombly* and *Iqbal*). Discussions in civil procedure treatises do not affirmatively identify the prevalence of either sort of argument. The closest this author has identified is an observation in MOORE'S FEDERAL PRACTICE, *supra* note 22, § 56.05, at 18 (in discussing *Twombly* and *Iqbal*, stating: "It is too early to assess the impact that the Supreme Court pleading cases will have on summary judgment practice.").

stringency across stages. Within a stage, as discussed in the examples just above, there can be an unambiguous comparison: if, under one rule, cases are terminated in all scenarios in which they are terminated under a second rule, but also in additional scenarios, we can say that the former rule is more stringent.[92]

Across stages, however, comparisons are more murky because the scenarios are different. For any scenario at which there was continuation at a prior stage, there will be sub-scenarios at a subsequent stage corresponding to differences in the information revealed as a consequence of the prior continuation decision. On one view, we could say that any terminations whatsoever (or, at the final stage, findings of no liability) imply greater strictness, for the decisionmaker would be passing forward no cases that were terminated earlier and also would be ending some cases that were previously continued. This comparison, however, is unilluminating. We could instead examine the overall percentage of cases terminated at different stages, but that comparison is rather arbitrary and in many respects may tell us more about the flow of cases than the stringency of rules. For example, suppose that large numbers of obviously frivolous cases are filed, and that a high percentage are terminated at the first stage even by an absolutely low standard for continuation. By contrast, application of a seemingly tough rule at a later stage that terminates many of the remaining cases, including some of fairly high merit, may nevertheless terminate a lower percentage than that for those terminated at stage one. Is it helpful to describe the second rule as more lenient?[93] The difficulty in making such comparisons reflects that the formulations for optimal decision rules developed earlier in this Part involve complex, multi-factor tests that do not entail any clear, unidimensional target, such as a continuation percentage or even the more seemingly plausible criterion of a target probability that the case involves a harmful act rather than a benign act.[94]

It remains interesting to consider whether differences across stages bear in systematic ways on the stringency of their respective optimal decision rules, even if clean comparisons cannot be made. Four differences are examined: in the information available, in the case mix, in continuation costs, and in the deterrence and chilling punches.

Perhaps most obvious — and underlying the familiar intuition favoring stricter rules for continuation at later stages — more information becomes available as a case proceeds through adjudication. Although qualifications might be noted, this predicate seems ordinarily to be true, on average, and perhaps for the bulk of cases. It does not, however, carry the supposed

---

[92]By contrast, two rules that each terminate cases in some scenarios in which the other rule continues them cannot be ranked in this fashion.

[93]A further problem with this measurement approach concerns settlement. If, for various reasons, most strong cases settled prior to the stage in question, the remaining set would be very weak and a high portion should be terminated. But if mostly the weak cases settled and only strong cases remained, then a low portion should be terminated. Anything like a target percentage thus can have extremely different consequences depending on settlement behavior. (Of course, the analysis is more complicated because, in doing system design, including the choice of continuation rules, one would want to take account of the effect of various rules on settlements. *See infra* subsection III.D.2.)

[94]As mentioned in note 85, none of the optimal decision rules reflect a targeted Bayesian posterior probability. Moreover, as explained in note 70, because the tests at all but the final stage involve continuation costs, it is not even true that they can be formulated as targeting a likelihood ratio (likelihood ratios being directly relevant in comparing aspects of deterrence and chilling). For a more complete and formal statement, see Kaplow, *supra* note 41.

implication. That is, there is no general proposition that a better informed decisionmaker should apply more stringent decision thresholds. To start, given that it is often possible to frame movements in a threshold in either direction as tougher (for example, a plaintiff might find a stronger evidence requirement tougher, a defendant more lenient), there cannot exist such a broad truth. Consider an analogy to the medical decisionmaking context: Does better information (less noise) on average favor having surgery more often? Less often? Or just as often, but in a different subset of cases? Clearly, if having surgery is costly and dangerous, and the benefits are confined to special cases, surgery would only be undertaken with good information — and then, in few cases, but still a greater percentage than when there is almost no information. But if a treatment was low cost and low risk and the potential benefits were great (perhaps an immunization), it may nearly always be undertaken when there is little information, but omitted more frequently when there is very accurate information that indicates it is unnecessary or detrimental in an identifiable subset of cases.

Second, at later stages, there is a different mix of cases. If earlier termination/ continuation decisions are made rationally, we would generally expect the mix of cases that remain in the system to be stronger as we move to later stages. If nothing else were different, this would seem more favorable to continuation. However, combining this point and a preceding one, we can see that there is a connection with the aforementioned ambiguity of defining what it means for stringency to change over time — namely, that every case initially terminated cannot, by definition, be continued at a later stage whereas any case initially continued might be terminated later. Specifically, when a scenario is associated with early-stage continuation, this reflects that the cases in that scenario were sufficiently promising. At the next stage, more is learned: some cases will look more promising than they previously did and others less so. Those in the latter group will be terminated if the subsequent information is sufficiently negative. We can see that the overall mix of cases at the later stage is stronger, but, in light of what is learned in the interim about some of them, there will be cases that now are weaker than some that were terminated previously. Accordingly, on average cases may be stronger and on average we have additional information, but some of that additional information tells us that certain of the cases are weaker, and those are ripest for termination.

Third, continuation costs differ across stages. Most obviously, the more stages a case has survived, the more aggregate adjudication costs have been sunk, so the lower the cost of continuing further. The extreme is after trial, when all costs are sunk: removal of the continuation costs component from the formula, all else equal, favors continuation, that is, liability. This simple but potentially powerful point seems to have been largely overlooked in past thinking about the subject.[95] To be sure, continuation costs are not the only consideration. Nor is this point as simple as it first appears. As already noted in section A, continuation costs

[95]Edward Cooper, in addressing (prior to *Anderson*) the question whether the standards for directed verdict (now, judgment as a matter of law) and summary judgment should be equivalent, raises but then immediately dismisses as obviously wrong (but without explanation) the point that granting summary judgment saves trial costs whereas a directed verdict does not. *See* Cooper, *supra* note 18, at 953–54. Although the notion that it may optimally take stronger evidence to proceed to trial than to prevail at trial, once the costs of trial are sunk, may seem jarring, we can see that it is hardly illogical (particularly in cases in which the same tribunal makes both decisions, such as when a judge rather than a jury will preside at trial in the U.S. legal system).

borne by defendants contribute to deterrence, which is valuable, although they likewise contribute to chilling, which is detrimental. Additionally, it is not necessarily true that expected continuation costs are falling as we move to later stages. In some instances, expected (that is average) continuation costs, looking forward, may have been low, but for some information sets at the next stage, what was learned in the interim indicates that subsequent continuation costs will be atypically high. Of course, this cannot be true on average, but only in certain scenarios. Also, in an early-stage scenario, expected continuation costs may not have been very high because, although continuation was net socially beneficial, this was true despite the fact that it was anticipated that most cases will be terminated at the next stage. (Perhaps there is a large deterrence deficit, and the costs of continuing for just one stage are fairly low.) Then, for the subset that are not terminated subsequently, it may be that the likelihood of further continuation, across many stages, some of which are expensive, would be high.[96] Again, this would not hold on average.

Fourth, previous analysis indicates that the contributions of continuation to the deterrence punch and the chilling punch differ across stages. The increment to the expected costs of an act, recall, is the sum of defendants' expected adjudication costs and the expected sanction. The former, paralleling the above logic, tend to be falling as one moves to later stages, because more costs are sunk, whereas the latter tends to be rising because there are fewer remaining hurdles — and, in the last stage, of course, the expected increment to the sanction from proceeding rather than terminating is simply the full sanction. Because of these two competing effects, this factor is of indeterminate direction with regard to both deterrence and chilling. Moreover, we have a further ambiguity due to the fact that greater deterrence is desirable whereas greater chilling is undesirable.

In all, we can see that whether optimal stringency rises as cases move to later stages is far more complicated than is generally appreciated. Note that the foregoing analysis takes the structure of the legal system — of staging in particular — as given. As mentioned previously, the analysis to follow in section III.A of staging will suggest that it tends to be optimal to arrange the sequencing such that high diagnosticity/cost activities come earliest in the process. If that is done, then for the reasons given, optimal stringency would tend to be lower at earlier stages and higher at later stages than otherwise.[97]

---

[96]Moreover, if for that subset the continuation costs are high enough to justify termination, then the earlier-stage decision to continue was probably suboptimal because, as described, it was largely predicated on the assumption that the subset revealed to be of high merit would be continued.

[97]The reader might note that the phrasing of this sentence in the text is stated in "than otherwise" form, which is to say that the point allows us to compare the stringency of, say, the stage-one threshold in cases in which a high diagnosticity/cost ratio segment is placed next rather than one with a low ratio. The sentence does not state that the stringency is high earlier *than later*, for reasons that are now apparent.

III. VARIATIONS AND EXTENSIONS

A. Optimal Staging

In some legal settings, such as U.S. civil litigation, legal proceedings consist of few stages that have a substantial all-or-nothing character. For example, a motion to dismiss is either granted, terminating the case, or denied, which then allows full discovery. The high stakes of this choice, moreover, underlie much of the tension in the Supreme Court's recent decisions in *Twombly* and *Iqbal*.

This extreme concentration of activity into a single stage or only a couple is suboptimal in many situations and does not characterize decisionmaking processes in various other realms. Medical decisions are often sequential. Typically, a doctor or nurse does not, after a limited initial exam, decide either to refrain from any further scrutiny or treatment, on one hand, or order every conceivable test, perhaps supplemented by second and third opinions, on the other hand. Instead, initial symptoms generate a closer assessment, followed perhaps by preliminary tests, the outcomes of which feed into follow-up decisions regarding further investigation and possible treatment. Likewise, in many aspects of ordinary life or the conduct of complex business activity, investigation and decisionmaking involves more finely graded steps.

Some legal systems also have these features. Continental legal proceedings in many jurisdictions have a more graduated character: even though final decisions are not ordinarily rendered at interim points, choices about the extent and nature of further information-gathering are more often determined sequentially. More broadly, police, prosecutors, and government agencies often proceed in multiple, smaller steps — sometimes nearly a continuous flow — with informal and occasionally formal decisions made along the way about whether to terminate a case deemed to be unpromising, to collect additional information, or to deem the amassed body of evidence sufficiently complete to move to a subsequent stage or a final one at which a decision on liability will be made. Furthermore, some contemplated reforms embody more incremental staging,[98] such as the possibility of allowing limited discovery to assess the adequacy of allegations in settings in which key information is solely in the possession of defendants.

This section analyzes optimal staging in three steps. First, taking a particular ordering as

---

[98]

> If pleadings were used to focus legal and factual disputes before discovery began, or if discovery alternated with legal resolution, constantly paring away issues, the process would be more tolerable. Civil law systems do this, and so do many common law systems. . . . Once the [judicial] officer decides what is likely to make a difference to the outcome, it is easy to decide what ought to be acquired in discovery. If the case is unresolved after the judge applies the law to what is discovered in response to this request, the officer ascertains the next appropriate inquiry and directs discovery concerning it. . . . Litigants in nations that employ this system do not complain about abusive discovery; to the contrary, they like their process and complain when portions of our system (which they think barbaric) begin to intrude. The judicial officer does not make impositional demands, and the link of discovery to the merits greatly cuts down on the number of demands made for any purpose.

Frank Easterbrook, *Discovery as Abuse*, 69 B.U.L. REV. 635, 644–45 (1989); *see* Epstein, *supra* note 16, at 206–07 (proposing staggered discovery).

given, it asks whether adjacent pieces should be separated or combined.[99] Specifically, following the analytical framework of Part II, it considers whether, with respect to two adjacent stages, it is best (A) to conduct the first segment, make a termination/continuation decision, and then conduct the second segment if the interim decision was to continue, after which there is a further termination/continuation decision, or instead (B) to conduct both segments as one, making a single termination/continuation decision at the conclusion. Next, taking particular groupings as given, it considers how they are optimally ordered. Finally, it addresses what sorts of evidence gathering and assessments should be assembled into a stage. If these analyses are aggregated, one can in principle assess and compare different proposed schemes and determine which is superior.[100] As will become clear, however, the analysis is more complex than one might have expected, particularly because, as we learned in Part II, decisions in particular scenarios and at particular stages optimally depend on how decisions in other scenarios and at other stages are made.

For the first question, the combination or separation of adjacent stages, begin from a baseline that has the two components combined. The direct benefit of separation involves an option value: if the information obtained during the first stage is sufficiently negative, it will be optimal to terminate at that point, thereby saving the continuation costs that would have been associated with the second stage. The magnitude of this option value will depend on the likelihood that the information will be negative enough to justify termination and the size of the cost savings from omitting the second stage.[101]

The direct cost of separation is any loss in synergies from information collection.[102] This magnitude could be large or small, and it will depend on the nature of the separation involved. For example, if one sequences document exchange, followed by expert reports that use the documentary information as inputs, there may be little lost synergy, but if one instead sequences by subject matter (perhaps elements), it may be that the need to review documents twice or

---

[99]In the law and economics literature, the seminal paper is William M. Landes, *Sequential Versus Unitary Trials: An Economic Analysis*, 22 J. LEGAL STUD. 99 (1993); *see also* Kong-Pin Chen, Hung-Ken Chien & C.Y. Cyrus Chu, *Sequential Versus Unitary Trials with Asymmetric Information*, 26 J. LEGAL STUD. 239 (1997) (extending Landes's model to incorporate asymmetric information in settlement bargaining), although the focus in that work is on how reductions in expected trial costs affect settlement and filing decisions (in manners that could raise system costs) in a setting in which legal outcomes are unaffected and there is no explicit attention to ex ante behavior.

[100]An apparently excluded category (although it might be viewed as embodied in the third step) is the possibility of omission — notably, evidence that is never worth collecting. If, however, the present analysis is joined with that in subsection C.1, wherein early-stage liability is also contemplated, one can consider moving such low-payoff efforts to late stages, where it would emerge (if such was optimal) that either termination with no liability or termination with liability would always happen before such a stage was reached. *See infra* note 135.

[101]Each of these components involves further subtleties. The termination threshold needs to be optimally determined, which would be done by application of Part II's analysis to each of the scenarios that might arise. The cost savings, note, may be less than meets the eye when there are synergies (explored in the text to follow). For example, if the continuation cost in the combined version is 10 and that for each segment with separation is 6, the savings when there is termination is not half of 10, but only 4.

[102]One source of synergy loss is the cost of making two decisions rather than one when the single decision, despite involving more new information, does not cost as much as the two separate decisions regarding subsets of that information. For ease of exposition, possible incremental decisionmaking costs from separation will not be delineated separately.

depose some witnesses twice would entail large additional costs. Note importantly that this additional cost is borne only with a probability, namely, when there is continuation rather than termination at the interim stage. Hence, a high probability of termination at that point implies both a high option value and a small expected cost in terms of forgone synergies.

Separation can also have important direct and indirect effects on decision quality, which in our framework influences deterrence and chilling. Focusing initially on the stages in question, separation may sometimes involve interim termination when the information that, as a consequence, is never collected would have revealed the case to be strong. If the interim termination/continuation decision is made optimally, this may not be very frequent. Moreover, cases that would thus be terminated but would have been continued if all the information from the two stages had been obtained will not ordinarily be the strongest: more often, if they were below the termination threshold part way through, the stronger showing in the second segment, even if enough to pull them above the decision threshold, will not typically indicate that they are the strongest cases. Hence, those cases that might be viewed as regrettably terminated at the interim stage will tend to be ones that have a poorer contribution to deterrence versus chilling than average.

There are also potentially important indirect effects when we look to other stages in the system as a whole. Changing whether two consecutive stages should be combined or separated will influence how optimal decisions are made at earlier and later stages. For example, returning to the introductory example involving motions to dismiss, if subsequent proceedings will be in smaller steps, it may be optimal to continue more cases at the first stage since some valuable information can now be obtained at relatively low cost, whereas if the subsequent stages were combined, the large continuation cost may have favored a tougher stage-one continuation threshold.[103] Observe that this sort of argument may readily be scenario specific: in some scenarios, stage-one termination may remain optimal; in others, continuation involving only a short segment, followed by reassessment, may make sense; and in still others, perhaps cases initially presenting themselves as strong and that involve nontrivial synergies, continuation without any interim decision may be optimal. In other realms, like our medical decisionmaking illustration, all of these variations are frequently employed, depending on the nature of the ailment and the strength of the preliminary indicators.

This example illustrates the broader and by now familiar point, emphasized in section

---

[103]Anticipating the subject of subsection D.1, structural decisions regarding separation versus combination can also influence filing decisions. For example, a meritorious case that cannot be demonstrated to have potential without some discovery might become more viable if cases proceed in smaller steps, which makes the first-stage screening decision more lenient. Strike suits may be affected in different ways depending on the context. On one hand, smaller steps might make such suits less promising because, after some discovery, as it becomes ever more clear that the case is without merit, termination may be quite likely, whereas if the suit would have appeared just strong enough to survive the first-stage hurdle in a system with full discovery after that, the impositional threat may be stronger. On the other hand, the resulting softening of first-stage screening may make more strike suits viable, at least for awhile. (Some readers may be familiar with the argument that more finely dividing the stages in which litigation costs are incurred may enhance the credibility of negative expected value suits. *See* Lucian Bebchuk, *A New Theory Concerning the Credibility and Success of Threats to Sue*, 25 J. LEGAL STUD. 1 (1996). However, that analysis assumes that there are no interim termination/ continuation decisions.)

II.D, that optimal decisions at any stage depend on how decisions are understood to be made at other stages. Here, if one combines or separates a pair of consecutive segments, that difference in decisionmaking feeds back to earlier decisions and forward to later ones.[104]

Let us now move to our second structural question: taking as given how many steps there should be, in what order should they be undertaken? In particular, examine two consecutive steps, involving the acquisition of different information at each, and consider which should come first. As an initial, rough cut at the problem, it seems that the step with a higher diagnosticity/cost ratio should be earlier — a prescription that is intuitively appealing and undoubtedly characterizes many sequencing choices in other contexts, like medical decisionmaking.

The cost component is most straightforward. In the simplest case, suppose that we have two steps that are expected to be equally informative, but one is lower cost. Obviously, it should come first, for the savings from interim termination is the reduction in the continuation cost from the latter step, and this savings will be larger when the higher-cost step comes second.

Turn now to diagnosticity. For this discussion, assume that the two consecutive steps have equal cost but differ in the information they are likely to generate. In the present setting, what matters most directly is the likelihood that the information learned in the first step will optimally lead to termination at that point.[105] Again, the benefit to be obtained is from avoiding the continuation costs of the next stage, and this arises only when there is a termination. Hence, what matters is not the overall informativeness of what is generated at the next stage of the proceedings but, instead, how likely one will learn sufficiently negative information to justify termination.[106]

Just as with the question of the optimal decision to combine or separate adjacent stages, the foregoing analysis is oversimplified with regard to information and cross-stage interdependencies. With the former, we care not only about the likelihood that the information yielded by whichever stage comes first results in termination at that point but also about how often those decisions would have been regretted (and the welfare consequences thereof) had the information at the other stage been obtained first. For the latter, changes in sequencing will alter what decisions are optimal at other stages. For example, considering motions to dismiss in U.S. civil litigation, if the stage immediately following a continuation decision is inexpensive and

---

[104]To illustrate the latter, additional interim terminations (recall in particular the preceding discussion of regrettable terminations) will tend to reduce deterrence and chilling somewhat, which will tend to make subsequent continuations more beneficial, all else equal. Cutting the opposite way, the resulting cost savings means that the expected adjudication costs of both harmful and benign acts will be lower, which reduces net deterrence gains and raises net chilling costs, making continuation less beneficial.

[105]This idea is the basis for the view that it tends to make sense, if bifurcating, to have a trial on liability precede that on damages since refinement regarding the latter would not obviate the need to determine the former, unless damages are zero (which, if likely, might make the reverse ordering sensible).

[106]If one introduces the possibility of settlement, explored in subsection D.2, then it is possible that nondecisive information will be valuable to the extent that it reduces information asymmetries that otherwise impede settlement (although in such settings the pertinent information may also be revealed voluntarily; *see* Steven Shavell, *Sharing of Information Prior to Settlement or Litigation*, 20 RAND J. ECON. 183 (1989)).

highly informative, then continuation at the first stage will be attractive, whereas if the subsequent stage is expensive and not very informative, then termination at stage one is more likely to be optimal.

The analyses of optimal combination versus separation and of optimal sequencing also have implications for our third question: how to think about the composition of possible stages of legal proceedings.[107] Conventional discussions tend to focus on the use of bifurcation that has particular features.[108] First, attention is often devoted to the conduct of trial whereas much of the savings may be at the pretrial phase, notably, the conduct of discovery in the context of U.S. civil litigation. Second, it is supposed that separation will involve liability and damages or, if liability is divided, that the divisions will track distinct legal elements. If the system is constrained to reach outcomes that, in principle, precisely mimic what would happen with full litigation of all issues, this focus makes sense. However, in a broader optimal decisionmaking framework, this limitation excludes potentially valuable alternatives.[109]

Specifically, it often may make sense to organize staging by type of evidence. With discovery, one might begin with key documents or only a few central witnesses, even if they pertain to multiple issues. There are two virtues of this approach — pertaining to diagnosticity and to cost — that emerge directly from the preceding analysis. Option value is the greatest when more is learned early, and the most informative evidence is not always neatly divided by independent issues. Relatedly, lost synergies from sequencing are smaller if one does not, say, have to depose or call at trial the same witnesses multiple times.

A corollary is that the method of construction of possible stages has a large influence on the optimality of separation versus combination and on optimal sequencing. Taking the last point about costs, suppose that in some contexts one can organize the inquiry into distinct phases such that the synergy loss is close to zero. In that event, the cost-benefit assessment of optimal separation will almost certainly be favorable to employing distinct stages. This point does not imply that such organization (minimization of potential synergy losses) is always best; sometimes it may well be optimal to cluster a small set of evidence that has very high diagnosticity, and to order it first, even at some synergy cost. The clear lesson is that these

---

[107] In strict logic, third question is entailed by the first two. If one begins with each speck of information as a primitive (each type of question that might be asked of each possible witness, and so forth), then by considering all permutations of possible orderings and possible combinations of adjacent fragments, one would also have considered all possible compositions of stages. Nevertheless, it is helpful to make this inquiry into composition separately, particularly since it emphasizes aspects that are often overlooked, certainly in U.S. writing on procedure.

[108] *See, e.g.*, Kötz, *supra* note 28, at 68–69 (illustrating sequencing in Continental procedure with an example having three issues: whether a contract was formed, whether delivered goods were defective, and whether the claim was barred by the Statute of Limitations); Landes, *supra* note 99; Stürner, *Transnational Civil Procedure*, *supra* note 28, at 228 (referring, with regard to proposed transnational principles of civil procedure, to the court determining the order in which issues should be resolved).

[109] That is, we have seen throughout that it often proves to be optimal to make interim decisions without having obtained full discovery and conducted a full trial based on all conceivably relevant evidence pertaining to an independently dispositive legal issue. Similarly, countless medical treatment decisions — overwhelmingly, decisions not to treat or proceed further — are made every day even though there is a conceivable possibility that exhaustive, costly, and possibly painful further testing would reveal an ailment that, ex ante, has a minuscule probability. Even when lives are at stake, decisionmaking systems do not and should not operate otherwise.

decisions are all importantly interdependent. Relatedly, it also appears that they may optimally be made quite differently not only in different areas of law but also in different cases (scenarios). Investigative processes no doubt often have this character, as do, to some extent, less structured types of formal adjudication, such as some arbitration systems and pretrial proceedings in Continental legal systems.[110]

As suggested at the outset of this section, the present analysis helps illuminate potential reforms, such as increasing suggestions in the wake of *Twombly* and *Iqbal* to allow limited discovery in some types of cases before ruling on a motion to dismiss.[111] For example, in a medical malpractice case regarding an operation with a bad outcome, one might initially allow access to all hospital records pertaining to the care of the injured patient and a couple depositions of those in the operating room, but not full access to all information pertaining to surgery at the hospital or all documents and witnesses knowledgeable about procedures with regard to patient care, training, and so forth, any of which might in principle be relevant in establishing negligent practice by the hospital. In an employment discrimination case involving a single employee, one might include internal reviews of the employee and depositions of a supervisor and perhaps an additional key witness but, again, not all internal material pertaining to all employment matters.

Although the impetus for reform might arise from scenarios in which all information is in defendants' possession,[112] many of the ideas on optimal staging developed in this section are not limited to such situations. For example, in cases in which not nearly all information is in a defendant's possession, one might still consider providing access to key documents or witnesses before proceeding to a stage at which expensive expert reports would be prepared. Even in the

---

[110]A related point is that more customized and intricate staging requires a stronger supervisory role, which itself has costs and may help to explain the fact that many Continental systems both have this feature and also a substantially higher ratio of judges to cases. *See* Stürner, *Transnational Civil Procedure*, *supra* note 28, at 227 (suggesting that the trend toward more intensive judicial case management in the United States had been limited by its relatively smaller number of judges); *see also* Easterbrook, *supra* note 98, at 645 ("Perhaps a system in which judges pare away issues and focus investigation is too radical to contemplate in this country — although it prevailed here before 1938, when the Federal Rules of Civil Procedure were adopted. The change could not be accomplished without abandoning notice pleading, increasing the number of judicial officers, and giving them more authority (the system depends on the presiding officer having the power to decide).").

[111]*See also infra* section IV.A (on what *Twombly* and *Iqbal* permit); section IV.D (on how judges might exercise their discretion, including possibly in ways like those suggested by the examples in the text here). Note also that similar de facto outcomes might be achieved in different ways. Suppose, for instance, that judges were understood to make negative inferences from a defendant's failure to voluntarily turn over or provide access to key information that may quickly resolve the matter in their favor if indeed the plaintiff's case had little merit. An early-stage decision standard that was understood to be more lenient in this fashion might induce some defendants — particularly those favored by the facts — to offer such information at the outset. Plaintiffs who thereby obtained access to such information and still lacked factual support would no longer be in a position to complain that their meager pleadings did not reflect a lack of merit but rather that all the information was in defendants' possession.

[112]*Cf.* Stürner, *Transnational Civil Procedure*, *supra* note 28, at 234–35 ("Like their Anglo-American counterparts continental legal systems have always been aware that a party's pleading cannot make the required presentation of detailed facts or offer of detailed means of evidence, if the pleading party has not sufficient knowledge of the necessary facts, especially when the relevant facts occurred in the sphere of the opponent or third persons. In this case a court operating within the continental tradition may permit more general factual assertions and a more general description of a class of documents needed to be produced by the opponent or third persons, and the court may order the taking of evidence without detailed factual or evidentiary contentions. Otherwise, the establishment of the truth would be blocked in a very unfair and intolerable way.").

case of an automobile accident, information from certain witnesses might be obtained and assessed before full-scale accident reconstruction.[113] Of course, there also will exist situations in which one cannot identify a handful of key documents or witnesses, thereby limiting the opportunity for streamlining through thoughtful sequencing.

B. Interaction with Substantive Law

Until now, the substantive law has been taken as given — it involves the prohibition of some set of harmful acts, which now will be denoted as the set X — with the analysis focusing on how thresholds for termination/continuation decisions (including determinations of ultimate liability at the final stage) are optimally set. A more complete analysis would also consider adjustments to substantive law. Suppose, for example, that the existing legal rule prohibiting acts in the set X, enforced by the legal system with given decision thresholds at each stage, appears to err too much in the direction of chilling benign activity. Is it better to address such a problem by making these thresholds tougher or by narrowing the substantive prohibition, perhaps by adding elements or otherwise restricting the acts deemed to be illegal? Relatedly, when are transsubstantive procedural rules optimal (relegating adjustments to substantive law), and what precisely is the meaning of this notion? These questions will be considered in turn.

Beginning with our initial set of harmful acts X that give rise to liability, consider some subset x that is readily identifiable from the outset[114] and seems potentially appealing to carve out, that is, to exempt from liability, leaving only the complementary subset, $X \setminus x$,[115] subject to liability. Specifically, assume that the potential benefit of this strategy derives from two features of the subset x: a large portion of the benign acts that may be subject to liability when the law targets X are readily mistaken for acts in x but not for those in $X \setminus x$, and x does not contain very many harmful acts.[116] In these circumstances, it might seem appealing to confine liability to acts in $X \setminus x$: excluding acts in x greatly reduces chilling effects without much undermining deterrence.

In many situations, however, there will not exist a subset x with these highly attractive features. Suppose instead that we can still identify a target subset x and that this subset is a relatively good candidate for exemption — in that the magnitude of chilling versus deterrence is more adverse than for acts in $X \setminus x$ — but that the difference is only moderate. In these circumstances, instead of exempting acts in x, consider an alternative strategy: toughening the

---

[113]Another important implication concerns the value of a tough initial threshold designed in part to induce prefiling investigation. *See infra* subsection D.1. An offsetting cost is that it may be far more expensive to insist that enforcers undertake great efforts to dredge up information that might cheaply be obtained directly from defendants. Of course, government enforcers are often endowed with powers giving them such direct access. More broadly, providing such access, even to private plaintiffs, may be more appealing if such access can readily be limited to key information rather than making a continuation decision that entitles a plaintiff to wide-ranging discovery.

[114]Often, it may not be obvious whether an act is in the subset x rather than in $X \setminus x$, which situation is readily encompassed by the analysis to follow in the text. This simpler case is examined for ease of exposition.

[115]Formally, $X \setminus x$ is termed the relative complement of x in X, which includes all of the elements in set X that are not also in the subset x.

[116]In lieu of the latter, x may contain mainly harmful acts that pose atypically low net social harm (lower external harm or higher private benefits than the average for X as a whole).

decision thresholds for cases falling in x while loosening them for cases in $X \setminus x$. The assumption that the subset x is readily identifiable means, in the terminology of Part II, that cases in x will present themselves to the tribunal as being in scenarios different from those for cases in $X \setminus x$. Furthermore, the analysis there made clear that the optimal decision rule at any given stage is scenario-specific. In particular, the toughness of the optimal rule for a given scenario at a given stage depends directly and importantly on the magnitude of the deterrence gain and the chilling cost. Therefore, if the decision rule is optimized for each scenario, it will already tend to be tougher on cases in x than on those in $X \setminus x$.

This strategy involving the use of optimal termination/continuation rules will be superior to modifying the substantive law to exempt acts in x, confining liability to those in $X \setminus x$. One way to see this point is to observe that exempting acts in x is tantamount to setting an infinitely tough stage-one decision rule for all scenarios involving acts in x. Now, as just stated, the optimal decision rule for those scenarios will already tend to be tougher than for cases in scenarios pertaining to the set $X \setminus x$. But, as long as some scenarios involving acts in x are ones in which continuation is optimal — which is likely under the stated assumptions — then it is, by definition, better that those cases be continued, not terminated. Suppose, for example, that the acts under consideration are quite harmful, there is a significant deterrence deficit, and there exist some scenarios in which very few benign acts would be confused for acts in x. Then continuation would be highly socially valuable, even if for most scenarios involving acts in x termination is optimal.[117]

Note further that if it did happen to be true that each and every scenario involving acts in x optimally requires termination at stage one, then the procedural approach, if implemented optimally, would necessarily be indistinguishable from the hypothesized modification of substantive law that exempts acts in x from liability. We might think of this extreme situation as that posed at the outset of this section, where the subset x was taken to be a highly favorable target for exemption. Indeed, this trait — that it is optimal to terminate in all scenarios associated with some subset x of the initial set X — should be taken as a precise statement of the conditions for that initial example.

Finally, let us briefly consider a situation even further along the continuum, one in which there is no especially appealing subset x of X to exempt from liability, but exemption of x is nevertheless contemplated because there is believed to exist a serious problem of excessive chilling of benign activity. It is obvious from the foregoing that a superior strategy would be to toughen the termination/continuation rules. That alternative entails removing from the legal system the weakest cases, those in which all the factors, including chilling costs and deterrence benefits, are least favorable to continuation. If one instead exempts an arbitrary subset x, one is

[117] An implication is that, if one was forced to employ a single, perhaps moderate or low, decision threshold to all cases, and thus to all acts in subset x, one would be forced to make a difficult choice: either retain liability for acts in x, which involves excessive chilling costs and modest deterrence gains for many scenarios, or exempt acts in x, thereby suffering the social losses described in the text. Observe further that it would be ironic if, say, the right ultimately to have one's case heard by a jury is understood to require a very low termination/continuation threshold, which in turn motivates the complete elimination of liability in various realms (thereby removing redress for the strong cases as well) in order to avoid the otherwise consequent chilling costs and legal system costs.

removing cases that vary greatly in quality, that is, some cases that are unfavorable to continuation but others for which continuation is quite favorable.[118] It is better not to remove cases essentially at random but instead to concentrate on those that have the worst ratio of benefits to costs.[119] If the system has been designed optimally, this filtering will already be reflected in the existing termination/continuation rules.

Summarizing the lessons from this range of situations, we can see that it is generally superior to adjust termination/continuation thresholds than to exempt subsets of acts from liability. If we had added the further, often realistic point that subsets for exemption are more easily identified in theory than in practice — that is, many scenarios will involve acts in x as well as acts in $X \setminus x$, to varying degrees — then the strategy of exempting subsets of activity from liability will inevitably be implemented imperfectly and thus at greater social cost. Therefore, narrowing substantive law is in principle sensible only when there exists a readily identifiable subset x and when the cases in that subset are particularly unfavorable to liability, although in that instance optimal procedural design would accomplish the same result.[120]

Why, then, do we have circumscribed substantive law? Put another way, why not have just a single legal prohibition that subjects every harmful act to liability, so long as the proof is sufficiently strong under the circumstances? The fundamental justifications relate to various reasons that procedural rules are not completely and optimally adjustable. Important limits may exist due to economies of scale and scope in employing transsubstantive procedure, on which more in a moment. Additionally, optimal adjustments pose great informational challenges (and, attendant administrative costs), as will be emphasized in section IV.C. Relatedly, the very promulgation of substantive law — which defines categories of behavior subject to liability, provides for exemptions, specifies remedies, and so forth — may usefully be viewed as a very important form of information provision, both to adjudicators and to private parties.[121] An

---

[118]One can imagine even worse reforms, wherein the subset x that is targeted for exemption from liability actually has a more favorable ratio of deterrence gains to chilling costs than does the remaining subset $X \setminus x$. This relationship is argued to characterize the approach of many commentators to competition law's prohibition on price fixing. *See* Louis Kaplow, *Direct Versus Communications-Based Prohibitions on Price Fixing*, 3 J. LEGAL ANAL. 449 (2011).

[119]Put another way, in this final situation the fundamental problem is not at all with the definition of acts subject to liability, that is, with the set X, but rather one with distinguishing acts truly in the set X from those that are not. Actually, this is always true unless the proposed carve-out subset x itself consists of benign acts, rather than of harmful acts but ones that many benign acts might be mistaken for.

[120]In practice, it is possible that courts would err substantially in their implementation of the optimal procedural approach but not in their assessments of whether acts are in the subset x, in which case narrowing the substantive prohibition may be superior. Note, however, that if acts in subset x are known by decisionmakers to be poor candidates for continuation, then they could use that very information in making their case-by-case decisions, even in the absence of an absolute exclusion from the prohibition, and this ability may mitigate the problem.

[121]*See generally* Louis Kaplow, *Rules Versus Standards*, 42 DUKE L.J. 557 (1992) (advancing this perspective in analyzing the familiar choice between rules and standards). In the present setting, if some types of harmful acts should almost never be subject to liability, perhaps in large part because there are too many benign acts that might readily be confused with them, it may save substantial resources for those harmful acts to be excluded from any legal prohibition. Also, it often may not be clear to a tribunal which acts are harmful. When one adds that strike suits are possible and that some trial court judges are inclined to procrastinate in ruling on motions to dismiss while not staying discovery, or simply to deny motions the merits of which may be strong, *see infra* section IV.D, the argument for clear exemption is more powerful.

additional set of reasons, constituting yet another generic benefit of employing rules, is to constrain the abuse of discretion.[122] Finally, if for whatever reasons a legal system employs procedural rules, including a termination/continuation rule at the first stage, that do not permit consideration of deterrence and chilling (not even approximately), then limiting the reach of substantive legal rules will tend to be desirable when there are significant concerns with chilling effects as well as system costs and the sacrifice of deterrence is not too great.

Turn now to the question of the desirability of transsubstantive procedural rules, which seem to be widely favored for ease of implementation by judges and the lawyers who practice before them and because of the common view that substance-specific fine-tuning of legal rules is best undertaken by adjusting the substantive rules themselves.[123] The former point obviously has merit, whereas the latter was just seen to be dubious. The Federal Rules of Civil Procedure are widely regarded to be a largely transsubstantive set of legal rules; exceptions, such as Rule 9 on pleading special matters, are few. The legal system as a whole, however, shows much greater deviation. In addition to subject-specific adjustments to the these Rules, such as those contained in the Private Securities Litigation Reform Act,[124] we have bankruptcy courts, a tax court, myriad administrative tribunals (such as for social security disability determinations and immigration cases), and, at the state level, probate courts, small claims courts, housing courts, formal arbitration systems (such as for certain consumer protection claims), some business courts, and so on.[125] And, for the most part, they each operate with customized procedures. In addition, even the administrative merits of generalized procedural rules that do exist can readily be overstated: some lawyers specialize by field in any event,[126] and the seeming simplicity of common rules, such as for discovery, may require that judges repeatedly reinvent the wheel in adapting them to situations that, system-wide, may be quite numerous.

Setting these points to the side, let us focus on the decision criteria for termination/

---

[122]*See* Kaplow & Shavell, *supra* note 33, at 1220–21 n.633 (citing and quoting a wide range of sources); *id*. at 1326–28 (offering a welfare economic perspective on the abuse of power); Frederick Schauer, *The Calculus of Distrust*, 77 VA. L. REV. 653 (examining the role of judicial abuse of discretion in John Hart Ely's constitutional theory).

[123]*See* Miller, *supra* note 14, at 90–94 (discussing the threat of *Twombly* and *Iqbal* to transsubstantive procedural rules and referencing literature on the issue generally). For a view endorsing substance-specific procedure but opposing its creation through judicial interpretation of the Federal Rules of Civil Procedure, see Stephen B. Burbank, *Pleading and the Dilemmas of "General Rules,"* 2009 WISC. L. REV. 535 (advancing the theme that contextual adjustment to pleading rules should be undertaken as a substantive law enterprise); *id*. at 541 ("[T]he Rules Enabling Act's reference to 'general rules' forecloses the promulgation of different prospective rules for cases that involve different bodies of substantive law."); *id*. at 557 (stating, in reference to a requirement of heightened pleading for a specific substantive context, that "the Court seems to have made it impossible for the judiciary to openly impose such a requirement other than through the Enabling Act Process").

[124]Pub. L. No. 104-67, 109 Stat. 737 (codified as amended in scattered sections of 15 U.S.C.).

[125]Part of the justification for separate, specialized systems is the informational challenge addressed in section IV.C. Note further that private dispute resolution systems also are often specialized; for example, different industries' trade associations provide different rules, and within, say, the American Arbitration Association, there are different procedural rules in different fields. *See* American Arbitration Association, *Rules and Procedures*, http://www.adr.org/aaa/faces/rules/searchrules?_afrLoop=273139571257825&_afrWindowMode=0&_afrWindowId=10gi7erfp8_34#%40%3F_afrWindowId%3D10gi7erfp8_34%26_afrLoop%3D273139571257825%26_afrWindowMode%3D0%26_adf.ctrl-state%3Dtkh1xhb9t_4 (last visited Apr. 26, 2012).

[126]And many other lawyers practice in multiple existing systems. Also, even within federal courts, there are detailed local rules that lawyers who appear in many districts need to confront.

continuation decisions (including final assignments of liability) in multistage adjudication. Part II reveals that optimal decision rules depend on numerous factors that vary, often greatly, across different fields of law. Imposing a one-size-fits-all constraint would accordingly be quite costly. Contemplate, for example, averaging chilling costs across discrimination cases, environmental enforcement, antitrust battles, and all manner of contract disputes. Furthermore, the analysis postulated a single type of harmful act, with a specified level of harm — akin to a single subfield of law, or, really, a narrower category of cases — and examined how the many factors varied across different scenarios. That is, just looking at case-specific idiosyncracy, heterogeneity will often be vast.

This Article's analysis nevertheless supports, in principle, the application of a single *algorithm* for decisionmaking in multistage legal proceedings.[127] In that sense, its implications are consistent with a single, transsubstantive set of procedural rules.[128] Therefore, depending on how we look at the question, the same analysis indicates that transsubstantive procedure is significantly suboptimal or just fine. Probing more deeply, this paradox can readily be resolved by recognizing that the distinction between transsubstantive and nontranssubstantive legal procedure is largely illusory, specifically, when considering rules that depend on one or more circumstances of individual cases (rather than being entirely context-independent rules, such as when allowing twenty-one days to file an answer to a complaint in any sort of case).

To elaborate this claim, consider a formally transsubstantive procedural rule, call it T, for the termination/continuation threshold at some stage of adjudication, and suppose that it depends on a single factor, F. We can then write the threshold as a function of the factor: $T(F)$. Now, examine two fields of law, A and B, and let us suppose that the value of the factor F in these fields is a and b, respectively. Therefore, the threshold under our transsubstantive procedural rule is $T(a)$ in field A, and it is $T(b)$ in field B.

Next, compare another legal system with nontranssubstantive procedural rules. Specifically, in this regime the termination/continuation threshold at the stage in question differs in the two fields of law: it is T_A in field A and T_B in field B.

Finally, let us set $T_A = T(a)$ and $T_B = T(b)$. (Or, if we started with the nontranssubstantive version, we could set $T(a) = T_A$ and $T(b) = T_B$.) Obviously, there is no difference between these two legal systems. To restate, once we are contemplating a procedural rule that depends on at least one factor, which in turn differs across substantive fields of law, we can convert transsubstantive procedural rules into ones that differ by area of law, and vice versa.[129]

Returning to the formulations for optimal decision rules presented in Part II, we can ask

---

[127] Actually, as section II.C indicates, the calculus for the final stage is somewhat different, although it can be viewed as equivalent to that for interim stages, but with some values set equal to zero or one.

[128] At this level of generality, for example, even the formulas for motions to dismiss and at summary judgment are the same, although the outputs generally differ in the two settings.

[129] A moment's reflection should make apparent that this point holds for differing and more complex variations of the same point. For example, if the relevant factor F also varies within a field of law, then the nontranssubstantive procedural rule could make that variation relevant as well. Likewise if there are multiple factors.

whether they should be deemed transsubstantive, since the same calculus is applicable to different types of acts involving what would conventionally be viewed as different areas of law, or they should be deemed nontranssubstantive, since they depend on many components, each of which tends to differ across fields of law.[130] This question is one of semantics, not content. Similarly, one could choose to describe the altered pleading rules of the Private Securities Litigation Reform Act as changes in substantive law or in procedural law.[131]

The mixing of procedure and substance has been present from the outset of this Article because the analysis of these conventionally termed procedural rules has very much to do with their effects on behavior, which is the focus of the substantive law that determines liability. It is understood that procedures matter in large part because of their impact on substantive outcomes, and that substantive law may be empty in practice — a mere paper tiger — if the pertinent procedures render it unenforceable. In this light, the conclusion here regarding how to think about transsubstantive procedure is unsurprising.

C. Additional Enforcement Instruments

1. Early-Stage Liability. — The analysis to this point supposes that the choice at any nonfinal stage is between termination and continuation, which reflects the operation of many preliminary and intermediate stages in both informal and formal legal proceedings. However, an additional option, assignment of liability, sometimes exists and in any event is worth considering explicitly in a foundational examination of system design.[132] For example, in U.S. civil litigation, summary judgment is also available to plaintiffs, and indeed its origin is in settings such as debt collection where it was thought possible and important to make speedy, inexpensive legal remedies available in cases in which no serious defense was forthcoming.[133] More broadly, affirmative decisions are often made before collecting complete information. In our medical decisionmaking example, if limited examination or testing makes it apparent that treatment is almost certainly optimal, it will be undertaken without further ado, especially if subsequent information gathering would be costly.[134] And such final, active choices might be made at a preliminary stage (particularly if the treatment is low risk and inexpensive) or at various points along the way, if sufficient evidence accumulates in favor of that course of action.

---

[130]Similar questions arise with the law/fact distinction. The legal category determines whether the rule T_A or T_B applies, but under rule $T(F)$, whether $F = a$ or $F = b$ is a question of fact.

[131]Since the changes do involve pleading rules, which are traditionally viewed as procedural law, that characterization may seem more apt. On the other hand, because passed by Congress and codified as part of the securities laws rather than promulgated through the special process for procedural rules and made part of the Federal Rules of Civil Procedure, they might be viewed as substantive.

[132]This point is reinforced if one adds that some cases involve affirmative defenses, counterclaims, and alternative framings (either party might sue first in many contract skirmishes; there may be actions for declaratory judgment), all of which give reason to attend less to the particular labeling of the parties and to view adjudication as more symmetrical.

[133]*See, e.g.*, FED. R. CIV. P. 56, advisory committee's note (to the original 1937 enactment); Charles E. Clark & Charles U. Samenow, *The Summary Judgment*, 38 YALE L.J. 423 (1929); Shapiro, *supra* note 18, at 359–61; Steinman, *supra* note 3, at 88–90.

[134]In this example, costs may include pain and reduced efficacy through delay as well as direct expenditures, which have analogues in adjudication as well.

Moreover, the intuition that earlier imposition of liability is sometimes optimal tracks that applicable to termination at nonfinal stages. Specifically, the benefits of further refinement resulting from additional information acquisition may not be worth the costs. Accordingly, it should be unsurprising that the analytical framework from Part II can readily be adapted to cover the possibility of early findings of liability.

To simplify the exposition and because the analysis of early termination has already been presented in sections II.A and II.B, let us focus only on the choice between imposition of liability and continuation.[135] Compare, now, the decision to continue *against a benchmark of concluding the case at the current stage with an assignment of liability*. Relative to that baseline, continuation involves a deterrence cost: rather than imposing sanctions for certain, continuation implies that they will only be imposed with a probability.[136] Likewise, there is now a chilling gain: although continuation does impose liability in some instances, we are comparing that outcome to imposition of liability with certainty; hence, continuation reduces the expected cost for benign acts. Finally, as before, we have continuation costs. Accordingly, in contrast to our prior formula, continuation is optimal if and only if the following inequality holds:

Chilling Gain > Deterrence Cost + Continuation Costs

To restate: continuation entails a benefit (reduction) with regard to chilling but a cost with respect to deterrence and also an increase in adjudication costs.

The determination of the relevant subcomponents of each of these three elements is almost precisely analogous to that in section II.A, so the previous discussion will not be replicated.[137] One difference is that, as suggested by the preceding paragraph, the calculation of the changes in the chilling and deterrence effects is somewhat different. Before, we had the probability of subsequent liability as a determinant of the increase (from a probability of zero, which arises from termination) in expected sanctions for each type of act, harmful and benign. Here that probability constitutes a decrease, from a probability of one, that arises from the

---

[135]One can imagine that early termination is not an option, or simply that cases that are close calls on the liability/continuation margin will not also be close on the termination/continuation margin. The latter rationalization does omit the important situation in which continuation costs are sufficiently high and the value of additional information sufficiently low that it makes sense to either assess liability or exonerate, with no further consideration of continuation. Combining all the foregoing analysis with this possibility generates the idea that the very notion of having some final stage would in principle make sense precisely when there was no plausible scenario in which the optimal decision would be to put off the choice in order to collect further information. In many systems of decisionmaking, one may not know in advance where that point will be but instead conclude, upon examination of the latest additional wave of information, that such a point has been reached. In that sense, such a final stage may occur at different points in different scenarios — that is, depending on what information has been revealed thus far.

[136]It is logically possible for continuation to be more, not less costly for a defendant because, although the sanction is not imposed with certainty, additional continuation costs are borne, and the latter could exceed the former savings. Note that, in such instances, defendants would prefer to concede liability.

[137]Likewise, many of the subtleties would arise in essentially the same manner. The optimal decision rule in the scenario under consideration will depend on how decisions are made in other scenarios, and similarly for the interdependence of decisions across stages. Introducing early impositions of liability changes expected adjudication costs, which influences the net impact on social welfare per deterred and chilled act, and so forth.

immediate assignment of liability. As before, one would use that change in probability to calculate the change in the expected sanction for each type of act, which in turn would allow one to determine the magnitude of the change in the number of acts chilled and deterred, and so forth.

Given the symmetry of the problem and the relevant analysis, as well as the widespread use of early affirmative decisions in many important decisionmaking settings, including some parts of the legal system, it is interesting to inquire into why early terminations, informal and formal, are so much more prevalent in many legal settings than are early assignments of liability. Probably the explanation with the widest application concerns institutional competence. For example, the police, prosecutors, and many agencies do not have the authority to impose sanctions without the decision of an independent body.[138] Even judges are often distinct from final decisionmakers, such as in the U.S. legal system when that power has been vested in a jury (although U.S. judges are often constrained even when there will be no jury)[139] and in Continental systems when there is a single examining judge whereas the decision will be rendered by a panel of judges.[140] The analysis of this section highlights an important cost of this functional separation. Obviously, there are benefits as well, the analysis of which is beyond the scope of this Article (although a few elements will be touched on in Part IV).

The existence of such institutional constraints suggests the relevance of a hybrid approach under which further interim proceedings are short-circuited at an optimal point, as indicated by the preceding analysis, in order to bring the case immediately to the authorized final decisionmaker. On reflection, it is apparent that this practice is routine: police, prosecutors, and agencies do not continue to collect additional information, at a real cost, indefinitely; nor do they necessarily follow a predefined series of steps. Instead, they often collect and analyze information sequentially, terminating some cases along the way and, when they have accumulated what seems to be a sufficiently powerful mass, initiating formal proceedings before the requisite tribunal. In Continental systems, an examining judge does not collect further evidence without limit, either on direct initiative or at the parties' behest, but at some point deems the matter to be ready for a final decision by the tribunal.[141]

---

[138]There are some compromises, such as when a violation is defined to contain fewer elements or when irrebuttable presumptions are employed (which can amount to the same thing), which narrows the scope of the choice before the final decisionmaker. In addition, some systems employ mixed approaches. For example, one may have to pay a traffic fine or a tax penalty imposed by the pertinent government authority unless one files a challenge before a tribunal.

[139]*See also infra* note 286 (further discussing the extent to which the constitutional guarantee of trial by jury constrains pretrial decisionmaking).

[140]Interestingly, systems of alternative dispute resolution often vest substantial discretion in the decisionmakers regarding the conduct of proceedings, yet in many systems early-stage dispositive rulings for either party are infrequent or, on many issues, not entertained, although this may be changing. *See, e.g.*, THOMAS J. STIPANOWICH & PETER H. KASKELL, EDS., COMMERCIAL ARBITRATION AT ITS BEST: SUCCESSFUL STRATEGIES FOR BUSINESS USERS — A REPORT OF THE CPR COMMISSION ON THE FUTURE OF ARBITRATION 203–06 (2001); STEPHEN J. WARE, PRINCIPLES OF ALTERNATIVE DISPUTE RESOLUTION 103–04 (2d ed. 2007). In addition, tribunals may signal their evolving views to parties in order to induce settlement, and, given the wider use of discretionary cost-shifting, these signals may sometimes have substantial force.

[141]That tribunal might disagree and examine further evidence, but this point does not fundamentally change the character of the process, and that inquiry as well is finite.

2. Enforcement Effort.[142] — Tradeoffs among deterrence, chilling, and system costs are not only influenced by termination/continuation decision rules (or, from subsection 1, liability/continuation decision rules) at various stages of legal proceedings. Most obviously, they are also determined by the level of enforcement effort, considered in this subsection, and the magnitude of sanctions, examined in the next. To simplify the discussion, enforcement effort will be treated as if it is under the control of the state; subsection D.1 on the initiation of cases will address briefly how, when private suits are involved, the state's control will be indirect, such as by altering filing incentives with fee-shifting rules and the like.

A helpful thought experiment in comparing enforcement effort and continuation rules is to contemplate changes that keep deterrence constant. For example, we could imagine raising enforcement by some amount, which would increase the fraction of acts that enter the legal system, and simultaneously raising the toughness, say, of the stage-one termination/continuation decision rule in an amount that (together) generates the same level of deterrence as before. Note initially that both adjustments would also influence chilling and system costs. The increased flow would increase chilling whereas the tougher continuation rule would reduce it. And the direct effects of raising enforcement effort would be to increase system costs whereas the reduced continuation rate would reduce them. (We know from Part II that there are also indirect effects on system costs. These ideas also apply to enforcement effort: its effects on deterrence and chilling change the number of acts and thus expected adjudication costs as well.)

In examining this experiment, we would like to know whether the increased chilling cost from greater enforcement effort is larger or smaller in magnitude than the reduced chilling cost from the tougher continuation rule.[143] (And likewise for the differential impact on system costs, the comparison of which will depend on the particulars of enforcement technology and on what is involved with continuation at the stage or stages where the rules are toughened, as well as at subsequent stages.[144]) In some settings, it seems that the net effect would be favorable. For example, if enforcement is through random audits, greater effort might correspond to a higher audit rate, which would increase the fraction of harmful acts and of benign acts that enter the legal system by the same proportion. By contrast, a tougher termination/continuation threshold would eliminate the weakest cases from among those that previously were continued.[145] Combining the two effects, the average strength of cases in the legal system that are continued would rise. If deterrence is being held constant, this implies that chilling will fall. Hence, the

---

[142] *Cf.* Kaplow, *supra* note 30, at 815–19 (comparing enforcement effort and the burden of proof); sources cited *infra* note 146 (comparing enforcement effort and accuracy).

[143] Throughout this subsection and the next, it will be assumed that chilling a marginal act is net socially costly, whereas it was explained in note 61 that it is possible that chilling at the margin is desirable, which arises when the forgone benefit from a marginal benign act is less than the expected system cost generated from a benign act.

[144] As mentioned, any net change in chilling will feed back on total system costs in a now familiar manner. Because deterrence is held constant in our hypothetical construction, there is no feedback effect from it.

[145] The argument that follows in the text assumes that the weakest cases tend to be ones that are worse on the merits — concretely (but roughly), scenarios with a lower ratio of harmful to benign acts. However, as the formulation for the optimal decision rule in subsection II.A.2 indicates, weakness (for a given deterrence effect) depends also on system costs. Therefore, it is possible that there are marginal scenarios where the chilling cost is not very high but, instead, the continuation costs are large. Of course, it remains true that marginal scenarios are marginal scenarios, which is to say that their overall benefit-cost ratio is worse than average for scenarios that optimally involved continuation.

proposed experiment would advance social welfare if the net chilling gain exceeded the net system cost increase (if any) from substituting more enforcement for more generous continuation decisions.

Increased enforcement effort can also operate on other margins.[146] For example, many audits (which include various forms of inspections) are not conducted entirely randomly. Instead, they are prioritized: targets that look most suspicious or for other reasons are known to be more likely to involve harmful rather than benign acts are examined first. Similarly, with investigations, cases with the best leads are pursued first. In such settings, raising enforcement effort will have diminishing returns with regard to case quality: marginal cases will be weaker than those higher on the targeting list. This pattern is qualitatively like that for the termination/ continuation decision: as one relaxes that threshold, one is retaining weaker and weaker cases. In this sort of situation, whether increased enforcement effort combined with a tougher termination/continuation decision — calibrated to keep deterrence constant — reduces chilling effects will depend on the comparison of these two margins. When enforcement effort is very low and the decision rule quite lax, one would expect the experiment to reduce net chilling relative to the situation in which enforcement effort is already high and the termination/ continuation threshold is strict.[147] And, whatever the net chilling effect, it would need to be combined with the net effect on the costs of operating the legal system (preliminary enforcement and adjudication) to determine whether the hypothetical adjustments were an improvement.

The latter comparison can usefully be restated entirely within the framework from Part II by taking the enforcement effort decision simply as an earlier, additional stage in the multistage legal system — we could call it stage zero. At that point, there is an initial indicator (the scenario), reflected perhaps by an audit prioritization score or a summary of the leads to date. The decision is whether to continue (to conduct the audit, undertake further investigation) or to terminate. Continuation will generate further information and involve a direct cost. One then arrives at the next stage, what was until now viewed as the first stage in our legal proceedings. Alternatively, if the audit or investigation has already been completed and the only question is whether to move the case forward, into the formal legal system, we likewise could view that as a termination/continuation decision.[148] In either situation, the preceding formulation of the question — should we employ greater enforcement effort combined with a tougher stage-one

---

[146]The main additional margin not examined in the text is that one could also raise deterrence through enforcement effort by increasing the intensity and hence the accuracy of the enforcement apparatus. For example, one might employ more talented auditors and investigators, or give them more time per case. Accuracy has received some attention in prior work. *See, e.g.*, Louis Kaplow & Steven Shavell, *Accuracy in the Determination of Liability*, 37 J.L. & ECON. 1 (1994) (examining accuracy versus enforcement rates in a setting with only one type of act, so that tradeoffs involve deterrence and system costs, but not chilling); Kaplow, *supra* note 33, at 345-62 (same); Kaplow, *supra* note 30, at 825–29 (informally examining the tradeoff between greater accuracy and a tougher burden of proof); Louis Kaplow, *On the Optimal Burden of Proof* 28–29 (Nat'l Bureau of Econ. Research, Working Paper No. 17765, 2012) (formally examining that tradeoff).

[147]These statements are explicitly relative because there could be contexts in which enforcement effort was superior to continuation rule adjustments across a very broad range, or conversely.

[148]If continuation was costless, we could simply collapse that decision with our stage-one decision; if additional types of cases are included, we would simply contemplate additional possible scenarios. If there is instead some cost — for example, the cost of the stage-one decision itself (which has been abstracted from for simplicity until now) — then viewing the decision as an additional, preceding stage would be appropriate.

termination/continuation rule — would be stated as whether we should lower the stage-zero threshold and raise the stage-one threshold. Section II.D has already analyzed this sort of comparison in broad terms, drawing on the analysis of sections II.A – II.C.[149]

3. Sanctions.[150] — This subsection first analyzes sanctions as an additional enforcement instrument and asks, in our basic framework, how changing sanctions compares to adjustments to termination/continuation rules with regard to the tradeoff among deterrence, chilling, and system costs. Second, the subsection examines how various forms of sanction costs may be incorporated into the basic framework.

On the first topic, we can consider an experiment parallel to that in subsection 2, now raising sanctions (if they are not already at the maximum feasible level) and toughening, say, the stage-one termination/continuation decision so as to keep deterrence constant. As before, there would tend to be a net favorable impact on chilling effects because the heightened sanctions apply to all cases in the system that ultimately are subject to sanctions whereas the tougher continuation rule tends to eliminate the weakest, which is to say, those with an above-average ratio of benign acts to harmful acts.[151] In addition, unlike greater enforcement effort, increasing sanctions (still, for the moment, taken to be monetary sanctions that are costless to impose) does not raise system costs, whereas the tougher stage-one rule saves continuation costs. This tendency of higher sanctions to be a favorable enforcement strategy arises in many of the settings examined in prior literature.[152] Interestingly, the present context has the feature that the concern for mistaken imposition of sanctions on benign acts may actually strengthen the case for higher sanctions. The reason — implied by the preceding argument that the chilling effect would fall — is that, although benign acts are sanctioned more heavily, the tougher continuation decision means that they are not only sanctioned less often but that this latter effect is relatively larger.[153]

On the second, recall that until this point sanctions have been taken to be monetary payments — fines or damages — that are socially costless to impose. Monetary payments per se

---

[149]It should be evident that the analysis from Part II and the sketch in the preceding text in this subsection are, on reflection, equivalent. (That presented here was much more abbreviated, because the ideas are now familiar.)

[150]*Cf.* Kaplow, *supra* note 30, at 819–24 (comparing sanctions and the burden of proof).

[151]Another benefit of relying more on sanctions than more generous continuation rules to achieve a given level of deterrence is that the former tend, in an important sense, to be better targeted. Sanctions are actually imposed only in cases that survive all continuation decisions and then result in liability. By contract, more generous continuation rules not only raise the likelihood of explicit sanctions but also raise defendants' adjudication costs, the latter of which are borne even in cases that do not ultimately result in a sanction. A greater fraction of latter, compared to cases that do result in imposition of the sanction, involve benign rather than harmful acts.

[152]The seminal paper that compares the tradeoff between higher sanctions and enforcement effort is Gary S. Becker, *Crime and Punishment: An Economic Approach*, 76 J. POL. ECON. 169, 183-84 (1968). The application of the traditional Becker argument in the present context is examined in Kaplow, *supra* note 41. For a survey, which also addresses reasons not to raise sanctions, see A. Mitchell Polinsky & Steven Shavell, *The Theory of Public Enforcement of Law*, *in* 1 HANDBOOK OF LAW AND ECONOMICS 403, 413-16, 431-34 (A. Mitchell Polinsky & Steven Shavell eds., 2007).

[153]Note that if, indeed, the net chilling effect is negative from the substitution, this means that individuals contemplating benign acts would face a lower expected cost of committing their benign acts. Hence, ex ante, they would all prefer the regime with higher sanctions and tougher continuation rules (although individuals who are chilled regardless would be indifferent between the two regimes).

are transfers: the prospect of payment contributes to deterrence, and also to chilling, but the payment itself is not socially consequential.[154] Now assume instead that sanctions are socially costly, such as with imprisonment, where individuals' sacrifice of liberty is a pure loss and, moreover, prisons are costly to operate.

Introduction of sanction costs alters the analysis in two ways. Most obviously, in our formulas in Part II, continuation costs are now larger — and, in the formula in section II.C for the optimal final-stage rule, we now add a continuation cost where there was none previously. As one would have anticipated, this addition makes continuation less attractive. Sanction costs, however, also raise the deterrence gain and reduce the chilling cost from continuation. These components, recall, include, for each act deterred or chilled, the savings in expected system costs from fewer acts entering the legal system. This savings per act is now greater because, in addition to avoiding adjudication costs, social sanction costs are also spared with respect to acts that are discouraged. Because these competing effects depend on different factors (recall the many subcomponents of each that are identified in subsection II.A.2), either could be greater a priori.

Because the foregoing logic applies equally to harmful acts and to benign ones, some additional lessons are implied. For example, if sanctions for harmful acts were per se beneficial (that is, a negative social sanction cost, or at least a benefit that partly reduces the direct cost) — perhaps on account of incapacitation effects or retributive value — the implication for optimal termination/continuation rules would be ambiguous. Likewise, if the imposition of sanctions on individuals who actually committed benign acts is additionally costly, either tougher or weaker continuation rules could be optimal. This last point may seem surprising, but it follows if the hypothesized social cost applies to the actual application of sanctions to benign acts: with tougher continuation rules, these mistakes happen less often per act that enters the legal system, but because of the reduced chilling effect, more benign acts enter the system, which itself increases the number of benign acts that might be sanctioned.[155]

---

[154]Even with monetary payments, however, there are additional social consequences if individuals are risk averse. *See, e.g.*, Polinsky & Shavell, *supra* note 152, at 414–16. In the case of actual harmful acts, liability would be beneficial to the extent it provides compensation to risk-averse victims. However, it also adds corresponding costs if actors are risk averse. Moreover, for benign acts, the prospect of compensating individuals who were not truly victims (consider feigned injury) is also negative with regard to risk imposition (a lawsuit is like a lottery ticket: a risky prospect worth less than its expected value). Regarding all, insurance and diversified ownership often reduce substantially the importance of risk aversion and compensation. Finally, note that risk aversion also is relevant to behavior: specifically, risk-averse individuals will be deterred and chilled more, especially when sanctions are significant relative to their wealth (so, for example, in the analysis of subsection II.A.2, both the preexisting degree of deterrence and chilling, that is, with termination, would be higher, and the contribution of continuation to deterrence and chilling would be larger).

[155]In contrast, from the perspective of individuals who contemplate committing benign acts, tougher rules are favorable in that they reduce the expected costs associated with such acts, even though those who ex post are unlucky are still worse off. Of course, if such individuals also suffer the adverse consequences of harmful acts and pay taxes to fund the legal system, their interests will align more closely with the overall social welfare calculus employed throughout this Article.

D. Endogeneity of Cases

1. Initiation.[156] — In the setting articulated and analyzed in Part II, the fractions of harmful acts and of benign acts that enter the legal system and thus present themselves at the first stage was taken as given. In this Part, subsection C.2 considered how it may be optimal to adjust enforcement effort, raising or lowering those two fractions. To a substantial degree, however, the decision to initiate cases is endogenous, determined by actors in the system.

For private suits, this point is straightforward. A prospective plaintiff will tend to sue when its expected recovery exceeds the expected adjudication costs it bears, all of which will reflect not only what is predicted to happen at trial but also the likelihoods of moving through each stage of litigation.[157] Public enforcers face different incentives, but nevertheless their behavior is hardly mechanical. Depending on their motivation — whether for immediate compensation, prospects for promotion, possibilities of reprimand, reputation that may influence future employment, or personal pride — expected system outcomes may well influence decisions to initiate cases.[158]

Although the particulars may differ in important ways, a fair generalization is that, in many instances, the prospect of more generous continuation decisions will induce cases to be filed at a greater rate. Accordingly, consider how that effect changes the analysis of what termination/continuation criterion is optimal.

Most obviously, the deterrence and chilling punches of continuation are both increased. Before, we had an increment to actors' expected costs from committing harmful and benign acts as a consequence of their being more likely to bear sanctions as well as there being higher expected defendants' legal costs per case. Now we must add that, for a given act, a regime more generous toward continuation in various scenarios at various stages means that there is a greater likelihood that one's act will enter the legal system in the first place. That too raises the expected cost of committing either type of act and hence the increment to deterrence and to chilling that results from continuation rather than termination. The consequence of these

---

[156] *Cf.* Kaplow, *supra* note 30, at 848–55 (examining how adjustment of the burden of proof influences the initiation of cases).

[157] *See, e.g.*, BONE, *supra* note 32, at 139–46. There are a number of ways in which more generous termination/continuation rules induce additional private suits. In addition to the most direct explanation — a greater expected recovery makes more suits viable — there can also be influences related to the timing of possible terminations and the sequential imposition of costs. For example, a more generous stage-one rule that enabled a plaintiff to impose large discovery costs on the defendant, even in instances in which termination or ultimate defeat after discovery was certain, may enable the plaintiff to extract a larger settlement, the prospect of which makes the suit viable even though the expected value of the suit assuming no settlement remains less than the plaintiff's expected litigation costs. *See, e.g.*, David Rosenberg & Steven Shavell, *A Model in Which Lawsuits Are Brought for Their Nuisance Value*, 5 INT'L REV. L. & ECON. 3 (1985); David Rosenberg & Steven Shavell, *A Solution to the Problem of Nuisance Suits: The Option to Have the Court Bar Settlement*, 26 INT'L REV. L. & ECON. 42 (2006); *see generally* Spier, *supra* note 32, at 268–80, 305–07 (surveying literature on settlement and on negative-expected-value claims). Settlement is discussed further in subsection 2.

[158] As suggested in subsection C.2, when analyzing how legal system design may change enforcement effort, we can imagine that many of the instruments will be indirect. For example, if additional filings are desirable, one might provide means for compensating private plaintiffs for legal fees (fee-shifting, subsidization) or for supplementing public enforcement by allocating more resources to increase staffing. This subsection inquires into how termination/continuation decisions influence system inflow when such features are taken as given.

phenomena, however, is ambiguous regarding the optimal decision: the deterrence gain will be greater, which favors continuation, and the chilling cost will be larger, which favors termination. Which effect will dominate obviously depends on the numerous subcomponents of each and how those may differ for the additional marginal cases induced by the more generous rule.

There is an additional set of effects, on expected system costs. On one hand, because the fraction of cases that enters the legal system rises, this means that the expected total costs of legal proceedings for each act committed also rise. On the other hand, the aforementioned increase in deterrence and chilling reduces the number of cases that potentially enters the legal system and thus arises in each scenario. Even supposing, for example, that the net effect is to increase the number of cases in all scenarios, we know from the analysis in section II.A that the implications are ambiguous: continuation costs rise, favoring termination, but greater expected system costs make deterrence more valuable and chilling less costly, favoring continuation.

Viewing these effects together, the influence of an induced increase in the case filing rate on the optimal rule will obviously depend on the circumstances. Suppose, for example, that there is a large deterrence deficit, due in part to the fact that many meritorious cases do not enter the legal system because their prospects of success are low. In that event, the additional increment to deterrence could be especially valuable in two respects: the social gain per additional act deterred would be large and the average quality of cases drawn into the legal system may be high. By contrast, if deterrence is already robust and the additional cases filed as a consequence of more a generous termination/continuation rule are mostly meritless, the main additional effects would be greater chilling and system costs.

To expand on this latter point, it is useful to focus on the scenario-specific responsiveness of case filings to anticipated termination/continuation decisions. Suppose, for example, that the scenario in question is one in which a broad swath of benign acts might readily be characterized as falling. In that event, allowing continuation in that scenario may induce a large number of meritless cases to be filed, substantially augmenting both chilling and system costs.[159] If, instead, a scenario is one that covers many harmful acts whereas few benign acts might even arguably be alleged to fit within it, then the prospect of continuation would primarily enhance deterrence.[160] We can see that, even if more generous continuation decisions in general were to

---

[159]Such a view of cases involving stock price drops shortly after disclosures (including in connection with initial public offerings) seems to have been a central motivation for the Private Securities Litigation Reform Act of 1995. *See infra* note 228. Likewise, the Supreme Court in *Twombly* seemed worried that if firms' decisions not to enter others' markets were sufficient proof of conspiracy to survive a motion to dismiss and reach discovery, large numbers of strike suits might be filed. Similarly, in *Iqbal*, the Court may well have been concerned about how readily others could file civil rights suits that would impose significant costs on high government officials.

[160]It is also useful to recall the theme of subsections C.2 and C.3: greater enforcement effort and stronger sanctions are substitutes for laxer continuation thresholds in achieving deterrence. When it may be compelling to terminate in many scenarios because of the concern for strike suits, it may be possible to make up some of the deterrence deficit by devoting greater resources to investigation and other enforcement activity and also to heighten sanctions. When concentrated on the reduced subset of scenarios that involve a higher proportion of harmful acts, the deterrence-chilling tradeoff will be more favorable. This strategy assumes that there are not important subsets of harmful acts that prospective violators can anticipate will fall in the scenarios involving termination, for higher sanctions on other acts that are expected to present themselves in different scenarios cannot make up for the deterrence deficit associated with such violations.

induce similar average filing responses, the elasticities for harmful versus benign acts may differ greatly across scenarios.[161]

Filing incentives are not the only channel through which litigation behavior can be influenced by termination/continuation rules. Parties may also adjust the intensity of their efforts, notably, to develop evidence, for cases that are in the system.[162] Whether their incentives will be stronger or weaker when continuation rules are more generous is hardly obvious. In some settings, greater prospects of success may induce enforcers to invest more; perhaps defendants, in response, will match these greater efforts, or perhaps they will subside if their prospect of success is becoming hopeless. In other instances, we may expect the opposite: more generous rules may make enforcers' prospects for success high enough that it is no longer worthwhile to press as hard to succeed, and we can imagine varied defendants' reactions here as well. In sum, there may be important implications for parties' conduct of litigation, but it seems difficult to draw any clear, general lessons regarding how this consideration bears on optimal termination/continuation rules.

There is, nevertheless, one aspect of litigation incentives that has drawn particular attention. Especially at the first stage — the point of initial screening — it may seem attractive to lean against continuation when the information presented is thin in order to induce greater prefiling efforts by enforcers. The reasoning behind this supposition, however, requires further scrutiny. First, it matters how expensive it is to obtain information directly (for example, by trying to locate and interview former employees of the defendant) versus through formal legal processes, such as discovery. On pure cost grounds, particularly when information is primarily in defendants' possession, it is not obvious that encouraging self-reliance by enforcers is efficient. Indeed, in many settings government agents are given powers to obtain information directly from targets of investigations based on little if any preliminary demonstration of cause, which they then employ in deciding whether to initiate a formal legal action. In considering how to make such investigative decisions optimally, these inquiries can simply be viewed as stage zero of legal proceedings (recall subsection C.2), and the analysis of Part II is applicable.

Second, imposing a tougher continuation threshold regarding initial investigation, particularly when private plaintiffs may be in a position to know whether their cases are valid, may induce useful sorting. Even if relatively expensive, investigative avenues that will likely prove fruitful may only be available (or may be more likely to bear fruit) for those with meritorious claims. If this is true, strike suits may become relatively less attractive, which will decrease overall system costs and chilling effects, even if the system cost for truly meritorious cases is increased. Note that, in this instance, both incentives to generate evidence and to file cases are endogenous, and in a manner that is interrelated.

---

[161]The extent to which this is true will depend in significant part on whether enforcers know the scenario at the time of initiating a case, which for the first stage they typically would.

[162]It is unclear whether stronger incentives to win always lead to the development of truly better evidence or evidence that, when a decisionmaker considers what is offered by both parties, leads the tribunal to a more accurate assessment. Even taking an optimistic view, it seems that there will often be diminishing returns, at least after some moderate point, regarding the contribution to accuracy. In addition, one must take into account the costs of developing additional evidence.

These ideas are also relevant to public enforcers, even though their incentive structures differ. One point of particular note is that we are sometimes concerned about various forms of abuse: governments may seek to target political opponents, or officials with aspirations to fame (possibly because they later expect to run for higher office) can direct formidable government resources at targets that will generate publicity.[163] One set of protections is high proof standards at trial, but it is often possible to inflict great damage and achieve substantial personal gain even if a case is ultimately dropped, whether or not a guilty plea has been obtained. Such concerns underlie the origins of the grand jury, which is enshrined in the U.S. Constitution for infamous crimes and dates to Magna Carta. In addition, civil investigations, which can inflict significant financial and reputational costs, can be ill-motivated or nevertheless pursued for more mundane reasons, with limited external constraint.[164]

2. Settlement. — In addition to influencing parties' filing choices and investigative efforts, termination/continuation rules may also affect settlement decisions, including plea bargains. Negotiated case closures also influence what decision rules are optimal. Let us begin with the latter.

Suppose first that settlement values approximately equal the expected sanction; perhaps each parties' expected litigation costs are similar and they settle at the midpoint in the bargaining range. This seemingly neutral outcome nevertheless decreases deterrence and chilling relative to a no-settlement benchmark because, recall, actors' aggregate expected costs include not only expected sanctions but also their own expected adjudication costs, which settlement reduces. In a world in which settlement is more frequent, therefore, the deterrence deficit will be somewhat larger and the cost of chilling the marginal act somewhat lower, both of which favor continuation, all else equal.[165] Another possibility is that settlements in some settings would be skewed relative to the expected sanction: for example, if defendants face asymmetrically large litigation costs — perhaps due to discovery, for cases reaching that stage — then settlements would tend to be for more than the expected sanction. Put another way, when the prospect of continuation without settlement would impose high costs on defendants and when settlement amounts tend to reflect that fact, the contribution of continuation to both deterrence and to chilling will be maintained to that extent.[166]

---

[163] *See* sources cited *supra* note 122.

[164] *See, e.g.*, Epstein, *supra* note 16, at 207–12 (criticizing the potential for overly intensive antitrust investigations using civil investigative demands).

[165] As explained in subsection II.A.2, it is also true that the number of acts deterred and chilled per unit change in their respective expected costs may differ, and in either direction.

[166] If it is believed that, even though ultimate decisions on the merits will be fairly accurate, early continuation decisions advance large numbers of meritless suits, then it may be that the primary determinant of deterrence is the expected sanction whereas the primary contributor to chilling is expected adjudication costs, a perspective that seems implicit in some past concerns expressed with regard to lenient first-stage rules, that is, readily denying motions to dismiss. *See, e.g.*, Bell Atlantic Corp. v. Twombly, 550 U.S. 544, 557–58 (2007) ("We alluded to the practical significance of the Rule 8 entitlement requirement in *Dura Pharmaceuticals, Inc. v. Broudo*, 544 U. S. 336 (2005), when we explained that something beyond the mere possibility of loss causation must be alleged, lest a plaintiff with '"a largely groundless claim"' be allowed to '"take up the time of a number of other people, with the right to do so representing an *in terrorem* increment of the settlement value."' *Id.*, at 347 (quoting *Blue Chip Stamps v. Manor Drug Stores*, 421 U. S. 723, 741 (1975)). So, when the allegations in a complaint, however true, could not raise a claim of entitlement to relief, '"this basic deficiency should . . . be exposed at the point of minimum expenditure of time and money by the parties and the

Settlements also tend to reduce system costs that, as we saw at many points in Part II, will have an ambiguous impact on optimal termination/continuation rules. On one hand, expected continuation costs are lower when there is a prospect that cases will subsequently settle, which makes continuation more attractive. On the other hand, the expected system costs of undeterred and unchilled acts are lower, which reduces one of the benefits of deterrence and the offset to the cost of chilling benign acts, and this effect makes continuation less attractive. Taking everything together, particularly since in some legal systems most cases are eventually settled, settlement may have a significant influence on optimal decision rules, but it is difficult to say a priori whether this consideration favors tougher or more lenient thresholds.

The matter is more complicated because of the reverse consideration: termination/ continuation rules influence settlement behavior, which generates feedbacks that influence what rules are optimal. Most obviously, the rules will affect settlement amounts, taking as given whether settlements take place. For example, the prospect of more generous continuation decisions will raise settlement values. These effects, however, are essentially mediated versions of the originally postulated effects. In the extreme, if there was a one-to-one translation of expected outcomes for defendants to settlement amounts, much of the analysis of Part II would be unchanged.

Termination/continuation rules may also affect whether cases settle. Here again there are competing effects. For example, more generous continuation raises both parties' expected adjudication costs, which makes settlement more attractive for each. Cutting in the other direction, potential disagreements on the ultimate outcome may be of greater significance because it is more likely that final adjudication will be reached.[167] Accordingly, it is difficult to say in general how the feedback effects from settlement behavior influence the optimal toughness of termination/continuation rules.[168] Yet another factor is that effects on settlements — frequency and amounts — will influence decisions to initiate cases, the subject of subsection 1.

---

court."""); *id*. at 559 ("And it is self-evident that the problem of discovery abuse cannot be solved by 'careful scrutiny of evidence at the summary judgment stage,' much less 'lucid instructions to juries,' *post*, at 4; the threat of discovery expense will push cost-conscious defendants to settle even anemic cases before reaching those proceedings."). Such concerns were an important motivation for enactment of the Private Securities Litigation Reform Act of 1995, Pub. L. No. 104-67, 109 Stat. 737 (codified as amended in scattered sections of 15 U.S.C.). *See infra* note 228.

[167]Similar logic applies to subsequent nonfinal decisions. Put another way, more generous continuation at a given, nonfinal stage tends to raise stakes with regard to expected sanctions, so disagreements (asymmetric information) with respect to such outcomes will become more important. *See, e.g.*, Lucian Bebchuk, *Litigation and Settlement Under Imperfect Information*, 15 RAND J. ECON. 404, 409–10 (1984); I.P.L. P'ng, *Strategic Behavior in Suit, Settlement, and Trial*, 14 BELL J. ECON. 539, 546 (1983).

[168]An additional source of complexity arises from the fact that the parties' information may differ from the decisionmaker's information. This possibility, in turn, may influence which subset of cases that otherwise would present themselves in some particular scenario at a given stage in the legal proceedings will settle and which will remain. If the decisionmaker could infer how it differed, this different case mix would influence the optimal decision rule. Moreover, the anticipation of any such effect would in turn influence parties' settlement behavior, all of which makes optimal rule formulation even more challenging.

E. Regulation of Future Conduct

Until now, all the analysis addresses settings in which, aside from its direct costs of operation, the legal system matters because of its effects on behavior: deterrence and chilling. In other settings, however, the focus of legal proceedings is on what future conduct will be permitted, such as with licensing, zoning, drug authorizations, merger approvals, and many injunctions.[169] In such instances, the analysis is more straightforward, and the analogy to medical decisionmaking (really, the classic valuation-of-information problem from decision analysis), noted at various points throughout,[170] is quite close.[171]

For convenience, consider a two-stage system, and let us inquire when it makes sense at the first stage to continue when the optimal action the decisionmaker would take, as currently informed, would be to allow the activity to take place. (When the default would be prohibition, the analysis would change in an obvious manner, as mentioned below.) Here, continuation will raise social welfare when the expected gain from prohibiting actually harmful acts at stage two exceeds the sum of the expected cost from mistakenly prohibiting benign acts at that stage and the system costs of continuation itself.

To decompose the first two components further, note that the welfare effects from continuation, aside from the direct continuation costs that are borne regardless of the stage-two outcome, arise only in those instances in which a different decision would be made. We can, as in section II.A, take the stage two decision rule as given (whether it is optimal or operates in some other fashion). At that stage, the additional information learned as a consequence of continuation will be employed, and sometimes it will lead to prohibition. When prohibition does occur and it is indeed a harmful act that is prohibited, we have a prohibition gain, which equals the product of the likelihood that this outcome occurs and the net social gain per prohibited harmful act. The latter will be the harm avoided minus the forgone benefit from the act.[172] When it is a benign act that is prohibited, we have a prohibition cost, which equals the product of this likelihood and the net social loss per prohibited benign act, which is simply the forgone benefit from such an act.

The basic determinants of when continuation is optimal are straightforward and intuitive, much more so than for our basic case. Here, high diagnosticity and low cost (a high diagnosticity/cost ratio) favors continuation. And, as already stated, the measure of diagnosticity

---

[169]When the prospect of the latter induces settlement that substitutes a payment of damages, the result may be closer to the sort of liability considered previously.

[170]*See, e.g.*, *supra* note 37.

[171]For prior, more extensive analysis of the difference in settings (but not examining multistage decisionmaking), see Kaplow, *supra* note 30, at 837–48 (examining the distinction when analyzing the burden of proof), and Kaplow, *supra* note 33, at 369–81 (exploring the contrast when addressing accuracy in adjudication).

[172]By contrast to subsection II.A.2, we do not have the additional reduction in system costs from fewer acts entering the legal system because, in the pure case under analysis, we are assuming that there are no deterrence effects. Also, here the magnitude of the forgone benefit from the act will be the average benefit for this class of acts (not the benefit for a marginal act, for the latter is appropriate only with deterrence, wherein actors themselves, who know their private benefits, decide which acts to commit). This same point applies to the forgone benefit when benign acts are prohibited by mistake.

concerns decisions that would be changed, in this instance relative to our benchmark of authorization at stage one. If, instead, the stage one decision, without continuation, would have been prohibition, then we would have an authorization gain when an actually benign act is permitted as a consequence of what is learned upon continuation, and an authorization cost when an otherwise prohibited harmful act is permitted by mistake.

Furthermore, we could reassess other subjects addressed in our basic case. For example, we could contemplate multiple nonfinal stages, compare the final-stage decision rule,[173] consider whether adjoining stages should be kept separate or combined, and determine how stages should be ordered. The greater simplicity of the present setting — particularly the lack of behavioral feedbacks — both eases such analysis and, in general, may alter what is optimal.

This section's analysis is streamlined in a number of respects, mainly in being confined to a simple, polar case when in fact many situations are intermediate. For example, the prospect of drug authorization versus prohibition also has important ex ante effects, namely, on incentives for research and development. In such situations, one would need to aggregate the two types of analysis, that in the present section and that in Part II. For example, the prospect of prohibition would have deterrence effects — discouraging the development of dangerous and ineffective drugs — and chilling effects — discouraging the development of beneficial drugs. To determine the optimal rule at any stage, one would add these benefits and costs of prohibition to those examined earlier in this section regarding future conduct and continuation costs.

IV. APPLICATIONS AND IMPLICATIONS

Parts II and III present a conceptual analysis of multistage legal proceedings. This Part applies the framework to elucidate existing rules and discourse, with a focus on U.S. civil procedure for concreteness and due to the author's familiarity.[174] That said, the object remains one of clarifying thought, expanding perspective, and raising alternatives rather than establishing correct legal interpretations or advocating particular reforms.

Section A considers possible interpretations of *Twombly* and *Iqbal*'s plausibility standard for deciding motions to dismiss and how it might relate to the decisionmaking criterion developed in this Article. The next three sections elaborate additional dimensions of this relationship: Section B delves further into the nature of facts, which appear to be central to current formulations of pretrial decision rules as well as key elements of Part II's apparatus. Section C emphasizes the informational challenges confronting decisionmakers in applying all but the most minimalist and formalistic versions of these rules and discusses possible responses. Section D highlights the extent of judicial discretion in pretrial decisionmaking and comments on how decisionmakers might choose to exercise it. Finally, section E turns to the summary

---

[173] *See* Kaplow, *supra* note 30, at 839–43.

[174] The latter is qualified in two respects: my not being a specialist in the field and the fact that important aspects of the actual conduct of even formal legal proceedings has an informal, discretionary quality that existing scholarship does not sharply illuminate.

judgment standard.

A. Motion to Dismiss

Twombly and *Iqbal* hold that a complaint must be plausible in order to survive a motion to dismiss under Rule 8 of the Federal Rules of Civil Procedure.[175] This section addresses possible meanings of this standard and how they might relate to this Article's analytical framework for decisionmaking at preliminary stages of a multistage legal system.[176] Although at first glance there may seem to be little connection, upon examination there is appreciable affinity.

Interpretation of the plausibility requirement is challenging for two reasons. First, the Court in *Twombly* and *Iqbal* does not substantially elaborate the concept. Second, having clearly rejected that the test is merely one of logic (referring merely to whether a plaintiff's claim is possible or conceivable),[177] the Court also seems to eschew another natural meaning, one

---

[175]This section considers only the plausibility standard, which may not be implicated in many motions to dismiss (and the many more potential motions not filed) because it is readily met or clearly fails under any likely interpretation or because the motions involve jurisdiction or other purely legal matters.

[176]Among the issues not considered is the extent to which the plausibility standard departs from prior understandings or practice. *See, e.g.*, POSNER, FEDERAL COURTS, *supra* note 32, at 180 (writing before *Twombly*, states: "Increasingly, district judges, sometimes abetted by court of appeals judges, dismiss a complaint because it fails to allege facts critical to the plaintiff's claim . . . ."); *id*. ("More important, district judges are increasingly prone to evaluate . . . complaints . . . as if [they] were a summary of evidence. If the judge is not impressed . . . he dismisses the suit . . . . And this irregular practice the courts of appeals are increasingly inclined to condone too."); 5B CHARLES ALAN WRIGHT & ARTHUR R. MILLER, FEDERAL PRACTICE AND PROCEDURE § 1357 (3d ed. 2004) (writing before *Twombly*, states: "In more recent years, however, a number of federal courts, as has been true with summary judgment motion practice, have been more willing to dismiss under Rule 12(b)(6), particularly in certain substantive contexts such as securities litigation."); Bone, *supra* note 12, at 3; Christopher M. Fairman, *The Myth of Notice Pleading*, 45 ARIZ. L. REV. 987, 988 (2003) (arguing, pre-*Twombly*, that "notice pleading is a myth" because "substance specific areas of law," including antitrust law, environmental law, conspiracy law, and copyright law, "are riddled with requirements of particularized fact-based pleading"); Richard L. Marcus, *The Revival of Fact Pleading Under the Federal Rules of Civil Procedure*, 86 COLUM. L. REV. 433 (1986). For preliminary empirical evidence on the effects of *Twombly* and *Iqbal*, see the sources cited in note 9. Another subject set to the side is the consistency of the current standard, however it may differ from what existed previously, with the Seventh Amendment jury right. *See, e.g.*, Suja A. Thomas, *Why the Motion to Dismiss Is Now Unconstitutional*, 92 MINN. L. REV. 1851 (2008); *infra* note 286 (discussing the question in the context of motions for summary judgment and judgment as a matter of law).

[177]*See, e.g.*, Bell Atlantic Corp. v. Twombly, 550 U.S. 544, 570 (2007) ("Because the plaintiffs here have not nudged their claims across the line from conceivable to plausible, their complaint must be dismissed."); Ashcroft v. Iqbal, 129 S. Ct. 1937, 1950–51 (2009) (quoting some of this language from *Twombly*); *id*. at 1951 ("To be clear, we do not reject these bald allegations on the ground that they are unrealistic or nonsensical. . . . It is the conclusory nature of respondent's allegations, rather than their extravagantly fanciful nature, that disentitles them to the presumption of truth."). Relatedly, both opinions' emphasis on facts strongly suggests that the test is not just a matter of logic.

In addition, the *Twombly* Court did make clear that prominent language of *Conley v. Gibson*, 355 U.S. 41 (1957), had "earned its retirement." 550 U.S. at 563. To glean a better understanding of what was being rejected, which presumably informs what was embraced, it is helpful to review the pertinent passage from *Conley*.

> In appraising the sufficiency of the complaint, we follow, of course, the accepted rule that a complaint should not be dismissed for failure to state a claim unless it appears beyond doubt that the plaintiff can prove no set of facts in support of his claim which would entitle him to relief.

355 U.S. at 45–46. In addition to the oft-mentioned "no set of facts" language that the Court explicitly references in *Twombly*, note that the passage contains the phrase "beyond doubt" (not even the familiarly qualified "beyond a reasonable doubt"). Taken together, this language suggests an extremely low floor with regard to probability, implying

involving a threshold probability, although this latter pronouncement is more equivocal.

To examine the probability question, begin with *Twombly*'s key statement: "Asking for plausible grounds to infer [a violation] does not impose a probability requirement at the pleading stage; it simply calls for enough fact to raise a reasonable expectation that discovery will reveal evidence of [a violation]."[178] The initial clause seems to constitute a definitive rejection of a probabilistic interpretation, yet one wonders if this statement is not taken back by the subsequent reference to a reasonable expectation. That is, while the Court purports to hold that there does not need to be any probability of a violation, it nevertheless appears to demand that there be a sufficient probability that evidence of a violation will be found if the case proceeds. (Compare: there need not be any probability that oil lies below, but there must be a decent expectation that our pilot hole will find evidence of oil there.[179]) This puzzling passage on probability is immediately followed by the further statement that "a well-pleaded complaint may proceed even if it strikes a savvy judge that actual proof of the facts alleged is improbable, and 'that a recovery is very remote and unlikely.'"[180] This clarification suggests either that the requisite probability is extremely low or that even an essentially zero probability suffices. But the latter understanding is inconsistent with the Court's many contrary statements, including its rejection of *Conley*'s formulation.[181]

---

that the pleading requirement was entirely one of logic or mere possibility (i.e., existence). For example, it is possible that lightening will strike twice, even five times, in a particular place over a short period of time, although unlikely in the extreme. One might still say that it is beyond doubt that no plaintiff could prove this, suggesting that the probability floor is above zero, even if barely so. Either way, the rejection of *Conley*'s phrasing does make clear that logical possibility, by itself, is insufficient.

Another challenge for interpretation concerns the possible interplay with the notice function of pleading, which by itself seems uncontroversial even if unclear in its requirements. For example, one can hardly say that, beyond doubt, the complaint "X is liable for injuring me" (where X might be some individual, organization, or government entity) could not be supported by any set of facts, yet it is deficient on notice grounds. Although not the focus of *Twombly* and *Iqbal*, notice concerns may underlie some of what moved the Court in these cases. For example, in *Twombly*, the Court was bothered by the broad, vague possibility of an unspecified conspiracy spanning seven years, which contributed greatly to the concern that discovery would be difficult to cabin. 550 U.S. 560 n.6 (quoted *supra* note 13); *see also infra* note 197 (addressing the interaction of the notice function of pleading and the Court's distinction between allegations of facts and of legal conclusions). A separate problem is that it is difficult, for example, for a defendant to move for summary judgment after discovery when the basis for liability is unclear, because it then cannot be known very well which facts need to be undisputed. As a doctrinal matter, it would seem helpful to have a clearer understanding of the relationship between notice and plausibility.

[178]550 U.S. at 556. Regarding the centrality of this passage and the next, the sentences immediately follow the paragraph's topic sentence containing the phrase "we hold."

[179]As just stated, the Court's second clause does not, in our analogy, require discovery of oil but only evidence of oil. The difference is that there could be false or misleading evidence of oil (or a violation) in addition to true indications thereof. But it would be odd to imagine that this difference was central to the rule — that is, that plaintiffs need not show a probability of a true violation but only a probability of either a true violation or that they will come upon misleading evidence regarding a phony violation. Another distinction is that a probability of evidence of a fact may refer to a lower likelihood than is meant when referring to a probability of the fact itself. Nevertheless, a lower probability is still a probability, and since little clue is given as to the magnitude intended by either reference, merely knowing that one is below the other provides little guidance as to the threshold.

[180]550 U.S. at 556 (quoting Scheuer v. Rhodes, 416 U. S. 232, 236 (1974)).

[181]*See supra* note 177. Another puzzle is that the complaint in *Twombly* alleged a conspiracy that may not have been supported by many particular allegations but almost surely does not fall below the level of being merely improbable, remote, and unlikely. *See infra* notes 220 & 221. Hence, the application of the standard to the facts of the case seems to belie the notion that there is not some sort of nontrivial probability floor.

The *Iqbal* Court makes specific reference to *Twombly*'s holding, encapsulated in these sentences, as follows: "The plausibility standard is not akin to a 'probability requirement,' but it asks for more than a sheer possibility that a defendant has acted unlawfully."[182] It too, in the same sentence that rejects that plausibility means probability, seems then to ask for what amounts to a probability after all. What else could be meant by "more than a sheer possibility"? That is, on what dimension is the Court asking for more, if not with regard to the allegations' likelihood of being true? (Compare: I am not requiring that there be any probability of striking oil, but there's no way we are going to spend all that money on a test hole unless you can tell me that there is more than a sheer possibility of success.) Other language in *Iqbal* likewise suggests the relevance of probability to plausibility.[183]

In addition, the incorporation of a probability factor seems difficult to avoid. Suppose, for example, that a medical malpractice complaint is accompanied by an affidavit from an impeccable expert stating that, in light of all available information, the probability of negligence is X%. Surely, if X equals 0.0001% (one in a million), the case would be dismissed under the plausibility standard, but if X equals 99.9999%, the motion to dismiss would be denied. It follows, therefore, that in this type of situation, there should exist some probability above which the case would be continued and below which it would be terminated.[184]

Perhaps modest guidance can be gleaned from standard definitions of the word "plausible." These definitions cluster on a handful of meanings.[185] One pertains to superficial

---

[182] 129 S. Ct. at 1949 (quoting *Twombly*).

[183] "Taken as true, these allegations are consistent with petitioners' purposefully designating detainees 'of high interest' because of their race, religion, or national origin. But given more likely explanations, they do not plausibly establish this purpose." *Id*. at 1951; *see id*. at 1950 (stating that *Twombly* found the complaint inadequate because the conduct alleged "was more likely explained by" competitive behavior). That is, plausibility fails because alternative explanations are "more likely." Hence, this passage is not merely a probability requirement, but one notably stronger than suggested by the *Twombly* passages quoted in the text to the effect that even allegations that are improbable, remote, and unlikely may be sufficient. Specifically, it seems to require that the plaintiff's explanation be the most likely one, or at least tied with any alternative. (Note that such comparative statements are, on reflection, strange. If there are 10,000 explanations, one with a likelihood of 0.02% might be the single most likely, whereas if there are only two possible explanations, one with a likelihood of 49.9% would not be.) A similar sort of requirement is imposed under the Private Securities Litigation Reform Act in *Tellabs, Inc. v. Makor Issues & Rights, Ltd.*, 551 U.S. 308, 324 (2007), holding that "[a] complaint will survive . . . only if a reasonable person would deem the inference of scienter cogent and at least as compelling as any opposing inference one could draw from the facts alleged." *See also* Ronald J. Allen & Alan E. Guy, Conley *as a Special Case of* Twombly *and* Iqbal*: Exploring the Intersection of Evidence and Procedure and the Nature of Rules*, 115 PENN ST. L. REV. 1, 35 (2010) ("The Court concluded that the factual allegations were insufficient, which sounds like a factual judgment about their probability, which it is, but again the Court denied this. . . . The logical contradiction is evident and the conclusion obvious that 'plausibility' incorporates a 'probability requirement.'"); *id*. at 37 ("Rather obviously, one cannot at the same time rationally dispense with a 'probability requirement' to determine 'plausibility' yet conclude that something is not 'plausible' because there are other 'more likely explanations.' No sense can be given of 'more likely' except 'more probable.'"); Clermont & Yeazell, *supra* note 10, at 833 (referring to "plausible" as an "unavoidably probabilistic standard").

[184] As a matter of strict logic, if the initial two cases are decided as suggested and yet this proposition is false, then there must exist two possible cases such that, with all else being the same, there is termination in one with a higher probability than that in another in which there is continuation.

[185] In advancing these definitions, presented without quotation marks, I am drawing on a number of readily available dictionaries (Merriam-Webster's Collegiate Dictionary, 11th ed., CD-ROM; Dictionary.com; M-W.com; MacmillanDictionary.com; and OED online). Because the differences are modest across many sources, it seems best to

appearance, recognizing the possibility of deceit. In a vacuum, this could be understood as the pertinent meaning. Indeed, this could be viewed as equivalent to a minimalist legal requirement, such as what some associated with *Conley*. That is, a complaint must, on its face, appear to state a violation, granting that the impression may be false. If logical coherence was sufficient, this definition of plausible might fit best. Yet it seems clear that this demand, although necessary, is not sufficient.[186]

Other definitions of plausible, not surprisingly, move more in the probability direction (for example, likely to be true). In particular, some definitions suggest the requirement of a reasonable probability, which aligns with *Twombly*'s "reasonable expectation" language. More specifically, some of these sorts of definitions suggest fitness for a purpose, such as forming a belief or engaging in some action (for example, serious consideration for a job).

It is notable that this concept of plausibility combines probability and consequences. That is, in some instances a fairly low probability might make a decision plausible: one should have a flu shot even if the chances of becoming ill without one are only ten percent. In other instances, a somewhat higher probability may suffice: perhaps a thirty percent chance of rain justifies bringing along an umbrella. And in other cases, an extremely high probability may be necessary: one would not cancel a trip to a best friend's wedding unless it was essentially certain not to take place. In sum, any rational decision rule that is concerned with consequences will depend on their probabilities, but it is equally true that such a test will not depend only on probabilities (that is, and not also on the consequences).

This interpretation seems to fit rather well with most statements in both opinions.[187] The reasonable expectation language specifically refers to the prospect that discovery will generate supportive evidence. Furthermore, the notion that the Court envisions a broad balancing test that is attentive to consequences is suggested both by the two cases' emphasis on the costs of discovery — financial and otherwise — and by the rejection of a pure probability requirement, which might have suggested that there was a one-size-fits-all minimum likelihood, without regard to competing considerations such as cost.

Perhaps the strongest language indicating that the plausibility inquiry directs courts to examine every aspect of the particular case before deciding whether, all things considered, a motion to dismiss should be denied, is *Iqbal*'s directive in interpreting *Twombly*: "Determining whether a complaint states a plausible claim for relief will . . . be a context-specific task that

---

present the core variations without making any particular phrasing canonical.

[186]As a matter of clarity in communication, it is unfortunate that the Court chose as its key term one having as a standard definition the very notion it meant to reject.

[187]As expressed at the outset of Part IV, the claims advanced here are not meant to be definitive, but more as provocative or suggestive of possibilities. Hence, the present assertion and the supporting evidence are not presented to demonstrate that the proffered interpretation is the only available one, the one actually on various Justices' minds, and so forth.

requires the reviewing court to draw on its judicial experience and common sense."[188] That is, plausibility depends on the specifics of the context, and the inquiry is one based on experience and common sense. The latter, in turn, is generally understood to involve practical judgment, a mix of both logic and an understanding of how the world works. One with common sense attends to probabilities but does not fixate only on them, disregarding the magnitude of consequences.

If the Supreme Court in *Twombly* and *Iqbal* did indeed have in mind a rather open-ended, case-specific balancing test, this would help explain much of what it did state, including that quoted in the preceding paragraphs and notes, and also some of what it did not. If such was the intent, it would not make sense to articulate a precise, canonical, and perhaps necessarily formalistic definition.[189] One would not assert that plausibility was simply a probability test, but one also would not disclaim the relevance of probabilities. And in applying the test, one would adduce whatever factors seemed most relevant to making an all-things-considered judgment.[190] Of course, this sort of legal test embodying a pragmatic approach, while hardly ubiquitous, is entirely familiar.

By this view, the plausibility requirement is some sort of reasonableness or balancing test that attends to likelihoods and costs. *Twombly* and *Iqbal*, however, do not come close to enumerating the relevant factors or indicating how they are properly weighed in making a judgment, leaving a large gap regarding how motions to dismiss ought to be decided. Indeed, one might view the proffered test as so open-ended as to be empty, essentially question-begging: When should a motion to dismiss be denied, thereby allowing the case to proceed to discovery? Answer: when a complaint is plausible. And when is a complaint plausible? Answer: when it is reasonable to allow the case to proceed to discovery.[191]

The framework of this Article may be useful at this point, for it indicates what factors are

---

[188]129 S.Ct. at 1950. Consider also: "It is true that Rule 9(b) requires particularity when pleading 'fraud or mistake,' while allowing '[m]alice, intent, knowledge, and other conditions of a person's mind [to] be alleged generally.' But 'generally' is a relative term." *Id*. at 1956. Presumably, the relativity of the term "generally" refers to the context of the particular case. And, although *Twombly* did not refer specifically to "judicial experience," it instead, in an antitrust case, refers to "common economic experience." 550 U.S. at 565. (Interestingly, in describing the manner in which Continental judges make decisions regarding the conduct of cases, Rolf Stürner refers to "judicial discretion and common sense." Stürner, *Transnational Civil Procedure*, *supra* note 28, at 220.)

[189]From this perspective, one would not expect commentators' attempts to propose some alternative phrasing of the official test, *see* sources cited *supra* note 16, to succeed in offering a silver-bullet solution to the test's ambiguity.

[190]As mentioned in the Introduction, it seems difficult from many perspectives to explain the seemingly great fixation with precisely what the complaint states. Perhaps this attention may best be understood as focusing on what facts a court is to have in mind in defining the scenario in which to make the plausibility assessment.

[191]Consider also *Iqbal*'s concluding statement of the "working principles" that underlie *Twombly*: "But where the well-pleaded facts do not permit the court to infer more than the mere possibility of misconduct, the complaint has alleged — but it has not 'show[n]' — 'that the pleader is entitled to relief.' Fed. Rule Civ. Proc. 8(a)(2)." 129 S. Ct. at 1949–50. *See also Twombly*, 550 U.S. at 557 ("The need at the pleading stage for allegations plausibly suggesting (not merely consistent with) [a violation] reflects the threshold requirement of Rule 8(a)(2) that the 'plain statement' possess enough heft to 'sho[w] that the pleader is entitled to relief.'"). This purportedly workable criterion, in asking whether the complaint contains enough to show an entitlement to relief, patently assumes that we already know what must be demonstrated for the plaintiff to be so entitled.

relevant, and how so, if the object is to conduct multistage legal proceedings so as to promote social welfare. As mentioned in the Introduction, Rule 1's statement of the objectives of the Federal Rules of Civil Procedure, which are to be the basis for construing and applying these rules, are loosely in accord with this purpose. Furthermore, although the Supreme Court did not specifically articulate a set of objectives, there can be little doubt that much of what it said in *Twombly* and *Iqbal* (in the majority opinions and the dissents) reflects a concern for whether meritorious cases will succeed, benign activity will be excessively encumbered, and system costs will be extravagant.[192] Hence, without by any means suggesting that the plausibility test is tantamount to the decision rules developed here, some harmony in aims seems to exist, and the Court's test appears to be of a sort that directs judges to attend to pertinent consequences rather than to employ a decision rule that is formalistic, purely logical, or focused on some target probability that is independent of the context.

B. Nature of Facts

Facts are central to legal decisions not only in final adjudication but also, at least under current doctrine in U.S. civil litigation, at preliminary stages. Both *Twombly* and *Iqbal* focus intensively on the adequacy of the facts alleged in complaints,[193] and the standard for summary judgment under Rule 56 concerns the existence of a "genuine dispute as to any material fact." Yet cases and commentary are insufficiently precise, sometimes inconsistent, and occasionally confused about the nature of facts, which inhibits analysis, understanding, and the ability to

---

[192] *Cf.* Tellabs, Inc. v. Makor Issues & Rights, Ltd., 551 U.S. 308, 322 (2007) ("Our task is to prescribe a workable construction of the 'strong inference' standard, a reading geared to the PSLRA's twin goals: to curb frivolous, lawyer-driven litigation, while preserving investors' ability to recover on meritorious claims.").

[193] *See Twombly*, 550 U.S. at 548–49 (requiring sufficient "factual context" to suggest illegality); *id*. at 555 ("Factual allegations must be enough to raise a right to relief above the speculative level . . . ."); *id*. at 555 n.3 ("The dissent greatly oversimplifies matters by suggesting that the Federal Rules somehow dispensed with the pleading of facts altogether."); *id*. at 556 (requiring "enough factual matter"); *id*. at 557 n.5 (stating that the line determining the adequacy of a complaint "lies between the factually neutral and the factually suggestive"); *id*. at 561–62; *id*. at 563–64 n.8 ("[B]efore proceeding to discovery, a complaint must allege facts suggestive of illegal conduct."); *id*. at 569; *Iqbal*, 129 S. Ct. at 1942–43 ("This case instead turns on a narrower question: Did respondent, as the plaintiff in the District Court, plead factual matter that, if taken as true, states a claim that petitioners deprived him of his clearly established constitutional rights."); *id*. at 1946 (stating that a complaint's adequacy "cannot be decided in isolation from the facts pleaded."); *id*. at 1948–49 ("[T]o state a claim . . . , respondent must plead sufficient factual matter . . . ."); *id*. at 1949 ("To survive a motion to dismiss, a complaint must contain sufficient factual matter, accepted as true, to 'state a claim to relief that is plausible on its face.'" (quoting *Twombly*, 550 U.S. at 570)); *id*. ("Where a complaint pleads facts that are 'merely consistent with' a defendant's liability, it 'stops short of the line between possibility and plausibility of "entitlement to relief."'" (quoting *Twombly*, 550 U.S. at 557)); *id*. at 1950 ("While legal conclusions can provide the framework of a complaint, they must be supported by factual allegations."); *id*. at 1951 ("We next consider the factual allegations in respondent's complaint to determine if they plausibly suggest an entitlement to relief."); *id*. at 1952 ("He would need to allege more by way of factual content . . . ."); *id*. ("Yet respondent's complaint does not contain any factual allegation sufficient to plausibly suggest petitioners' discriminatory state of mind."); *id*. at 1954 ("But the Federal Rules do not require courts to credit a complaint's conclusory statements without reference to its factual context."); *id*. ("We hold that respondent's complaint fails to plead sufficient facts to state a claim for purposeful and unlawful discrimination against petitioners."). A complication is that the Court's per curiam opinion in *Erickson v. Pardus*, 551 U.S. 89, 93 (2007), issued shortly after *Twombly*, states: "Specific facts are not necessary; the statement need only '"give the defendant fair notice of what the . . . claim is and the grounds upon which it rests."' *Bell Atlantic Corp. v. Twombly* . . . (quoting *Conley v. Gibson* . . .)." *See also supra* note 177 (discussing the interplay between the plausibility requirement and the notice function of pleading); *infra* note 197 (addressing the relationship between the notice function and the Court's unwillingness to credit pleadings of ultimate facts).

articulate rules and methods for decisionmaking. Accordingly, this section explicates some basic questions and distinctions, including the relationship between facts and evidence and the inputs that might provide the requisite information or ground the necessary beliefs that are at the core of existing tests. These matters are especially important regarding decisions on motions to dismiss due to the juxtaposition of, on one hand, the centrality of facts and the aforementioned relevance of probabilities under the plausibility test and, on the other hand, the conventional supposition that evidence need not be presented and should not be evaluated at this preliminary stage.

The distinction between a fact — something true about the world — and evidence — something that furnishes proof of a fact — is familiar. In a simple auto accident case, an allegation that the light was red is an assertion of fact, whereas reference to Jones, a bystander, who purports to have seen that the light was red, is an identification of evidence of that fact. Whether Smith was motivated by racial animus in taking an action is a question of fact; an email in which Smith uses a racial epithet in discussing the action is evidence of the fact. Also, there often exist intermediate facts, such as Smith having acted differently when dealing with a similar situation except for differences in race (itself something to be established by evidence), that might be a basis for inferring the ultimate fact. In light of this distinction, insistence on a statement of facts (notably, regarding allegations in a complaint) does not in itself constitute a demand for evidence.[194]

This conclusion, however, is substantially misleading — that is, as long as allegations based on pure fantasy are impermissible. Proffering a fact entails holding some basis for acceptance of the fact as true, which in turn requires possessing some evidence. For example, Rule 11(b) refers generally to "knowledge, information, and belief," and subsection (3) explicitly requires that "factual contentions have evidentiary support or, if specifically so identified, will likely have evidentiary support after a reasonable opportunity for further investigation or discovery."[195] Knowledge presupposes some evidentiary basis, information is tantamount to evidence, and belief in the truth of a fact presupposes evidence.[196] As a consequence, although insistence on particular facts is not per se a demand for evidence, for all practical purposes it requires evidence. Perhaps the evidence need not be stated, at all or in detail, but it is presumed to exist in some quantum. (If someone told you "the light was red" and, when pressed for the basis for the assertion, stated that there was none whatsoever, you would deem the speaker to be either playful or a liar, depending on the context.)

---

[194] *See, e.g.*, AREEDA & HOVENKAMP, *supra* note 71, at 101 (stating that, although a court should ensure that a plaintiff has "pl[ed] its claim," there is no requirement that the plaintiff "plead its evidence"); Steinman, *supra* note 10, at 1339 ("The statement does not need to provide any kind of evidentiary support."). Even before considering the point in the text to follow, however, the matter is murkier than meets the eye. The reason is that one central meaning of the term evidence includes any fact from which one might infer a belief; by this definition, any relevant fact is also evidence. But in the legal setting, evidence is generally understood to refer to admissible testimony or documents that gives one reason to believe a fact, intermediate or ultimate.

[195] FED. R. CIV. P. 11(b)(3).

[196] There are some exceptions of limited relevance. For example, humans possess some instinctive understanding about basic elements of how the world works (or at least an innate capacity to produce such knowledge), and one may sincerely hold beliefs attributable to divine inspiration, but such alternative bases for belief are not taken to be significant for present purposes.

Moreover, it seems clear that *Twombly* and *Iqbal* demand more than this minimum to survive a motion to dismiss. Specifically, it is not sufficient to plead the ultimate fact (often a legal element in a claim), such as discriminatory intent as in *Iqbal*.[197] Pleadings must also contain some basis for believing the ultimate fact rather than accepting innocent explanations for the behavior in question. Such basis should be understood as requiring evidence. As already suggested, one type of evidence for an ultimate fact is another fact or set of facts from which one may infer the ultimate fact. It may well be that the evidentiary basis for such intermediate facts

---

[197] *Iqbal*, 129 S. Ct. at 1949 ("First, the tenet that a court must accept as true all of the allegations contained in a complaint is inapplicable to legal conclusions. Threadbare recitals of the elements of a cause of action, supported by mere conclusory statements, do not suffice."); *id*. at 1950 ("In keeping with these principles a court considering a motion to dismiss can choose to begin by identifying pleadings that, because they are no more than conclusions, are not entitled to the assumption of truth. While legal conclusions can provide the framework of a complaint, they must be supported by factual allegations."); *id*. at 1951, 1954; *see Twombly*, 550 U.S. at 555 ("While a complaint attacked by a Rule 12(b)(6) motion to dismiss does not need detailed factual allegations, . . . a plaintiff's obligation to provide the 'grounds' of his 'entitle[ment] to relief' requires more than labels and conclusions, and a formulaic recitation of the elements of a cause of action will not do, see *Papasan v. Allain*, 478 U.S. 265, 286 (1986) (on a motion to dismiss, courts 'are not bound to accept as true a legal conclusion couched as a factual allegation')."); *id*. at 557 ("[A] conclusory allegation of agreement at some unidentified point does not supply facts adequate to show illegality."); *id*. at 557 n.5; *id*. at 564 (describing the complaint's statements as "merely legal conclusions").

At some points, however, the *Twombly* Court is fuzzier on the distinction. Notably, the dissent argues that, "As relevant, the Form 9 complaint states only: 'On June 1, 1936, in a public highway called Boylston Street in Boston, Massachusetts, defendant negligently drove a motor vehicle against plaintiff who was then crossing said highway.' Form 9, Complaint for Negligence, Forms App., Fed. Rules Civ. Proc., 28 U. S. C. App., p. 829 . . . . The asserted ground for relief — namely, the defendant's negligent driving — would have been called a '"conclusion of law"' under the code pleading of old." 550 U.S. at 576 (Stevens, J., dissenting). The majority responds: "This lack of notice contrasts sharply with the model form for pleading negligence, Form 9, which the dissent says exemplifies the kind of 'bare allegation' that survives a motion to dismiss. *Post*, at 576. Whereas the model form alleges that the defendant struck the plaintiff with his car while plaintiff was crossing a particular highway at a specified date and time, the complaint here furnishes no clue as to which of the four ILECs (much less which of their employees) supposedly agreed, or when and where the illicit agreement took place." 550 U.S. 565 n.10. This response is remarkable because it seems implicitly to accept the adequacy of the naked assertion of negligence — the ultimate legal conclusion — but to find the complaint adequate, in contrast to that in *Twombly*, entirely because of its substantially greater degree of notice (that is, the specificity greatly narrows the focus, even while stating absolutely nothing about the alleged negligence itself), a function of pleading that the Court does not otherwise expressly advance in presenting its plausibility standard. (In addition to the *Twombly* majority's failure to feature notice in it's central analysis in the text of its opinion, the term is not even mentioned in the relevant sense in *Iqbal*.) *See also supra* note 177 (discussing the entanglement of the notice function of pleading and the plausibility requirement).

The distinction between facts and legal conclusions seems central to *Twombly* and *Iqbal* for two reasons. First and most obvious is the repeated emphasis on the point in the central portions of the Court's opinions. Second, more formally, if one were to reject this distinction and also to continue to adhere to the mantra that allegations must be accepted as true, then it would be hard to see how there could nevertheless be a nontrivial floor that must be surpassed to survive a motion to dismiss. A few commentators, however, do not believe that any distinction is possible. *See* Allen & Guy, *supra* note 183, at 33-34 ("In *Twombly* the Court thought that the allegation of a conspiracy was a 'legal conclusion' and not, apparently, a fact but that is simply impossible to understand. Asserting a conspiracy plainly asserts a state of affairs in the world independent of the law, and this remains true even if the existence of a conspiracy is an element of the substantive law. In any event, the complaint went further and asserted agreements not to compete, among other 'facts.' In *Iqbal*, a remarkable litany of what any rational person would deem 'facts' were found to be unacceptable conclusions, such as Ashcroft being the 'principal architect' of an invidiously discriminatory policy, that Mueller 'was instrumental in adopting and executing it,' and that this was done on account of the plaintiff's religion, race, and/or national origin. These are factual assertions distinguishable in no interesting qualitative way from an allegation like 'the defendant ran the red light' or that 'the defendant drove negligently.'"). Perhaps it would be helpful to distinguish not facts from legal conclusions but rather intermediate facts from ultimate facts, although it may remain unclear how far removed an intermediate fact must be from an ultimate fact to be credited under the Court's rubric.

need not be described, but, as just explained, it presumably must exist in order for the allegations to be taken seriously.

Nor does the convention of pleading based on "information and belief" avoid the need for evidence (whether it must be articulated or not); as mentioned, information is evidence, and a belief with no evidentiary basis is insufficient for the current purpose.[198] Furthermore, this need cannot be avoided by taking advantage of Rule 11(b)(3)'s option of asserting that one's claims "will likely have evidentiary support after a reasonable opportunity for further investigation or discovery" because asserting an affirmative likelihood of finding evidence presupposes that one already has some evidence.[199] (Recall section A's paradoxical example in which there need not be any probability that oil is present but there must be a decent probability that a pilot hole will in fact discover oil.[200]) Were there no such requirement, it would be difficult to understand the opinions and outcomes in *Twombly* and *Iqbal*, both of which do not allow plaintiffs to proceed to discovery because their existing facts are insufficient.

Put another way, what may appear to be courts and commentators' fixation with precisely what is stated in a complaint is in an important respect an indirect way of addressing what evidence needs to be present. A difficulty is that the emphasis on complaints' language obscures and thus averts attention[201] from the question of what a plaintiff actually needs to know

---

[198] *Black's Law Dictionary* indicates that information and belief is that "based on secondhand information that the declarant believes to be true." BLACK'S LAW DICTIONARY 783 (7th ed. 1999). By contrast, some authorities suggest plaintiffs can somehow pull themselves up by their own bootstraps. Wright and Miller's treatise, for example, states that "[p]leading on information and belief is a desirable and essential expedient when matters that are necessary to complete the statement of a claim are not within the knowledge of the plaintiff." 5 CHARLES ALAN WRIGHT & ARTHUR R. MILLER, FEDERAL PRACTICE AND PROCEDURE § 1224 (3d ed. 2004). In this usage, the "information and belief" seems to be a legal fiction, with the requirement of any factual basis being excused. Although it is difficult to see how such an approach could continue in light of the facts, reasoning, and outcomes of *Twombly* and *Iqbal*, it seems alive in some lower courts. *See, e.g.*, Arista Records, LLC v. Doe 3, 604 F.3d 110, 120 (2d Cir. 2010); Boykin v. KeyCorp, 521 F.3d 202, 215 (2d Cir. 2008).

[199] *See, e.g.*, Elan Microelectronics Corp. v. Apple, Inc., No. C 09-01531 RS, 2009 WL 2972374, at *4 (N.D. Cal. Sept. 14, 2009). It appears that the intention behind Rule 11(b)(3) is not to permit complaints to proceed when they lack an evidentiary basis in the epistemic sense but rather that such may be present but not in a form that may be appropriate for presentation at trial. *See* FED. R. CIV. P. 11, advisory committee's note ("The certification with respect to allegations and other factual contentions is revised in recognition that sometimes a litigant may have good reason to believe that a fact is true or false but may need discovery, formal or informal, from opposing parties or third persons to gather and confirm the evidentiary basis for the allegation."). This provision, which requires specific identification, still seems strange because all discovery is intended to lead to a party's ultimately obtaining evidence, not already in its possession, that may be so used.

[200] Or, to take a more mundane example, I cannot sensibly assert that my misplaced cell phone is likely to be found on the front seat of my car without some basis for believing that my phone was left there — whether I have a particular recollection, know that this is where I often find it when it has been misplaced, or have some other ground.

[201] This point should also remind the reader that the analysis in this Part does not purport to present a definitive legal interpretation of existing rules or to offer the best account of what was actually on the Justices' minds in writing the opinions in *Twombly* and *Iqbal*. The most plausible account may be that many facets of the question examined here were at best inchoate, in which case the most that could be said on the matter would relate to predictions of what the Court might say if they responded sharply to certain queries.

(assuming adherence to Rule 11[202]) in order to craft a legitimate complaint. The remainder of this section elaborates this knowledge base, setting aside the question of the extent to which the requisite foundation must be stated explicitly in a complaint rather than presented in a more summary form. Before proceeding, it should be noted that this sort of inquiry is also relevant at other stages — for example, at summary judgment and at trial — and in other settings with different tests, as long as the factual basis for a legal claim is at issue.

In examining the factual underpinning for legal decisionmaking, most have in mind whatever particular pieces of evidence are available, but it is important to consider as well, really in combination, background knowledge.[203] For example, the weight to be given to a witness's statement that the light was red will depend — along with other particulars, such as the time of day, whether the witness wears glasses, his viewing angle and distance, his relationship to the parties — on our general knowledge of the reliability of vision and of the honesty of individuals in pertinent settings. If we know that an individual suffered a bad outcome from a medical procedure, our inference about negligence would importantly be informed by how often the procedure, when conducted properly, generated such a result versus when it was performed negligently. Indeed, the term res ipsa loquitur, often translated as "the thing speaks for itself," is a misnomer: the thing (action) only speaks at all, much less with clarity and force, because of its interaction with the observers' substantial background knowledge.[204]

The importance of general information is difficult to overstate, and, as these illustrations suggest, it is manifest in numerous ways. First, it plays a central role in interpreting particular evidence: assessing weight and what inferences may be drawn. Second, it often constitutes a substantial basis for factual conclusions even when particulars are minimal (which can be seen as

---

[202]It is also important to contemplate the effects of decision rules at various stages for enforcers (private plaintiffs or government actors) who might contemplate bending or outright violating the rules. Because such behavior may not be readily detected or sufficiently punished to discourage it in all guises, and because largely baseless cases may significantly contribute to the problems that screening procedures, such as the motion to dismiss, are designed to mitigate, optimal decision rules will reflect this possibility. In part, pressing for more detailed factual allegations may have this motivation: forcing the initiating party to commit to details may make subsequent sanctions for misbehavior easier to apply. (The same can also be said of requiring defendants to offer more precise answers to complaints, a subject not addressed in this Article.) In addition, in contemplating different bases for inferences discussed in the text to follow, it is worth augmenting the analysis by consideration of what sorts of decisional bases are easier to fabricate. For example, it may be easier to make up fragments than to convince a decisionmaker that background knowledge of how the world functions is other than it really is. Or perhaps not: maybe the former, although simple to do, might more likely result in subsequent sanctions (depending, for example, on whether alleged rumors are a sufficient basis for particular allegations), whereas the latter might be generated, at some level of prima facie credibility, when one is able to find what superficially appears to be an expert who, for adequate compensation, can be induced to say just about anything.

[203]*See, e.g.*, Allen & Guy, *supra* note 183, at 28 ("'Evidence' is not packets of information. Those packets of information are completely meaningless unless analyzed by a human bringing to bear a vast conceptual apparatus including such things as the meaning of language, rules of logic, expectations and beliefs about the real world, and so on. 'Evidence' is thus not things produced at discovery or trial but the consequence of an interaction between those things and all the cognitive capacities of a person."); *id*. at 34 (referring to "the myth . . . that there is a useful analytical difference between 'evidence' and the background knowledge and experience brought to bear in appraising it").

[204]*Cf.* RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 17 cmt. c (2005) [hereinafter RESTATEMENT THIRD] ("Insofar as the basis for a res ipsa loquitur finding comes from the testimony of experts, in a sense the case is not one in which 'the accident speaks for itself.' Even so, expert testimony obviously can provide an appropriate basis by which the jury can make the general assessments called for by the res ipsa loquitur doctrine.").

a strong version of the first point). One does need some particulars: the plaintiff died during her minor medical procedure. Beyond that, the remainder of the inputs may involve background knowledge.

Consider the implications of these observations for decisions on motions to dismiss. Suppose that a complaint modestly exceeds a bare-bones conclusion ("defendant wrongfully injured plaintiff") — specifically, it identifies the core act but in highly general terms, such as in the example wherein all the plaintiff alleges is the occurrence of a certain injury during a particular type of medical procedure. Is negligence sufficiently pled? From *Twombly* and *Iqbal*, it may seem not.

Next, assume that there is also (recall the example in section A) an expert report to the effect that this outcome is rare when the procedure is done properly but the injury is typical when the procedure is botched. It may be that the likelihood of negligence is extremely high (say, 97%), easily enough to win at trial, even though no additional factual matter is offered. Now contrast a similar case, but where the probability of negligence is 0.0001% (as earlier, one in a million). This plaintiff, however, presents a number of additional, quite specific bits of evidence, all pointing toward negligence. That might seem sufficient to survive a motion to dismiss. Suppose, however, that the (not required) expert statement is compulsively complete and makes clear that, with this added evidence, the probability actually raises to only 0.0002% (two in a million). It seems hard to argue that the latter complaint — which has multiple particular factual allegations that all point toward negligence — is stronger than the first — which has none of them beyond the bare statement that the plaintiff suffered the injury from the stated medical procedure, but where the injury-procedure combination, by itself (although combined with background knowledge), indicates (really is evidence of) an overwhelming (rather than minuscule) likelihood of negligence.[205]

It would be nonsensical to use the existence or nonexistence of some fragment to determine when factual allegations were "enough" to be deemed plausible.[206] Many complaints' mere presentation of the setting incorporates, even if implicitly, many particular facts that may warrant denying a motion to dismiss. For example, knowing that a twenty-five-year-old individual bled to death while having a tooth removed, even if described as a mere setting of the

---

[205]The tendency to focus on particular fragments and underplay background knowledge is loosely related to the familiar base-rate fallacy (related to the representativeness heuristic) in cognitive psychology, where particulars, such as anecdotes, are given undue weight relative to statistical knowledge of general likelihoods (such as in believing that air travel is particularly dangerous based on reports of crashes, the vastly larger number of non-crashes being absent from the news, with overall statistics indicating that flying is actually much safer than alternative modes of transport). *See, e.g.*, Maya Bar-Hillel, *The Base-rate Fallacy in Probability Judgments*, 44 ACTA PSYCHOLOGICA 211 (1980); Daniel Kahneman & Amos Tversky, *Evidential Impact of Base Rates*, *in* JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES 153 (Daniel Kahneman, Paul Slovic & Amos Tversky eds., 1982); Daniel Kahneman & Amos Tversky, *On the Psychology of Prediction*, 80 PYSCHOL. REV. 237 (1973).

[206]*See* Bell Atlantic Corp. v. Twombly, 550 U.S. 544, 555-57 (2007) (using the term "enough" five times in discussing the requisite factual allegations). Also, even if required, a mere fragment is typically easy to allege. Strictly speaking, even the mundane assertion that the defendant was seen in the country during the year in question elevates the likelihood that the defendant committed the act, even if by a trivial amount (because it does rule out the possibly that the defendant was never present, which the example is supposing to be a predicate for the overall claim to be true).

stage, can be sufficient to proceed. (The aforementioned doctrine of res ipsa loquitur comes to mind.[207]) That is, virtually any actual complaint (including those in *Twombly* and *Iqbal*) contains numerous factual allegations that are relevant in assessing likelihoods, even if we are not used to conceiving of them as such. From this perspective, we can say that there typically exist both a number of particular facts and much general knowledge from which inferences can be made, which seems to shift attention away from the mere presence of one or a few fact fragments to the question of how strong of an inference can be drawn from them and whether that suffices. That is, once the analysis of facts moves beyond the existence of a relevant corpuscle of evidence, the focus would have to shift to an assessment of likelihoods where, depending on the circumstances, background understandings could play a central, even dominant role. In *Twombly* and *Iqbal*'s demands for sufficient facts, one might suppose — or perhaps hope — that sufficiency refers to their weight[208] rather than some minimum number of scraps (three?) that bear positively, even if minutely, on liability.

The foregoing suggests that general knowledge bearing on even the most stripped-down description of the situation often should be given more weight, even overwhelming weight, relative to the sorts of particular factual elaborations that often receive the most attention, all depending on the circumstances of the case. In both *Twombly* and *Iqbal*, plaintiffs' complaints were not of the extreme bare-bones variety; each described a concrete scenario in the world. There was no mystery about the general notion of what was alleged to have occurred. Rather, the Court in both cases believed that those descriptions alone did not provide a sufficient basis to infer the asserted violations. Much of the discussion in both majority opinions suggested — drawing on the Court's general understanding of the contexts — that other explanations for defendants' behavior were readily available, not ruled out by the allegations, and consistent with

---

[207]For example, according to the version of res ipsa loquitur advanced in RESTATEMENT THIRD, *supra* note 204, § 17, "The factfinder may infer that the defendant has been negligent when the accident causing the plaintiff's harm is a type of accident that ordinarily happens as a result of the negligence of a class of actors of which the defendant is the relevant member." Under this doctrine, a description of the setting combined with background knowledge, but without any further particulars, is deemed to be sufficient for liability in final adjudication. The argument in the text is that, if such information is sufficient to prevail at trial, it is a sufficient type of information to deny a motion to dismiss, as long as the strength of the inference it generates meets the pertinent standard in the particular context.

One could go further by arguing that this doctrine and others that allocate proof burdens between plaintiffs and defendants may have their primary impact not at trial, which is the standard context in which they are discussed, *see, e.g.*, 2 MCCORMICK ON EVIDENCE §§ 336–44 (Kenneth S. Broun ed., 6th ed. 2006), but in deciding motions to dismiss, precisely because discovery comes in between. Discovery provides access to defendants' information, enabling a more complete understanding of what happened, including via negative inferences, regardless of who has the production burden at trial. After all, if the burden remains with the plaintiff, the plaintiff can introduce the defendant's information, including non-information (such as the inability of the relevant employees to offer any other convincing explanation of the acts and events in question), from which a factfinder may make negative inferences. By contrast, before discovery, the plaintiff does not have such access, but placing the production burden on the defendant (supposing the triggering allegations are present) may mean directly that the plaintiff wins the motion or that, in coming to a decision, the tribunal should be drawing negative inferences (although the standard procedures for motions to dismiss, unlike those for summary judgment, do not contemplate affirmative representations by the defendant).

[208]This point relates to the analysis in section A indicating that it is difficult to interpret *Twombly* and *Iqbal*'s statements that the test is not a probability standard as rejecting the relevance of probabilities rather than the existence of some target probability that is independent of the context.

legal behavior.[209] How likely they were and just what was the general knowledge upon which the Court drew was less clear.[210]

This latter point presents an issue examined further in section C regarding how judges are supposed to make decisions that presume knowledge outside their ken. As explained in section A, *Iqbal* instructs judges to undertake context-specific inquiries that draw on their "experience and common sense." These founts of wisdom, however, do not provide a basis for understanding the relationships between medical procedures and outcomes (the preceding hypothetical example), strategic firm interaction in a recently deregulated and rapidly evolving telecommunications industry (*Twombly*), and perhaps as well the likelihoods of various mixes of motives in generating a massive multidimensional response to a unique historic event that generated strong emotional and sometimes irrational political forces (*Iqbal*, post 9/11).

The foregoing discussion of the nature of facts should also be related to the analytical framework in Part II. Facts and, relatedly, evidence are what constitute what were there referred to as scenarios, different facts and evidence corresponding to different scenarios. For example, in discussing the extent to which a decision to continue rather than terminate would contribute to deterrence, the contribution to the deterrence punch was explained to depend on the likelihood that a harmful act would enter the legal system and present itself as being in the scenario in question (and also on how much continuation would increase the expected costs). This likelihood, obviously, is a function of the facts and evidence. For example, if harmful acts would rarely generate such a factual configuration, the contribution to deterrence from continuation would be negligible, whereas if benign acts often did, then the contribution to chilling may be large.[211]

In making such statements, what matters are the pertinent likelihoods, which depend on all of the available facts and evidence analyzed as a whole. Some facts may be closer to ultimate ones and others intermediate; there may be many particulars or few, and background knowledge

---

[209]Even the *Iqbal* dissent acknowledged that one could draw on general knowledge to dismiss a complaint in sufficiently extreme circumstances. "The sole exception to this rule lies with allegations that are sufficiently fantastic to defy reality as we know it: claims about little green men, or the plaintiff's recent trip to Pluto, or experiences in time travel. That is not what we have here." Ashcroft v. Iqbal, 129 S. Ct. 1937, 1959 (2009) (Souter, J., dissenting). If this is indeed the "sole exception," it would seem that the two Justices (Souter and Breyer) subscribing to the proposition but voting with the majority in *Twombly* must have viewed allegations of businesses conspiring to maintain supracompetitive profits as hallucinations (in sharp contrast to how the *Twombly* majority, also authored by Justice Souter, described them).

[210]Regarding the likelihood, see section A and, for *Iqbal*, the discussion in note 183 about the Court's suggestion that the plaintiff's explanation had to be the most likely one. Recall that the *Twombly* Court purported to draw on "common economic experience," 550 U.S. at 565, but the question it faced was not of the entirely general sort, such as how often a firm's failing to enter another's territory is due to conspiracy, but, as elaborated in section C, a much more focused one — in the setting of recent telecommunications deregulation, where some had predicted such entry and where such entry, undertaken independently, would have been profitable (as the plaintiffs alleged), how likely was the abstention to reflect conspiracy? — on which the Court did not refer to any particular knowledge base.

[211]As explained in Part II, the scenario may have further relevance, for example, for the magnitude of continuation costs. For example, if a medical malpractice claim describes an injury of a type that could have been caused by doctor error, discovery would likely be fairly circumscribed, whereas if the claim is that the injury derives from an infection made more probable by poor training and faulty air circulation systems at a hospital, discovery (and the scope of expert reports) would probably be more wide-ranging.

that gives meaning to the specific facts that constitute the scenario may be extremely important or less so.[212] What has been referred to here as a description of the setting (for example, that some specific injury was suffered during some type of medical procedure) could readily be the complete depiction of a scenario in Part II, or a substantial portion thereof. The presence or absence of some fragments of evidence may well alter the scenario, but if the pertinent likelihoods, which affect the increments to deterrence and chilling that result from continuation, change only slightly, the optimal termination/continuation decision would not be influenced except in the closest of cases. In contrast, differences in the setting and thus in the implications of background knowledge could have huge effects on the social costs and benefits of continuation versus termination and thus on the optimal decision In all, the irrelevance per se of the various distinctions elucidated in the present section to any sensible test — however helpful the taxonomy may be in guiding our understanding of how facts are rationally processed — is implicit in and fully consistent with Part II's analysis.

Finally, it is illuminating to relate this discussion of the nature of facts to the familiar concern about information being solely in possession of defendants, aspects of which were addressed in subsection II.A.3. To begin, note that the extreme, pure case is not what is ordinarily contemplated, as can be seen by imagining what it would mean for literally all pertinent information to be confined exclusively to the defendant's knowledge. In such a case, a plaintiff could not even know whether she was injured, what sort of injury she suffered, whether she had any relationship to the defendant, and so forth — essentially, it would be a random lawsuit. Instead, what is meant is that, while the plaintiff does know, for example, that she suffered a particular injury while under anesthesia on the operating table for a given procedure, any knowledge of what precisely occurred during that time period is in the defendant's hands. As just explained, however, even this basic depiction of the setting contains substantial information: evidence and certain supporting facts that, together with background knowledge, have particular implications. Applying the decision framework of Part II, there would be scenarios in which this sort of information base was quite favorable to continuation and others in which termination was strongly indicated.

Another possible interpretation of the concept of information being solely in the defendant's hands is that it refers to the particulars of the case rather than the basic depiction of the setting and background knowledge. But we have seen that these distinctions, to a substantial degree, are neither meaningful (what is the difference between a feature of the setting and a particular about the case? can one ever give meaning to any particular fact without reference to background knowledge?) nor important (why do we care about the number of fragments rather than the overall likelihoods derived from all pertinent sources, appropriately combined?).

To further press these distinctions, consider the following facts that might appear in a

---

[212]It is worth noting that, even though the present discussion emphasizes the importance of background knowledge, the logic does not involve Bayesian analysis, specifically, the use of Bayesian priors to generate Bayesian posterior probabilities, which here would refer to the likelihood that the act in a given scenario is harmful rather than benign. Instead, it uses background knowledge pertaining to the likelihoods that harmful acts and benign acts would enter the legal system and present themselves in the scenario at hand, as stated in the preceding paragraph and throughout Part II. *See supra* note 85.

complaint alleging racial discrimination in employment: The plaintiff (*P*) applied for a job. The job was with the defendant (*D*). *D* actually had such a job available, as evidenced by a job posting. *P* is African-American. *P* has the qualifications listed in the job description, as evidenced by *P*'s resume. The interviewer was white. *P* was turned down. The job remained vacant for a month. The posting stated an immediate need. *P* had relevant experience in excess of what the listing had indicated. The job was ultimately filled by a white applicant. The individual hired did not have the stated minimum level of experience. *D* is a huge firm that employs many individuals of the type involved in the posting in question. Even though twenty percent of the labor pool in that field consists of African-Americans, they make up only three percent of that part of *D*'s labor force.

One can ponder if this entire set of facts, or various subsets thereof, would be enough to survive a motion to dismiss. Note that every one of them elevates the likelihood of discrimination relative to the likelihood if the fact is absent. (Even the seemingly bland facts, such as *P* applying for the job, there being an opening, and the like, are actually crucial.) There are, in this listing, none that directly reveal the actual thought process of the interviewer or of whomever made the rejection decision (an identity that may be unknown to *P*). In addition, the last handful of facts in some instances will be absent precisely because key pieces of information may be in *D*'s possession. On the other hand, it is also true that much that is unknown and thought to be exclusively in *D*'s possession may be only probabilistically so (even if the probability is high). For example, sometimes there will be disgruntled employees, former employees, or individuals who post incriminating information on the Internet. Also, if discrimination is rampant at *D*, there may be prior lawsuits in the public record, agency investigations, or news accounts that provide further information or leads that could produce additional evidence.

The point of this example is that the extent to which information is available to plaintiffs is a matter of degree, varying by the type of case and myriad particulars of a given situation. Even holding much constant, we can imagine different scenarios in which more or less information is accessible. Nor, as repeatedly emphasized, should it matter how much of the facts presented are part of the setting, further particulars, or background knowledge — by contrast to whether all the facts, regardless of their type, taken as a whole, imply that the scenario is one in which continuation makes sense.

Consideration of what may be in defendants' hands is relevant in other ways. Part of the decisionmaking calculus involves how much may be learned from continuation and how much this prospect may remedy deficiencies in deterrence, augment chilling, and impose costs in identifying additional evidence even when it does exist. In other words, this Article's framework does reflect, in various and sometimes subtle ways, what information one predicts might be revealed when a case is continued. In some instances (but not all), the overly simple suggestion that information is solely in the defendant's possession may be a shorthand for the expected information value of continuation being large, but we can now see that the idea needs to be unpacked, and the revealed elements combined with others, in order to make sensible judgments

about the value of continuation versus termination.[213]

C. Informational Challenges

Whether employing the explicit analytical framework of Part II or the possibly overlapping context-specific inquiry that seems to be demanded by *Twombly* and *Iqbal*, rather than a minimalist, formalistic legal test, the informational challenges confronting the decisionmaker are formidable. Likewise for rulings at summary judgment (see section E) or really at any stage in any sort of legal proceeding, if indeed termination/continuation decisions and those in final adjudication aim to further the legal system's objectives. A major lesson of this Article is that the task is substantially more demanding than seems to be appreciated.[214] This section elaborates this difficulty and then addresses possible responses, including how litigants and courts might attempt to adapt under the existing regime.

The informational requirements of a purposive decision criterion in multistage adjudication are high. This point is clear in reflecting on the analysis in subsection II.A.2, which makes explicit the pertinent factors. It is hard enough to know the likelihood that truly harmful and actually benign acts will present themselves in a given scenario and to estimate the continuation costs. These inputs, which come most readily to mind, are indeed central but, as explained, are not nearly sufficient. To quantify deterrence benefits and chilling costs, it is also necessary to estimate the level of expected sanctions for the two types of acts when cases in the scenario are expected to be terminated and also the benefit per deterred act and the cost per chilled act. These factors, in turn, depend on expected outcomes in other scenarios and on additional considerations, as explored in subsection II.A.3. For many areas of law and particular situations in each field, such information is neither already known nor readily obtainable.

To see this problem as a defect of the analytical framework, however, would be to blame the messenger. Supposing that the legal system is indeed meant to advance the sorts of purposes

---

[213]For example, section II.A.3 explains that it is quite important whether the plaintiff's inability to access additional information is a particular feature of the scenario at hand or is a general feature of the class of cases, the latter often being more favorable to continuation.

[214]Among the more direct prior recognitions of this point is Bone, *supra* note 12, at 7 ("The problem of determining the optimal role for strict pleading in a screening system is too empirically contingent and the cases too heterogeneous to be confident that a one-size-fits-all rule, like Rule 8(a)(2), is optimal. Moreover, the problem is too complex to resolve through case-specific decision making like that in *Twombly* and *Iqbal*."); *cf.* Bone, *supra* note 7, at 851 ("Screening weak lawsuits raises much more complex and controversial policy questions than screening meritless suits, and the Supreme Court is not well equipped institutionally to address those policy questions. They are better left to the committees involved in the formal rulemaking process or to Congress."). Burbank, *supra* note 123, advances a similar view, with a particular emphasis on courts' competency, or lack thereof, to make the types of decisions contemplated by *Twombly* (and, in anticipation at the time of his writing, *Iqbal*). *See also id*. at 559–60 ("The *Twombly* Court, by contrast [to *Tellabs*, applying the heightened pleading standards of the PSLRA], was not well positioned institutionally to evaluate even the procedural costs and benefits of tightening the pleading screws on plaintiffs, even in the isolated substantive-law context involved in the case. The Court acting as such under Article III was even less well positioned to estimate the procedural costs and benefits of a general rule of plausible pleading (if that is what *Twombly* gives us), let alone the nonprocedural costs and benefits of such a rule, substance specific or general."); *id*. at 561 ("In any event, from this perspective, it is again apparent that the policy questions are not the sort that should be answered by nine judges in the exercise of Article III judicial power, with little information, less experience, and no power to implement nonlitigation alternatives.").

articulated in Rule 1, only wishful thinking or willful obliviousness can evade the challenge. Generally, it is better to appreciate the consequences of important decisions and do the best one can rather than to pretend that the choice is simpler, ignoring some effects or imagining that they can be divined without offering any means by which that might be done. Part II supposes that we indeed care whether decisions in multistage legal proceedings advance social welfare and accordingly attempts to determine the implications for how such decisions should be made. If we were indifferent to whether harmful acts were deterred, benign acts chilled, and adjudication costs incurred, these decisions might be easy; otherwise, the challenge must be confronted.[215]

As mentioned in section A, *Iqbal* instructs judges to make decisions on motions to dismiss by drawing on their "judicial experience and common sense."[216] Unfortunately, these two sources of information and inspiration do not seem up to the task. Almost none of the various components and subcomponents that enter into the optimal decision rule seem within the realm of judicial experience. Whether considering time on the bench or prior activities, most judges would never have been in the position of most potential actors contemplating either harmful or benign acts. The various empirical questions pertaining to preexisting levels of deterrence and chilling and the net effect of deterring a marginal harmful or benign act are in a different arena. Nor is any of this information in the domain of common sense, except perhaps for simple forms of behavior that may be the subject of routine torts, such as automobile accidents. With regard to understanding of the universe, common sense refers to the generalized knowledge of an ordinary individual, which surely does not encompass the theory and empirics of complex economic behavior in particular industry settings or the internal operation of large bureaucracies, government or private, or the sorts of evidence likely to be generated by such behavior, harmful or benign, and with what probabilities.

Twombly itself illustrates judges' predicament.[217] The Court suggested that its situation was strong because, "[i]n identifying facts that are suggestive enough to render a § 1 conspiracy plausible, we have the benefit of the prior rulings and considered views of leading commentators, already quoted, that lawful parallel conduct fails to bespeak unlawful agreement."[218] But confidence that lawful conduct does not demonstrate unlawful conduct hardly takes one very far. What a court actually would need to have known — or formed a judgment about — to decide the matter was whether the parallel conduct in the circumstances of the case was suspicious, keeping in mind that it is blackletter law that an illegal conspiracy may be proved by circumstantial evidence.[219] Those circumstances would include the nature of the

---

[215]*Cf.* BONE, *supra* note 32, at 138 ("In any event, there is no way to avoid estimating these variables, no matter how one approaches the issues. Any policy analysis must compare the frequency and cost of error.").

[216]Ashcroft v. Iqbal, 129 S. Ct. 1937, 1950 (2009).

[217]Also recall the medical malpractice example from sections A and B, and consider a court's ability, when deciding on a motion to dismiss, to make the requisite empirical findings based on its own knowledge.

[218]Bell Atlantic Corp. v. Twombly, 550 U.S. 544, 556 (2007).

[219]*See, e.g.*, 1 ABA SECTION OF ANTITRUST LAW, ANTITRUST LAW DEVELOPMENTS 5–6 & n.29 (6th ed. 2007) ("Conspiracies can be proven either by direct or circumstantial evidence. . . . [C]ourts traditionally recognized that '[o]nly rarely will there be direct evidence of an express agreement' in conspiracy cases . . . . Circumstantial evidence as to this element of the offense is . . . not only admissible, but often dispositive.") (quoting Local Union No. 189, Amalgamated Meat Cutters & Butcher Workmen of N. Am., AFL-CIO v. Jewel Tea Co., 381 U.S. 676, 720 (1965) (Goldberg, J., concurring), and for the latter proposition, citing, inter alia, Monsanto Co. v. Spray-Rite Serv. Corp., 465 U.S. 752,

telecommunications industry at that point in time, viewed in light of the recent statute and regulations that dramatically changed the landscape and the accompanying expectations about their implications for firms' behavior, including in particular that they would lead to the incumbents entering each others' territories so as to create competition.[220] Neither prior rulings nor general commentary nor basic economic principles[221] indicated how plausible it was that an

---

765–66 (1984)); 6 PHILLIP E. AREEDA & HERBERT HOVENKAMP, ANTITRUST LAW 4 (3d ed. 2010); *see also* 2 WAYNE R. LAFAVE, SUBSTANTIVE CRIMINAL LAW 267 (2d ed. 2003) ("[I]t is thus well established that the prosecution may 'rely on inferences drawn from the course of conduct of the alleged conspirators.'") (quoting Interstate Circuit v. United States, 306 U.S. 208, 221 (1939)).

[220]For example, the *Twombly* complaint alleged (¶ 2) that "[t]he purpose, intent and requirements of the [Telecommunications Act of 1996] are to create competition without delay in the local telephone services markets," and that the FCC's actions implementing the Act were "intended to encourage the development of competition" (¶ 34 (quoting 61 Fed. Reg. 45476 (Aug. 29, 1996))). With regard to the latter, the complaint refers to the partially successful challenge to the FCC's regulations in *Iowa Utilities Board v. FCC*, 120 F.3d 753 (8th Cir. 1997). That court had stated that

> Congress clearly included measures in the Act, such as the interconnection, unbundled access, and resale provisions, in order to expedite the introduction of pervasive competition into the local telecommunications industry. *See* H.R. Rep. No. 104–204(*I*), 1995 WL 442504 at *202–03, 494 (1995) (explaining importance of resale provision for the early development of competition and indicating that the local competition provisions "create the transition to a more competitive marketplace"). Congress recognized that the amount of time and capital investment involved in the construction of a complete local stand-beside telecommunications network are substantial barriers to entry, and thus required incumbent LECs to allow competing carriers to use their networks in order to hasten the influence of competitive forces in the local telephone business. The Commission's unbundling rules facilitate the competing carriers' access to these networks and thus promote the Act's additional purpose — the expeditious introduction of competition into local phone markets.

Id. at 816. In the conclusion to its detailed review of challenged FCC regulations under the 1996 Telecommunications Act, the court further observed: "As an aside, and while we do not pretend to possess the Rosetta stone that reveals the true meaning of every portion of this Act, we hope that our review of the FCC's First Report and Order in light of the Act's provisions offers some guidance to the participants in the telecommunications industry as they continue its evolution into the competitive marketplace Congress intended." *Id*. at 820. In essence, the *Twombly* Court professed that federal judges (presumably including those issuing this Eighth Circuit opinion) should be able to determine, based on their general knowledge of the world, that Congress's and the FCC's predictions of what would happen in the absence of illegal behavior were sufficiently dubious that their failure to materialize did not render the explanation of conspiracy at least plausible. (Such may well be the case, among other reasons because statements in congressional and agency reports are sometimes public relations platitudes and can serve as a cover for giveaways to special interests. The present point is simply that the standard seems to suppose that federal judges are in a position to determine these likelihoods, largely on their own.)

[221]The central point of the Brief of Amici Curiae Economists in Support of Petitioners, *Twombly*, 550 U.S. 544 (No. 05-1126), was that parallel behavior is ubiquitous when behavior is competitive; hence, cases alleging mere parallel behavior should not be permitted to proceed. In a similar vein, some commentators suggest that *Twombly* was easy because firms often do not enter completely unrelated lines of business. *See* AREEDA & HOVENKAMP, *supra* note 71, at 120-21 (explaining that "Schwinn does not make pickup trucks and Chrysler does not make bicycles"); Hovenkamp, *supra* note 16, at 65 (complimenting Justice Roberts's question at oral argument about a grocery store not competing with a nearby pet store, and suggesting that "those allegations could be made about almost any firm"). *But see* Caitlin Moldvay, *Lucky Dog: Pet Owners Will Invest in Premium Products as Disposable Income Rises*, IBISWorld Industry Report 45391: Pet Stores in the US 4, www.ibisworld.com/gosample.aspx?cid=1&rtid=101 (last visited Apr. 20, 2012) ("Pet food, toys and accessories supplied by this industry compete with comparable products offered by supermarkets and grocery stores."); *id*. at 7 ("As such, grocery stores and mass merchandisers have become leading retailers of pet supplies over the five years to 2012."); *id*. at 8 ("However, the Pet Stores industry will continue to face strong competition from grocery stores and mass merchandisers, which to some extent will place a cap on the industry's growth."); First Research, *Pet & Pet Supplies Stores Industry Profile*, http://www.firstresearch.com/Industry-Research/Pet-and-Pet-Supplies-Stores.html (last visited Apr. 20, 2012)

illegal agreement explained the unanticipated noncompetitive outcome actually experienced. It is as if the Court was taking (and expecting lower courts in similar situations to take) judicial notice of complex matters that one might ordinarily have expected would be the subject of expert testimony,[222] none of which was in the record when originally deciding or reviewing the decision on the motion to dismiss.[223]

(mentioning grocery stores first in a list of pet stores' competitors). But the actual behavior in *Twombly* was radically different: it concerned nonentry in firms' existing line of business (an aspect Hovenkamp acknowledges), sometimes in areas in which they served nearly all customers except in isolated, surrounded regions (¶ 40), where the firms that abstained from entering were alleged to have "substantive competitive advantages" (¶ 41) that generated "an especially attractive business opportunity" (¶ 40) (quoted by the *Twombly* majority, 550 U.S. at 567), which, moreover, was one that it was asserted the legislature enacting the 1996 Telecommunications Act anticipated would be taken advantage of (¶ 2). If those allegations are taken to be true, one then must further inquire whether an illegal agreement thereby constitutes a plausible explanation for that behavior in the industry context in question. It is this specific, contextual question that a federal judge is presumed to be able to answer.

[222]Judicial notice of adjudicative facts is governed by Rule 201 of the Federal Rules of Evidence. If a judge were to take judicial notice, the court would be limited in what would be permitted: "The court may judicially notice a fact that is not subject to reasonable dispute because it: (1) is generally known within the trial court's territorial jurisdiction; or (2) can be accurately and readily determined from sources whose accuracy cannot reasonably be questioned." FED. R. EVID. 201(b). The sorts of facts under discussion in *Twombly* (or in *Iqbal*, or in the other examples discussed in this Part) do not readily fall within the Rule's ambit. Much of the information pertinent to applying the optimal decision rule involves legislative fact, which is not governed by any rule. *Id*., advisory committee's note. The distinction between adjudicative and legislative facts is unclear, but since the source for neither is apparent in the present setting, the difference seems secondary. More broadly, it seems widely accepted that factfinders will rely substantially on knowledge beyond that formally presented, a point that is made apparent by section B's analysis of the nature of facts, specifically, with regard to background knowledge. *See also id.* ("As Professor Davis points out, A System of Judicial Notice Based on Fairness and Convenience, in Perspectives of Law 69, 73 (1964), every case involves the use of hundreds or thousands of non-evidence facts."); JAMES BRADLEY THAYER, A PRELIMINARY TREATISE ON EVIDENCE AT THE COMMON LAW 279–280 (1898) ("In conducting a process of judicial reasoning, as of other reasoning, not a step can be taken without assuming something which has not been proved; and the capacity to do this with competent judgement and efficiency, is imputed to judges and juries as part of their necessary mental outfit."). *See generally supra* section III.B (discussing the difference between whether factual differences in legal settings should be seen as associated with substantive or procedural law).

[223]The *Twombly* Court writes as if it felt comfortable resolving these issues on its own.

> But it was not suggestive of conspiracy, not if history teaches anything. In a traditionally unregulated industry with low barriers to entry, sparse competition among large firms dominating separate geographical segments of the market could very well signify illegal agreement, but here we have an obvious alternative explanation. In the decade preceding the 1996 Act and well before that, monopoly was the norm in telecommunications, not the exception. See *Verizon Communications Inc. v. FCC*, 535 U. S. 467, 477–478 (2002) (describing telephone service providers as traditional public monopolies). The ILECs were born in that world, doubtless liked the world the way it was, and surely knew the adage about him who lives by the sword.

550 U.S. at 567–68. As mentioned in notes 220 and 221, some of these assertions are hardly obvious and were contested; for example, the plaintiffs specifically alleged that the purpose of the 1996 act was to make regulatory changes that would lead to competition. Even as a generalization, one might further question the empirical basis for the Court's judgment; for example, it is hardly the case that huge restaurant chains, broadly successful retailers, most large-scale manufacturers, major providers of cell phone service, leading banks post deregulation (eliminating barriers to cross-state expansion), oil companies, and other substantial firms, even in concentrated industries, routinely stick to geographically segregated territories. Moreover, even supposing that an express agreement must ultimately be demonstrated for plaintiffs to prevail, inferring its likelihood from industry conditions is an extremely complex, fact-intensive undertaking. *Cf*. Kaplow, *supra* note 118, at 488–508 (analyzing the problem with regard to coordinated oligopolistic price elevation).

A further complication in *Twombly* is that the meaning of the legal standard, which both the majority and dissent took to be uncontested, is murky and was stated in inconsistent ways in the Court's opinion, so the question of whether an illegal agreement was plausibly suggested not only raises complex, context-specific factual issues but also

There is one set of components of the optimal decision criterion that is within judicial experience (but not common sense): adjudication costs. Even here, however, judges' knowledge will often be incomplete. Many cases, especially those that are most complex, will be novel, at least for a particular sitting judge. In addition, judges do not directly see, for example, the costs of discovery, even when they have supervised its conduct in prior cases. Most of the expense, and especially any disruption, is out of judges' view. To be sure, they will hear parties' protests about their magnitude, and the oppositions' pooh-poohing thereof, but unless a judge has prior practice experience working with the sorts of parties bearing these costs, only rough, somewhat speculative guesstimates will be possible. A further irony is that some of the impetus for the Supreme Court's decisions in *Twombly* and *Iqbal* concerns judges' difficulty in managing discovery, which itself is partly a product of these very limitations, whereas similar knowledge is required to decide the motion to dismiss.[224]

Considering all dimensions of the decisionmaking problem, it is clear that in many settings in many areas of law, the informational challenges are daunting, much more so than seems to be recognized. In principle, a number of approaches might be employed, individually or in combination, in attempts to address the problem. Present discussion is confined to a brief sketch of some possibilities.[225] (Those concerning the manner in which adjudication is conducted in existing trial courts are deferred momentarily.) In considering different strategies, it is important to keep in mind that the magnitudes of the multiple factors that enter into the optimal decision rule are likely to vary tremendously not only across fields of law but also between individual cases, implying that more capable decisionmakers and not just more finely tuned rules capable of ready application are necessary to achieve many of the potential social gains.

---

posses a legal conundrum. The intersection of these two — neither directly confronted by the Court — adds to the difficulty of interpreting *Twombly*. *See generally* Kaplow, *supra* note 71 (examining the agreement requirement in general and how it was addressed by the opinions in *Twombly*). Interestingly, in *Iqbal* there was also (in that case explicit) disagreement about the legal standard, so there as well there is an additional level of complication in attempting to extract a clear understanding of the import of the case for the general legal standard for motions to dismiss.

[224]For example, *Twombly*, 550 U.S. at 560 n.6, quotes extensively from Judge Easterbrook, *supra* note 98, at 638–39, on judges' limitations, without appreciating that they may bear on the ability to make the decision at hand concerning whether to allow discovery in the first place: "The judicial officer always knows less than the parties, and the parties themselves may not know very well where they are going or what they expect to find." *Id*. at 638. "Judicial officers cannot measure the costs and benefits to the requester . . . ." *Id.* "[I]t is no wonder that the magistrates answered 'no' when Judge Weinstein asked them whether there is abuse in the Eastern District of New York. They have no way to evaluate the costs and benefits of discovery ex ante, and they rarely examine their handiwork ex post (because the case either settles or passes to the judge for disposition)." *Id*. at 639 (offering a further statement beyond the passage quoted in *Twombly*). There is an interesting contrast in the experiences of discovery management versus decisionmaking on motions to dismiss that may help to explain this tension. With the former, a trial court judge would have the ongoing experience of being faced with difficult decisions that involve a degree of micromanagement, whereas the latter requires only a single, all-or-nothing decision. It is also true that effects on deterrence and chilling, which greatly complicate making good termination/continuation decisions, are perhaps, to an even greater extent than the many costs of complying with discovery requests, less salient to judges making these decisions, so their complexity seems less bothersome, particularly if one does not in fact attempt to analyze them very carefully. Aspects of these differences are addressed further in section D.

[225]The reader is reminded that no reforms are being advocated here. The aim is merely to instigate further thought on questions of system design. Furthermore, even the brief attention offered is confined to the question at hand, ignoring that most of the possibilities examined may have other (possibly more important) benefits or costs.

First, one might turn to specialized courts. This strategy is used in varying degrees in some settings. In the United States, there is a separate tax court (but not with exclusive jurisdiction), the Federal Circuit for patent cases (but only for appeals), bankruptcy courts (but not specialized by industry and that have jurisdiction to consider issues in many areas of law), business courts (in some states, but ordinarily having broad jurisdiction), probate courts, and tribunals within many agencies or executive departments (such as for the determination of social security disability claims or immigration matters). Some of these courts have fairly narrow domains, which enables them to develop relevant expertise concerning pertinent empirical matters, but most do not. Greater specialization is possible, and systems of a given degree of generality could also assign cases to particular judges based on specific competency. If that were done, then the usefulness of "judicial experience" in the present setting would be considerably greater.

Expert agencies offer another route, they are often employed in parallel or exclusively, and they can potentially address the informational challenge in a number of ways. They might handle matters entirely, which is more often done outside the United States. Or early stages might be conducted internally, with informal or formal interim termination/continuation decisions, leaving subsequent stages to the courts. They might also act in ways designed directly to offer further information to courts that operate independently, such as by promulgating regulations, issuing reports, enhancing access to defendants' information in particular cases, or intervening (as a party, as an amicus, or even more informally) to offer a perspective on some of the empirical questions pertinent to a court's decision.[226] Recognizing the limitations of "judicial experience and common sense," agencies might act in ways that would enhance judges' information base. That is, they might attempt to transfer to courts some of the experience and expertise of their specialized staff (such as lawyers, economists, or scientists) that already guides their own screening and related decisions, regulation-writing, and so forth. Note also that, if an agency does undertake its own investigations and either pursues its own action in court or makes the information it obtains available to private parties, then rather different inferences and outcomes would be appropriate when, at the time of a motion to dismiss, an enforcer has little information suggestive of liability.[227]

It is also possible for legislation to address the problem. Section III.B examined whether

---

[226]For example, one of the factors under the optimal decision rule concerns the magnitude of the deterrence deficit — a factor influenced by the legal system as a whole, including agencies' own enforcement actions; *see supra* subsection II.A.3 — and this is a matter on which an agency should ordinarily have expertise, certainly more than possessed by a generalist court. (Also, a court may find it awkward on its own to assess the degree of enforcement success of a pertinent agency. On the other hand, if the reason for parallel private enforcement is to ensure aggressive action even when an agency might be captured by those it regulates, then relying on an agency's assurances that there is little problem would be inappropriate.) To take another illustration, although an agency may well lack the authority to promulgate regulations that would bind federal courts in the manner done by the Private Securities Litigation Reform Act of 1995, Pub. L. No. 104-67, 109 Stat. 737 (codified as amended in scattered sections of 15 U.S.C.), one could imagine the agency undertaking a general investigation of private litigation and issuing an opinion concerning the extent of meritless suits, what factors seem to be indicative of cases' merits, and other matters that would inform a trial court's plausibility assessment even if no special pleading rule is implemented.

[227]Keep in mind that the continuation cost may also be lower in such instances if much of discovery has essentially already been conducted.

it is better to alter substantive rules (such as elements of offenses) or procedures (such as through the Private Securities Litigation Reform Act of 1995[228] or Federal Rule of Civil Procedure 9(b) on allegations of fraud or mistake).[229] Regardless of which route is best or is in any event pursued, the obvious advantage is that the sort of information that is appropriate in determining how to make the relevant decisions may be brought to bear.[230] This approach, however, is only available to the extent that generalizations are possible, and hence it cannot readily address the many factual variations within an area of law.[231] Of course, even when there is substantial case-specific variation, one might attempt to adjust a decision threshold up or down to reflect averages across a given pool of prospective cases.

Another approach is that, over time, courts themselves may refine the application of rules in different legal contexts through what is essentially a common law process. As a consequence, what must be pled, or demonstrated to survive a motion for summary judgment, has long been understood to differ in various ways across such fields as employment discrimination, patents, antitrust (for example, the need to plead so-called plus factors in cases alleging horizontal agreement, at issue in *Twombly*, or to plead market definition[232]), securities law (even before the 1995 Act), and many others.[233] An obvious advantage is that collective wisdom and experience may greatly exceed that of any single trial court judge, especially one who is newly appointed or who, even after many years on the bench supplemented by substantial practice as a lawyer, may never have confronted the type of case at hand. This method of refinement also has obvious

---

[228]The Conference Report stated the Act's motivation as follows:

> Congress has been prompted by significant evidence of abuse in private securities lawsuits[, which] include: (1) the routine filing of lawsuits against issuers of securities and others whenever there is a significant change in an issuer's stock price, without regard to any underlying culpability of the issuer, and with only faint hope that the discovery process might lead eventually to some plausible cause of action; (2) the targeting of deep pocket defendants . . . without regard to their actual culpability; [and] (3) the abuse of the discovery process to impose costs so burdensome that it is often economical for the victimized party to settle . . . . At the same time, the investing public and the entire U.S. economy have been injured by the unwillingness of the best qualified persons to serve on boards of directors and of issuers to discuss publicly their future prospects, because of fear of baseless and extortionate securities lawsuits.
>
> In these and other examples of abusive and manipulative securities litigation, innocent parties are often forced to pay exorbitant "settlements."

H. R. REP. NO. 104–369, at 31–32 (1995) (conference report). In addition to requiring more specific pleadings, the Act provides for a stay of discovery pending decision on a motion to dismiss (thereby removing the option described in section D of intentionally delaying ruling to allow some discovery). 15 U.S.C. § 78u-4(b)(3)(B) (2006).

[229]*See* Clermont & Yeazell, *supra* note 10, at 854–55 (discussing expansion of the scope of Rule 9(b) to other substantive areas as an alternative to the general pleading standard promulgated in *Twombly* and *Iqbal*). The latter — which uses the Advisory Committee, judicial approval, and congressional review — is an instance of a quasi-legislative process, which will not be distinguished for purposes of this brief overview of possibilities.

[230]Other issues are set to the side, such as the susceptibility to capture by special interests on either side of an issue in the legislative process versus in agencies or courts.

[231]*See generally* Kaplow, *supra* note 121 (exploring the promulgation of substantive rules as information generation and dissemination).

[232]*See, e.g.*, EARL KINTNER, FEDERAL ANTITRUST LAW § 10.16 (1980).

[233]*See* sources cited *supra* note 176. On reflection, it is apparent that most published opinions in federal court involve decisions on motions to dismiss, summary judgment, and judgment as a matter of law, indicating that most court-made law is generated in these contexts and suggesting that much of that law will have particular implications for the decision standards in these settings.

limitations, among them the problem that if essentially none of the judges who have decided past cases really has pertinent knowledge on key dimensions, pooling their decisions cannot fill the gap. Likewise, systems of trial and error only generate advances when the successes and failures can be identified ex post, whereas most of the effects of termination/continuation decisions — particularly regarding deterrence and chilling — will never be observed by courts. And, as mentioned, even continuation costs are not necessarily perceived; also, to the extent that they are, they are not ordinarily the subject of judicial opinions that permit wisdom to accumulate and spread. Furthermore, settlements significantly obscure the legal system's effects from those who operate it.

Whatever systemic responses might be undertaken, at any particular point in time in a given area of law courts may need to make decisions for which they are not well equipped by their experience and common sense, and parties will seek to influence those decisions. Consider the predicament of a plaintiff who fears its case will be deemed too speculative to satisfy the plausibility requirement for surviving a motion to dismiss. Lawyers already, in some settings and to varying extents, advance ideas and present information that suggest a more favorable context, such as in describing background conditions or quoting sources in a complaint and referencing data and expert publications in briefs. Given what appears to be the current decision criterion and adverse outcomes in cases like *Twombly* and *Iqbal* themselves, it is unclear why advocates would stop there.[234] One might hire an expert before filing a complaint (something already done in some settings), have a report prepared, and insert it in various ways — attached to a brief or complaint, or even included in the body of the complaint itself.[235] If indeed the context-specific assessment depends on empirical questions largely beyond general knowledge, it may make sense to try to influence the decisionmaker's views regarding such matters.[236] Would a trial judge feel confident deeming a scenario to be implausible when a detailed expert report explained that it was likely?[237] Recall the medical malpractice illustration in section B and

---

[234]*See, e.g.*, Allen & Guy, *supra* note 183, at 30 ("[T]hey also have the opportunity to meet that burden by producing more 'evidence' in their pleadings. If parties fear a 'biased' reaction to a bare bones pleading, they can provide considerably more than they otherwise would."); Edward A. Hartnett, *Taming* Twombly*, Even After* Iqbal, 158 U. PA. L. REV. 473, 474–75 (2010) ("Second, the *Twombly* framework can be treated as an invitation to present information and argument designed to dislodge a judge's baseline assumptions about what is natural.").

[235]For prior suggestions of this possibility, see Kaplow, *supra* note 71, at 740 n.143; Louis Kaplow, *An Economic Approach to Price Fixing*, 77 ANTITRUST L.J. 343, 437 n.225 (2011). This possibility and others discussed in the text to follow raise the question of the circumstances under which a motion to dismiss might need to be treated as one for summary judgment in accord with FED. R. CIV. P. 12(d). That rule so requires when the court considers matters "outside the pleadings," suggesting that if the report's content was part of the complaint, the conversion would not be triggered. Furthermore, material can be influential without being deemed evidence that is part of a record, as suggested by the citation of expert and other non-strictly-legal sources in briefs and court opinions.

[236]Note that some of the empirical questions underlying the optimal decision rules in Part II are the sorts about which expert witnesses are often used whereas others (such as those pertaining to the prevailing level of deterrence) are not. As mentioned earlier in this section, specialized agencies and the like might be most useful along the latter dimensions.

[237]Suppose that the plaintiff had argued: "While our complaint's allegations of facts A, B, and C may seem innocuous to the uninitiated, experts in the field widely regard them as red flags. See the report of expert Lee, quoted in (or attached to) our complaint."

imagine various probability levels that might be submitted.[238]

With or without such proffers, defendants might behave likewise[239] — and, of course, already do so to some extent when they cite various background facts and articles in their own briefs in support of their motions.[240] And courts might do the same. Actually, of course, they already do, although the point does not receive much explicit attention in commentary on procedure. For example, in *Twombly* itself, the Supreme Court both referred to specialized sources and also had the benefit of amicus briefs containing much more material of this type, in addition to what was in the parties' briefs.[241] If an amicus brief by an industry, advocacy group, or independent experts is permissible and believed to have potential impact on what is, after all, the review of a decision on a motion to dismiss, we might expect parties at the outset to find it in their interest to present this type of content. Courts, perhaps well aware of the limitations of their direct knowledge of complex matters about which they have little experience, may listen — or, with the aid of their clerks, identify additional literature and cite its findings to bolster their opinions. Even if such material is not specifically referenced, parties may nevertheless act on the possibility that it would have a positive impact on their prospects.

It is also interesting to speculate how such developments would affect the quality of decisions at early stages. It may seem that additional information, particularly expert information focused on the scenario at hand, could only enhance the ability to make context-specific decisions when the decisionmaker initially knows little about the circumstances. An obvious shortcoming is that it is often possible for a party to hire a purported expert to opine

---

[238]As this example suggests, there will sometimes be important limitations on what an expert report can address because only the plaintiff's information and generalized knowledge may be available. Perhaps, without discovery, an expert could suggest the likelihood of malpractice but not be able to say much about precisely what mistakes were made. A similar distinction runs through the *Twombly* opinion, where the Court was unclear about how it thought a conspiracy might be proved when deciding a motion to dismiss: could it be shown purely through evidence of industry conditions and other circumstantial evidence (which an expert could address at that point and which, as mentioned earlier in the text, can be sufficient at trial under existing law), or must a complaint detail who said what to whom and on what date (which an industry expert could not illuminate)? Another point that has received little direct attention is that it may make sense to rely on certain types of evidence at early stages (before discovery) while insisting on different types of evidence at later stages. Outside the legal system, it is of course routine to make many preliminary decisions, including about whether to investigate a matter further, using generalized background knowledge but to choose ultimate actions based on the particulars subsequently revealed.

[239]"Although the plaintiffs would have you believe that facts *A*, *B*, and *C* are suspicious, experts in the field take them to be entirely typical of ordinary, benign conduct. See expert Shin's report."

[240]Other types of responses might be available to defendants in some instances. For example, if a plaintiff claims a need to discover information solely in the defendant's possession, a defendant could opt to provide the key information voluntarily, and then argue that the plaintiff's argument should be viewed as particularly implausible (since even armed with the most pertinent evidence it still has a flimsy case) and, moreover, that the plausibility threshold should implicitly be higher because the need for discovery has been reduced. Note further that if this practice became common, a judge might, as suggested in note 111, draw negative inferences against defendants who would not freely share such information. Both possibilities relate to the discussion in section D of judges delaying rulings while permitting limited discovery.

[241]*Cf.* Richard A. Epstein, Bell Atlantic v. Twombly*: How Motions to Dismiss Become (Disguised) Summary Judgments*, 25 WASH. U. J.L. & POL'Y 61 (2007) (arguing that the Supreme Court reached a correct outcome in *Twombly* by using something closer to a summary judgment framework); Epstein, *supra* note 16, at 199–207 (suggesting that the plaintiffs in *Twombly* failed to undertake adequate pre-filing investigation, whereas in *Iqbal* there was no real information that the plaintiff could have gleaned from public records).

favorably regardless of the actual state of affairs. Of course, if a judge, aided by whatever a defendant might offer in response, is unable to make sense of such a report for purposes of deciding the motion, it is unclear how the same decision could sensibly be made acting in a vacuum.[242] It is also possible that additional up-front investments by plaintiffs would serve some screening function.[243] Finally, all of these sorts of developments would tend to add to the costs of deciding motions,[244] although, aside from the caliber of the decisions made, the additional expenditures might focus subsequent development[245] and primarily serve to accelerate activity that would have come later in any event.[246]

D. Judicial Discretion

In federal civil litigation in the United States, judges have a great deal of discretion, including with regard to motions to dismiss. They may deny such motions, allowing cases to continue, without review because these decisions are not appealable (in the absence of special provisions or actions). Moreover, when cases continue, there is a high probability of settlement and thus a low likelihood that the judge will have to make any subsequent ruling that may then be appealed (and, even in that instance, the previous denial of the motion to dismiss would not then be examined). Accordingly, even if *Twombly* and *Iqbal* are regarded as having raised the threshold for continuation, lower court judges can de facto proceed as they always have. Put another way, there is a sense in which the standard for motions to dismiss forcibly binds lower court judges only when they wish to grant such motions. In this regard, one can view the decision rule as providing judges an option of granting these motions whenever the case falls below the cutoff.[247]

This structure accords additional leeway that a trial court judge could choose to use in

---

[242]Judges might make use of court-appointed experts, authorized by FED. R. EVID. 706, or special masters having appropriate expertise, but currently this is rarely done even at later stages of adjudication.

[243]In part, strike suits are aided by the ability to undertake negligible expenditures while imposing significant costs on defendants, so such suits may be less promising when plaintiffs need to make more expenditures at the outset. *See supra* subsection III.D.1. Also, the revealed willingness to make such expenditures before filing may tend to indicate credibility (although it is also true that sinking costs is a commitment strategy that can have some success even independent of the merits).

[244]*See, e.g.*, BONE, *supra* note 32, at 146; Bone, *supra* note 12, at 5. System costs will also be influenced by how many motions are filed. It is not obvious whether raising the bar induces more or fewer motions, which will depend on how many cases are close to the new standard versus the old one and also on how the rule change influences case filings (*see supra* subsection III.D.1). In this regard, it is important to keep in mind that adjudication costs play an important and multidimensional role in the analysis of the optimal decision rule in Part II (which abstracts from the cost of the decision itself).

[245]Clermont & Yeazell, *supra* note 10, at 840–41, object that the consequence of the *Twombly-Iqbal* standard may be to give defendants a cheap form of discovery in cases that will be continued, but it is unclear why this savings (which may come in substantial part from a clarification of issues, essentially boosting the notice function of pleadings) is considered a defect rather than a virtue.

[246]These latter points are particularly significant if the comparison is to a very low threshold that does not require much attention to facts and results in most cases being continued, at which point expert reports and other efforts would be required.

[247]Realistically, the location of the cutoff is unclear, and a district court judge cannot fully predict the decision of the appellate panel (itself a random subset of judges on the appellate court). Also, subsequent review of their decisions (en banc, to the entire appellate court, or through certiorari to the Supreme Court) is fairly rare.

addressing the dilemma of either dismissing what may actually be a meritorious case or continuing at great cost what may really be a frivolous case. Specifically, judges can delay their ruling on a motion to dismiss and, in parallel, allow discovery to proceed.[248] Moreover, in the interim, a judge who wishes to do so can manage that discovery so as to keep it to a minimum, focusing only on key documents or witnesses that may be thought critical in determining whether the case should be terminated or continued.[249]

In this manner, a judge could replace combination with separation early in the legal proceeding, dividing into two segments the stage between a traditional motion to dismiss — decided with no discovery at all — and a motion for summary judgment — decided after full discovery.[250] The merits of additional separation into distinct stages were assessed in section III.A. The most obvious attraction would lie, as just suggested, in cases in which a small quantity of highly probative evidence is solely in the defendant's possession, but denying the motion to dismiss and thereby authorizing full discovery could involve large impositional

---

[248]*See, e.g.*, Swanson v. Citibank, N.A., 614 F.3d 400, 412 (7th Cir. 2010) (Posner, J., dissenting) ("If the plaintiff shows that he can't conduct an even minimally adequate investigation without limited discovery, the judge presumably can allow that discovery, meanwhile deferring ruling on the defendant's motion to dismiss."); Miller v. Gammie, 335 F.3d 889, 899 (9th Cir. 2003) (en banc) ("Under the functional analysis laid out by the Supreme Court, the district court did not err when it deferred ruling on the motion to dismiss on the pleadings until the nature of the functions the defendants allegedly performed was sufficiently outlined . . . .") (note that the opinion predates *Twombly* and *Iqbal*); 5 CHARLES ALAN WRIGHT & ARTHUR R. MILLER, FEDERAL PRACTICE AND PROCEDURE 181–82 n.90 (3d ed. Supp. 2011) (citing cases taking different stands on the post-*Twombly*/*Iqbal* permissibility of discovery to support pleadings); Bone, *supra* note 16, at 932–33 ("There is precedent, however, for a more promising approach based on supplementing strict pleading with limited pre-dismissal discovery, and recently at least one trial judge has indicated a willingness to use this approach to address the information-access problems raised by *Twombly*'s plausibility standard."); *id*. at 933 n.250 (citing a case permitting such discovery and another disallowing it); *id*. at 935 ("Allowing pleading-stage discovery fits the current Rules awkwardly at best. Moreover, with a new rule, the procedure can be designed optimally and the provisions applied uniformly to all district courts."); Hartnett, *supra* note 234, at 509–10 ("Instead, the district court could deny the motion to stay discovery (or grant a motion to compel) and delay decision (either purposefully or simply due to competing priorities) on the motion to dismiss. . . . [Or, at a Rule 16 conference, t]he court could allow limited discovery, targeted at the identified allegation, and establish a briefing schedule for any motion to dismiss that follows the completion of that limited discovery. This scenario tames *Twombly* rather thoroughly; indeed, it resembles what Justice Stevens envisioned in his dissenting opinion in that case."); *id*. at 511 ("Lest anyone think that such an approach cannot be right because it guts rather than merely tames *Twombly*, bear in mind that the Federal Rules explicitly authorize a district court to defer hearing and decision on a 12(b)(6) motion until trial."); *id*. at 513–14 (noting that this use of discretion is effectively nonreviewable by an appellate court); William H. Page, Twombly *and Communication: The Emerging Definition of Concerted Action Under the New Pleading Standards*, 5 J. COMPETITION L. & ECON. 439, 466–68 (2009). A defendant can file a motion to stay discovery pending decision on the motion, but that ruling also can be delayed or the motion can be denied, in either instance with no review. Indeed, discovery stays are not routinely granted. *See, e.g.*, Hartnett, *supra* note 234, at 507–08. As mentioned in note 228, a significant feature of the Private Securities Litigation Reform act is to stay discovery pending a decision on a motion to dismiss.

[249]In addition to formal orders, a judge has a powerful threat (which does not need to be stated explicitly) over a plaintiff, who will know that the judge could stay discovery and decide the motion for the defendant if the plaintiff, in the judge's view, takes inappropriate advantage of the forbearance that has been accorded.

[250]*See also* Bone, *supra* note 16, at 933 n.251 ("Another possibility is to give discretion to the trial judge to stage discovery in increments, evaluating the strength of the case after each stage. . . . I am skeptical, however, that trial judges can make good decisions about discovery's benefits to be able to implement a staged approach like this effectively in a complex case. Moreover, additional discovery increases costs and enhances the plaintiff's settlement leverage."); *id*. at 934 ("In addition, it is important that the discovery be limited in a clear way. . . . One possibility is to give the plaintiff the option to take one deposition of each defendant and perhaps serve a narrowly tailored request for documents. The objective is to ensure that discovery is not so costly that it pressures the defendant to settle before it even takes place.").

costs.[251] Moreover, this option owes its existence to a nontrivial hurdle at the motion to dismiss stage, for if a judge has no realistic ability to dismiss the case, then delaying that ruling does not hold a significant threat over the plaintiff that may be used to induce sequenced, highly limited initial discovery.[252] Hence, one perspective on *Twombly* and *Iqbal*, to the extent that they do elevate the threshold for denying motions to dismiss, is that they create or enhance this alternative for district court judges who, as mentioned, are still de facto able to deny motions outright if they insist on doing so.[253] A further implication is that the criticism that the cases foreclose access to courts when key information is solely in defendants' hands, sometimes accompanied by proposals to allow perhaps limited discovery in such instances,[254] may be of less force to the extent that the aforementioned avenue is already available.[255]

Consider next how trial court judges might be inclined to use their latitude. A natural conjecture is that the answer will vary greatly by individual judge[256] and type of case in ways that will reflect substantial idiosyncracy that cannot readily be captured by oft-discussed ideological bias[257] — although that too, when present, will matter in light of the discretion that is

---

[251]Of course, in many cases it may be that important information is largely in defendants' hands, yet highly limited discovery — perhaps the inspection of a few documents and one or two depositions — would not be sufficient to ascertain the facts. For example, despite the contrary suggestion in AREEDA & HOVENKAMP, *supra* note 71, at 100–01, it would not generally be possible to ascertain "with only limited discovery" the profitability of hypothetical business decisions that defendants might otherwise have made.

[252]Prior discussion of the option of delaying a ruling to permit limited discovery, *see* sources cited *supra* note 248, generally does not emphasize this point. Even without the threat of granting a motion to dismiss, the ruling on which has been delayed, judges could greatly limit discovery through active management, keeping in mind that the option discussed in the text does presume some management. This route would require, however, that if the initial, key information suggested that the case was frivolous, then the judge would essentially have to cut off further discovery despite its seemingly being authorized by the procedural rules. Perhaps in certain circumstances this could be accomplished by threatening the use of Rule 11 sanctions. In some Continental systems and some modes of arbitration, in which factual development is more sequenced and where discretionary cost shifting is more widely authorized and employed, this sort of method may succeed in substantially reducing impositional costs.

[253]In offering this perspective, it is not suggested that the Supreme Court envisioned or intended this result.

[254]*See, e.g.*, Ray Worthy Campbell, *Getting a Clue: Two Stage Complaint Pleading as a Solution to the* Conley–Iqbal *Dilemma*, 114 PENN ST. L. REV. 1191 (2010); Scott Dodson, *New Pleading, New Discovery*, 109 MICH. L. REV. 53 (2010); Suzette M. Malveaux, *Front Loading and Heavy Lifting: How Pre-dismissal Discovery Can Address the Detrimental Effect of* Iqbal *on Civil Rights Cases*, 14 LEWIS & CLARK L. REV. 65 (2010) (expressing concern about the potential impact of *Iqbal* on civil rights cases but arguing that existing procedures permit a trial court judge to allow limited discovery before deciding a motion to dismiss); *see also* Bone, *supra* note 12, at 7 ("If strict pleading is desirable for some cases, moreover, it should be coupled with limited access to discovery before dismissal, such as one deposition of each defendant and perhaps a fixed request for a fixed number of specifically identified documents.").

[255]There are differences. At present, a motion to dismiss might be granted if a case is deemed to fail the plausibility test even though minimal discovery would be highly probative. However, given the open-ended nature of that test, as discussed in section A, such a denial could be reversed, and a trial court worried about that possibility may find it prudent to allow some carefully circumscribed discovery before dismissing a case of uncertain plausibility, which doubts may be resolved fairly readily and at low cost.

[256]*Cf.* MOORE'S FEDERAL PRACTICE, *supra* note 22, at 31 ("[S]ome judges enjoy deciding motions. They give summary judgment motions prompt and reflective attention. . . . At the opposite extreme are judges who dislike motion practice and even avoid it by allowing decisions on motions to pile up . . . .").

[257]*See, e.g.*, JEFFREY SEGAL & HAROLD SPAETH, THE SUPREME COURT AND THE ATTITUDINAL MODEL REVISITED (2002); CASS R. SUNSTEIN, DAVID SCHKADE, LISA M. ELLMAN & ANDREW SAWICKI, ARE JUDGES POLITICAL? AN EMPIRICAL ANALYSIS OF THE FEDERAL JUDICIARY (2006); Myron H. Bright, *Do Philosophy and Oral Arguments Influence Decisions?*, 77 A.B.A. J. 68 (1991) (in Eighth Circuit Judge Bright's survey of his colleagues, which presented

available.[258]

Dismissal helps judges clear their dockets. Furthermore, if a judge anticipates substantial discovery disputes that may need to be adjudicated, dismissal may be especially appealing. Additionally, judges' tastes (to put it bluntly) for different types of lawsuits vary greatly: what one judge sees as extremely important may to another judge be terribly boring. Perhaps many complex business disputes have this feature in the eyes of many.

Continuation also has its attractions. As mentioned, denying a motion to dismiss avoids appellate review and the possibility of being reversed. Moreover, because of this fact, one may find it much easier to write an opinion denying the motion. Because most cases will settle before any further need for a dispositive ruling, the overall workload may be less. Judges also vary regarding their attentiveness to discovery and other pre-trial wrangling; some remain largely aloof, either not ruling on motions, deciding them quickly, or in various ways signaling that they are unwelcome.

For complex cases in particular, most of the factors on both sides are magnified, the direct implications of which are unclear. Preferences about case types may be much stronger when a great deal of time needs to be spent, whether writing a defensible opinion dismissing a case or subsequently supervising it. Also, more mundane considerations, such as the tendency to procrastinate, bear mention. For court systems that monitor the number of unresolved motions (often broken down in categories by how long they have been unresolved), a pending decision on a motion to dismiss for a massive dispute may count the same as one for a minor dispute that can be decided almost instantly. Hence, nonruling, at least for extended periods to time — and quite possibly without any discovery stay — may frequently occur in complex litigation. In this regard, recall the observation in section C that many of the costs of discovery are out of judges' sight and thus perhaps largely out of mind as well. During even a few months before a court gets around to making a ruling, millions of dollars (or, in a few of the largest cases, much more) may be spent and much disruption incurred.

The great degree of discretion — which is likewise present in case management whenever motions to dismiss are denied — raises concerns about abuse and more mundane but frequent problems of decision quality and cost.[259] A possible, partial remedy would be to allow

---

them with a summary judgment case, finding that their decisions differed greatly as a function of the ideological and jurisprudential orientation of each judge); Theodore W. Ruger, Pauline T. Kim, Andrew D. Martin & Kevin M. Quinn, *The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking*, 104 COLUM. L. REV. 1150 (2004).

[258] *See* Stephen R. Burbank, *Pleading and the Dilemmas of Modern American Procedure*, 93 JUDICATURE 109, 117–18 (2009).

[259] Concerns expressed about consistency (*cf.* WRIGHT & MILLER, *supra* note 248, at 68 ("The subjectivity at the heart of *Twombly* and *Iqbal*, as embodied by 'judicial experience and common sense,' raises the concern that rulings on motions to dismiss may lead to inconsistent rulings on virtually identical complaints.")) and legitimacy, when examined more carefully, often involve the problems noted in the text. *See* Kaplow & Shavell, *supra* note 33, at 1222–23, 1328–29; Kaplow, *supra* note 33, at 395–96. Related, the use of open-ended and unelaborated criteria reduces transparency, which in turn inhibits accountability as well as the prospects for coherent rule refinement over time and the ability to conduct legal proceedings in a manner that induces the presentation of the most pertinent information to guide

interlocutory appeals of denials of motions to dismiss, or perhaps even of refusals to stay discovery in settings in which unfettered discovery may be particularly costly, although it would be necessary for such a channel to be rapid in order to limit strategic behavior by defendants.[260] The concern with constraining judicial discretion is related to protection of the right to a jury trial, noted further toward the end of the next section.

E. Summary Judgment

Under Rule 56(a) of the Federal Rules of Civil Procedure, "[t]he court shall grant summary judgment if the movant shows that there is no genuine dispute as to any material fact and the movant is entitled to judgment as a matter of law."[261] This section focuses on what is meant by "no genuine dispute as to any material fact," addressing questions similar to those examined in sections A and B with regard to the plausibility requirement to survive a motion to dismiss.[262]

As stated in the Introduction, this test is rather muddy.[263] *Anderson* interpreted the

---

sensible decisionmaking in individual cases. The current predicament may in part reflect judges' reluctance to be blunt about the demands imposed by their complex task, appreciating that onlookers will be aware of their limitations in meeting them.

[260]*See also* Gideon Mark, *Federal Discovery Stays*, 45 U. MICH. J. L. REFORM 405 (2012) (proposing that mandatory stays of electronic discovery be the norm in federal civil litigation). In addition to providing for expedited review, defendants could be required to bear a nontrivial cost if the reviewing court deems the interim appeal to have been unwarranted. Obviously, a number of considerations enter into whether interlocutory appeals should be allowed in various contexts. New York State does allow such appeals of denials of summary judgment (which seems much less valuable), but it is unclear whether the provision has had much impact. *See* Thomas R. Newman & Steven J. Ahmuty Jr., *Review of Denial of Summary Judgment on Appeal After Trial*, N.Y.L.J., Jan. 5, 2005 ("[U]nless you can persuade either the trial judge or the Appellate Division to grant a stay of the trial pending determination of the appeal (generally, a futile task), there is no point in perfecting the appeal from the intermediate order."). Whether considering interlocutory appeals or appeals of grants of motions to dismiss or for summary judgment, appellate courts operate at a disadvantage to the extent that their the decision rule is fact intensive and there were substantial prior proceedings that substantially exposed the facts of the case to the district judge who made the initial decision.

[261]FED. R. CIV. P. 56(a). Regarding the change in the 2010 amendments from the long-familiar language of "genuine issue" to the new "genuine dispute," the Advisory Committee states: "The standard for granting summary judgment remains unchanged. . . . Subdivision (a) carries forward the summary-judgment standard expressed in former subdivision (c), changing only one word — genuine 'issue' becomes genuine 'dispute.' 'Dispute' better reflects the focus of a summary-judgment determination." *Id*., advisory committee's note.

[262]The issues raised in sections C and D also have obvious relevance. Depending on how the summary judgment test is interpreted, many of the same sorts of informational challenges could be present (although, obviously, not those concerned with predicting the cost of discovery, which ordinarily would be completed at that time). Regarding judicial discretion, the factors bearing on judges' motivations are relevant, although with different weight, in the present context. Here, continuation may, despite the nontrivial possibility of a settlement in the interim, produce a trial, which could consume substantial time from the judge, whereas the amount of energy that must be devoted to resolving discovery disputes would ordinarily be far less. At the summary judgment stage, the primary attraction of delaying any ruling is that settlement may render a decision unnecessary.

[263]*See* sources quoted *supra* notes 18 & 24; *see also supra* note 87 (discussing how it is not obvious whether the *Twombly*-*Iqbal* plausibility requirement is tougher or weaker than the requirement of a genuine dispute about a material fact); *supra* section II.D (analyzing the optimal relationship between the strength of termination/continuation rules at different stages of adjudication).

requirement as congruent with Rule 50's standard for judgment as a matter of law,[264] which directs an inquiry into whether "a reasonable jury would . . . have a legally sufficient evidentiary basis to find for the party on that issue."[265] That is, the evidence presented by the nonmoving party is deemed legally sufficient to permit resolution by the factfinder when that evidence provides the factfinder with a legally sufficient basis to find for the party in question. Clearly, explication is in order.

To begin, consider the ordinary meaning of the key adjective "genuine," the most pertinent of which is real, true, or actual as distinguished from pretended or insincere.[266] This meaning must be joined with the modified term, dispute, which simply means a disagreement, argument, or debate. A literal interpretation of the two-word phrase is that the party opposing the summary judgment motion truly means to disagree with the movant's claims, which could readily be so even when there is no factual basis, but merely a frank desire to disagree, with the ultimate hope of victory (perhaps as a product of factfinder confusion or nullification) or simply to prolong the case in order to extract a settlement. The responding party's expenditure of effort to defeat the motion might be taken as powerful proof that it truly disagrees in this sense.

Because the context clearly suggests that a tougher standard is envisioned, the genuine dispute requirement presumably demands something more, something qualitatively different. Rule 50's formulation, although circular, does helpfully direct us to consider the sufficiency of the evidence (even though it is unilluminating as to the requisite quantity).[267] The question then

---

[264]Anderson v. Liberty Lobby, Inc., 477 U.S. 242, 249–50 (1986). On the change in language from "directed verdict" to "judgment as a matter of law," see note 20.

[265]FED. R. CIV. P. 50(a)(1); *see also supra* note 22 (commenting on the uncertain meaning of "reasonable jury"); *supra* note 23 (discussing the interpretation offered in MOORE'S FEDERAL PRACTICE, *supra* note 22, § 56.22). *Anderson*'s further elaboration is equivalent for all practical purposes:

> Formerly it was held that, if there was what is called a scintilla of evidence in support of a case, the judge was bound to leave it to the jury, but recent decisions of high authority have established a more reasonable rule, that in every case, before the evidence is left to the jury, there is a preliminary question for the judge, not whether there is literally no evidence, but whether there is any upon which a jury could *properly proceed to find a verdict* for the party producing it, upon whom the onus of proof is imposed.

477 U.S. at 251 (emphasis added). Similarly unilluminating are some other traditional statements of the rule. *See, e.g.*, 9B CHARLES ALAN WRIGHT & ARTHUR R. MILLER, FEDERAL PRACTICE AND PROCEDURE § 2524 (2011) ("The question is not whether there is literally no evidence supporting the party against whom the motion is directed but whether there is evidence upon which the jury might reasonably find a verdict for that party."); *id*. ("The question of evaluating the sufficiency of the evidence on a motion under Rule 50, when viewed in this way, is one that has been the subject of a great variety of verbal formulations, many of them couched in generalities that unfortunately can not be applied readily to any particular set of facts."); Cooper, *supra* note 18, at 919 (stating the question as whether a contrary jury verdict would be "against the weight of the evidence"); *id*. at 920 ("The most common current approach is often dubbed the 'substantial evidence' test . . . ."); *id*. at 921 ("In essence, this approach represents an attempt to limit the jury to its factfinding function by inquiring what is 'reasonable.'"); *see also id*. at 923 ("First is the fact that history, formally recognized as the primary measure of the seventh amendment right, simply does not provide any meaningful guidance in measuring the sufficiency of a case for jury determination. Second, and closely related, is the simple fact that it would be impossible to implement a uniform standard in a uniform manner, no matter how it were stated.").

[266]*See supra* note 185 (discussing the informal use of dictionaries to supply definitions).

[267]Both the language of the test in Rule 50 and the procedural stage make clear that, at summary judgment, one is examining (in some fashion) the evidence itself. *See supra* section B (exploring the relationship between facts, the focus in most discussions of the test for a motion to dismiss, and evidence).

arises whether this sufficiency is to be assessed in terms of probabilities or some other measure.

The former, probabilistic interpretation is supported by the Federal Rules of Evidence's very definition of relevant evidence in Rule 401(a) — which is whether the evidence "has any tendency to make a fact more or less probable than it would be without the evidence"[268] — as well as the absence of any sensible alternative.[269] This account suggests a similarity to the plausibility test for motions to dismiss that, as explained in section A, seems unavoidably about probabilities, with the explicit disclaimers in *Twombly* and *Iqbal* understood as rejecting the existence of a generic threshold rather than a context-specific one that is sensitive to consequences.

Matsushita further suggests a probabilistic orientation in its statement that, "[w]hen the moving party had carried its burden . . . , its opponent must do more than simply show that there is some metaphysical doubt as to the material facts."[270] Specifically, the Court indicates that, in doing so, it is necessary to combine the pieces of evidence in the particular case with background likelihoods in reaching an overall judgment:[271] "if the factual context renders [plaintiffs'] claim implausible[, plaintiffs] must come forward with more persuasive evidence to support their claim than would otherwise be necessary."[272] One might view this demand as advancing a context-specific probability requirement, but it might also be understood more narrowly as emphasizing that the probability that one logically infers from a set of evidence depends on the context (just as in standard Bayesian reasoning).

Although suggestive of a probability-based test, none of the foregoing shows how high the probability must be or whether and how any such threshold might vary with the context. The discussion in *Anderson* indicates that the standard is related to the decision criterion at trial:

> [T]he inquiry involved in a ruling on a motion for summary judgment or for a directed verdict necessarily implicates the substantive evidentiary standard of proof that would apply at the trial on the merits. . . . [I]n a run-of-the-mill civil case[, t]he judge's inquiry, therefore, unavoidably asks whether reasonable jurors could find by a preponderance of the evidence that the plaintiff is entitled to a verdict . . . .[273]

In concluding its discussion of the issue, the *Anderson* Court further stated:

> Thus, in ruling on a motion for summary judgment, the judge must view the evidence presented through the prism of the substantive evidentiary burden. . . . It makes no sense to say that a jury could reasonably find for either party without some benchmark as to what standards govern its deliberations and within what boundaries its ultimate decision must fall, and these standards and boundaries are

---

[268]FED. R. CIV. P. 401(a). The additional requirement in Rule 401(b) that the pertinent "fact is of consequence in determining the action" is close to Rule 56(a)'s requirement that the genuine dispute pertain to a "material" fact.

[269]Presumably, the rule does not require some count of the number of witnesses or documents, or any other measure independent of the extent to which the evidence bears on the matter in dispute.

[270]Matsushita Elec. Indus. Co. v. Zenith Radio Corp, 475 U.S. 574, 586 (1986).

[271]Compare the analysis in section B.

[272]475 U.S. at 587.

[273]Anderson v. Liberty Lobby, Inc., 477 U.S. 242, 252 (1986).

in fact provided by the applicable evidentiary standards.[274]

That the burden of proof at trial is understood to be probabilistic[275] and that the threshold at summary judgment varies positively with this probability reinforces the notion that the threshold is probabilistic, but all of this still does not indicate the height of that threshold.

We also know from *Anderson* that the trial court is not supposed to weigh the evidence in the manner that it would if it were the factfinder at the conclusion of a trial,[276] which seems inconsistent with any kind of probability assessment. Perhaps this sort of reassurance was intended primarily to assuage any sense that the opinion constituted a radical departure from existing understandings. In any event, it seems to contradict the instruction in *Anderson* that more than a scintilla of evidence is required to survive a summary judgment motion[277] and, as stated just above, the command that the judge must indeed decide whether there exists a

---

[274] *Id*. at 254–55.

[275] *See, e.g.*, MCCORMICK ON EVIDENCE, *supra* note 207, at 484 ("The most acceptable meaning to be given to the expression, proof by a preponderance, seems to be proof which leads the jury to find that the existence of the contested fact is more probable than its nonexistence." (citing MODEL CODE OF EVIDENCE R. 1(3))); David Kaye, *Naked Statistical Evidence*, 89 YALE L.J. 601, 603 (1980) (reviewing MICHAEL O. FINKELSTEIN, QUANTITATIVE METHODS IN LAW: STUDIES IN THE APPLICATION OF MATHEMATICAL PROBABILITY AND STATISTICS TO LEGAL PROBLEMS (1978)) ("A majority of courts and almost all commentators have concluded that [the preponderance of the evidence rule] is satisfied by evidence that indicates to the trier of fact that the event that must be established is more likely to have occurred than not.").

[276]

> Our holding that the clear-and-convincing standard of proof should be taken into account in ruling on summary judgment motions . . . by no means authorizes trial on affidavits. Credibility determinations, the weighing of the evidence, and the drawing of legitimate inferences from the facts are jury functions, not those of a judge, whether he is ruling on a motion for summary judgment or for a directed verdict. The evidence of the non-movant is to be believed, and all justifiable inferences are to be drawn in his favor.

Id. at 255. For pragmatic statements regarding actual practice to the contrary, see *Shager v. Upjohn Co.*, 913 F.2d 398, 403 (7th Cir. 1990) ("The growing difficulty that district judges face in scheduling civil trials, a difficulty that is due to docket pressures in general and to the pressure of the criminal docket in particular, makes appellate courts reluctant to reverse a grant of summary judgment merely because a rational factfinder *could* return a verdict for the nonmoving party, if such a verdict is highly unlikely as a practical matter because the plaintiff's case . . . is marginal."); POSNER, FEDERAL COURTS, *supra* note 32, at 179 ("Nowadays summary judgment is likely to be granted, and the grant upheld on appeal, if the district judge and the appellate panel are reasonably confident that the party opposing the motion has 'no case,' in the practical sense of being highly unlikely to win if the case is tried."); *id*. at 179 n.37 ("Since judges at best have only imperfect insight into the reactions of jurors, the criterion of 'plaintiff's likelihood of prevailing at trial' may in practice mean simply whether the judge thinks that the plaintiff's case has some merit."); WRIGHT, MILLER & KANE, *supra* note 18, at 217 ("[T]aken together, these three cases signal to the lower courts that summary judgment can be relied upon more so than in the past to weed out frivolous lawsuits and avoid wasteful trials, and the lower courts have responded accordingly."); Samuel Issacharoff & George Loewenstein, *Second Thoughts About Summary Judgment*, 100 YALE L.J. 73, 89 (1990) ("There is evidence in the post-trilogy case law that summary judgment has moved beyond its originally intended role as a guarantor of the existence of material issues to be resolved at trial and has been transformed into a mechanism to assess plaintiff's likelihood of prevailing at trial.").

[277] *See supra* note 265 (quoting *Anderson*); MOORE'S FEDERAL PRACTICE, *supra* note 22, at 92 ("[T]he opposing party's evidence must be sufficiently substantial to support a jury verdict in the nonmovant's favor. Evidence that is merely colorable, or is not significantly probative, is not enough. The mere existence of a scintilla of evidence in support of the nonmovant's position will not suffice.") (citing *Anderson*, 477 U.S. at 249–50). *Matsushita*, as quoted in the text just above, required in that context that plaintiffs had to present "more persuasive evidence to support their claim than would otherwise be necessary," which demands some degree of persuasive force but unhelpfully refers to the degree required as simply "more . . . than . . . otherwise," without saying either how much more or what is otherwise needed.

sufficient evidentiary basis to support a finding for the nonmoving party.

In all, the existing rules and the Supreme Court's interpretation thereof in its 1986 trilogy, on one hand, point to some sort of probability assessment, but, on the other hand, insist that this evaluation is to be performed without actual quantification and that its result is to be matched against a target that, although somehow positively related to the burden of proof, is unspecified.[278] Accordingly, it is difficult to say how this formulation relates to that developed in Part II[279] for optimal decisionmaking.[280] It is unclear what, if any, inference to draw from the fact that the Court's summary judgment decisions do not refer as explicitly and as often to the legal system's objectives[281] as do *Twombly* and *Iqbal* — keeping in mind that all the cases in principle should be interpreting the rules in question in accord with Rule 1's purposive command.[282] It is also noteworthy, however, that the Court's articulations of its test use the term "sufficient,"[283] variants of "reasonable," and in one instance a derivative of "plausible," all indicating a need for some sort of context-specific balancing judgment.

Commentators, as well as courts more broadly, have directed comparatively little attention to the competing factors that bear on when it makes sense to grant summary judgment. Most often mentioned, including in the trilogy, is the role of the jury. Note, however, that it is difficult to see this as the exclusive basis for determining the appropriate decision rule since the same rule is applied in cases in which there will be no jury.[284] In addition, it is unclear why

---

[278]*See, e.g.*, WRIGHT, MILLER & KANE, *supra* note 3, at 216 ("Another area of judicial disagreement is over the quantum of evidence that must be mustered in order to defeat a motion for summary judgment.").

[279]Because summary judgment is also available to plaintiffs, one must add the supplement in subsection III.C.1.

[280]As mentioned in note 87, it is not obvious that the summary judgment standard is tougher than that for a motion to dismiss — a difficulty that we can now see is compounded by the point that we cannot be sure that they each operate along the same dimensions. For example, *Twombly* and *Iqbal* indicate that the cost of continuation (in resources and disruption) is a relevant consideration, but its relevance under the trilogy is a matter of conjecture. Regarding the optimal relationship between the decision standards — and the question whether they can be meaningfully compared even if the same factors are relevant under both — see section II.D.

[281]There are some key references. *Matsushita* is undoubtedly concerned about the prospect of chilling procompetitive conduct, and *Anderson* with chilling speech.

[282]In examining the test under Rule 50 (before *Anderson*'s holding equating the test under Rule 56 to it), Cooper, *supra* note 18, at 960 states: "Decision between alternatives ordinarily involves a consideration of the expected desirability of each alternative and the expected losses resulting from a mistaken decision." He continues: "Decision must rest on an evaluation of the degree of uncertainty, the gains from correct decision, and the losses from mistaken decision." *Id*.

[283]Standard definitions of the term sufficient (just as was true of "plausible"; *see supra* section A) refer explicitly to purposive, means-ends reasoning.

[284]The combination of all of *Anderson*'s pronouncements in particular creates a quandary in light of the fact that the same summary judgment rule is applicable when the judge rather than a jury will be the factfinder. On one hand, we are told that the summary judgment test is the same as that for a judgment as a matter of law at trial, which for bench trials addresses, one might say, when a judge should feel compelled to substitute for his or her own judgment on the facts. On the other hand, these (equivalent) tests notably differ from how a judge would hypothetically decide if he or she did form a judgment on the evidence. More broadly, literature on summary judgment by both courts and commentators does not usually address bench trials and, in particular, does not explain why judges should be unwilling at the summary judgment stage to make judgments about their own judgments. (Note that the *Anderson* quotation in note 276, when applied to a bench trial, essentially states that the various tasks "are [judicial] functions, not those of a judge," even those of a judge considering a motion for a directed verdict after all the evidence has been presented.) For an exception, see *Shager v. Upjohn Co.*, 913 F.2d 398, 403 (7th Cir. 1990) ("A judge's decision to grant a motion for summary judgment

constraints on formulation of the decision rule involving protection of the role of the jury, if binding, receive so much less attention when analyzing the decision threshold for motions to dismiss: after all, if a judge is restricted from terminating a case on the eve of trial (or entering a judgment contrary to a jury's at the conclusion of a trial), it seems that this restraint may be eroded to the extent that the judge can terminate the case at an earlier stage.[285]

Setting such institutional limitations to the side (whatever their appropriate force might be),[286] let us apply the general framework in Part II, which encompasses decisions at any stage in multistage legal proceedings. Although the main considerations and subcomponents are precisely the same as those for motions to dismiss, their magnitudes in specified scenarios and their relative importance on average may differ greatly across the two contexts. For the latter, two points are clear.

---

may be a good predictor of the outcome of a bench trial before the same judge; it may not be a good predictor of the outcome before a jury.").

[285] *See, e.g.*, Thomas, *supra* note 176.

[286] Established practice in U.S. civil litigation does not appear to take the Seventh Amendment right to trial by jury as a tight constraint. In addition to its being inapplicable in many areas of law (understood as not being "suits at common law" and thus not under the Amendment), provisions for summary judgment and judgment as a matter of law (and for motions to dismiss) seem to be regarded as creating little tension. For example, the Advisory Committee's note on Rule 50 (accompanying the 1991 amendments) states:

> The expressed standard makes clear that action taken under the rule is a performance of the court's duty to assure enforcement of the controlling law and is not an intrusion on any responsibility for factual determinations conferred on the jury by the Seventh Amendment or any other provision of federal law.

FED. R. CIV. P. 50, advisory committee's note; *see* WRIGHT, MILLER & KANE, *supra* note 18, § 2714; *but see* POSNER, FEDERAL COURTS, *supra* note 32, at 182 ("The redefinitions of summary judgment and dismissal on the pleadings that I have been discussing are questionable in formal legal terms. They have the practical effect of amending the Federal Rules of Civil Procedure without the required formalities of amendment. And they step on the skirts of the Seventh Amendment . . . ."); *see also* FED. R. CIV. P. 56, advisory committee's note (making no mention of juries in discussing summary judgment). Given the question-begging character of Rule 50's standard and that evolving interpretations make little reference to the Seventh Amendment, however, one wonders whether this rationalization is much more than an attempt to change the subject. Cooper's article, *supra* note 18, which prominently discusses the Seventh Amendment at the outset, makes little mention of it in later assessments of how Rule 50 ought to be applied. For example, the special role of the jury does not seem to restrict judges' substitution of their own determinations of how the evidence is best assessed. *See id*. at 904 ("[G]eneral statements of deference to the jury cloak widely different degrees of deference according to the perceived consequences of possible jury error."); *id*. at 932 ("Is a jury, then, likely to make fewer mistakes if it is allowed free rein to reject uncontradicted testimony which is not at odds with any known facts, and which is given by an uninterested witness who has not been impeached in any way, than if it is required to accept such testimony? Several considerations suggest that more, not fewer, mistakes would result [from allowing the jury to decide such cases]."); *see also id*. at 968 ("Judges may nevertheless do well to continue to bury these considerations under a blanket of just such generalities. Reticence may be justified in part because litigation continues to serve a witch-doctor function, and it would lose much of its perhaps dwindling acceptability if it were frankly confessed that ordinary courts cannot really know what happened once upon a time and are prepared to act in states of ignorance which vary according to individual circumstances."). Additionally, Cooper advocates that, when entering a judgment as a matter of law rather than deferring to juries, it may be appropriate for the judge to advance certain substantive views as to proper outcomes but to write opinions that mask the actual basis for decision so as to preserve appearances. *See id*. at 970–71 (stating that "More leeway may be given if the plaintiff is badly injured and in great need, if the defendant enjoys a capacity for spreading losses which the law does not take into account, and so on." — but cautioning about mentioning such motivations); *id*. at 968 ("Some of the examples offered depend on considerations which are not formally incorporated in the relevant legal rules. The most poignant way of stating the resulting dilemma is to observe that some of the reasons for altering the permissible scope of inference depend upon matters which are formally ruled irrelevant.").

First, continuation costs will differ. For denial of a motion to dismiss, these consist of the various costs of discovery and also, with a probability (depending on the likelihood of summary judgment being denied), the costs of trial. For denial of a motion for summary judgment, the former costs are sunk, and the latter are certain (aside from the possibility of settlement, discussed in subsection III.D.2). As already mentioned in section II.D, we can readily imagine that these factors imply lower expected costs of continuation, which favors a more generous standard later than earlier.

Second, the information that may be acquired if a case is continued is qualitatively different. At the stage of a motion to dismiss, moving forward will, at least through the discovery phase, allow parties to uncover additional evidence. By contrast, proceeding from the summary judgment decision point involves trial. There, the primary additional enlightenment will be due to the processing of evidence: the factfinder will be exposed to a somewhat organized, live presentation of what each party believes to be the evidence that is most favorable to its cause. Perhaps as a rough cut, at the summary judgment stage (compared to decisions on motions to dismiss), both the costs and the informational benefits of continuation will be lower, with unclear implications for the diagnosticity/cost ratio that was offered as a heuristic in subsection II.A.3.

Returning to the case of a bench trial, note that a range of intermediate outcomes is conceivable at the summary judgment stage if the judge does not feel constrained to abstain from deeper examination of the evidence. For example, if the record is largely documentary,[287] spending additional time before trial to examine the documents — with the aid of the parties, through briefs and oral argument — could in the limit approach what might be done at trial.[288] In this imagined setting, a judge might be continuously deciding how much more effort to expend before being able to make a decision for one party or the other, asking periodically whether further study is likely to offer sufficient incremental understanding to justify the additional work. This depiction, note, is suggestive of the more sequenced manner of legal proceedings in some Continental systems, particularly those that place greater weight on the documentary record.[289]

V. CONCLUSION

This Article develops a conceptual framework for analyzing how decisions are optimally made at each juncture in multistage legal proceedings — be it a distinct formal step in civil or

---

[287]Witnesses may have already been deposed, and it is controversial the extent to which hearing live testimony enhances decisionmaking quality. *See, e.g.*, Olin Guy Wellborn III, *Demeanor*, 76 CORNELL L. REV. 1075 (1991) (reporting that experimental evidence consistently indicates that demeanor does not aid ordinary people in assessing credibility and in fact may somewhat diminish the accuracy of such judgments).

[288]Interestingly, former Rule 56(c) (then entitled "Motion and Proceedings Thereon") referred to a "hearing," which might have been understood to allow the hearing of witnesses and not merely oral argument by the lawyers. However, recent amendments have unceremoniously eliminated this reference, without mention in the Advisory Committee Notes. *See* FED. R. CIV. P. 56, advisory committee's note (omitting mention of the previously authorized "hearing" in discussing amendments to the rule).

[289]A notable distinction is when a panel of judges, not the examining judge, is the decisionmaker, producing a parallel with the distinction between jury and judge.

criminal adjudication; an informal, ongoing assessment during investigations by police, prosecutors, or regulatory agencies; or in the course of dispute resolution conducted in Continental legal systems or by arbitrators. Although the building blocks are elementary, the decision criteria that reflect their interaction are informationally demanding, surprisingly complex, and occasionally counterintuitive.

At nonfinal stages, such as motions to dismiss and for summary judgment in U.S. civil litigation, the decision to continue a case in some particular scenario rather than to terminate it has three central consequences: by raising the expected costs of prospective harmful acts, deterrence is enhanced; by increasing costs of benign acts, chilling is intensified; and by the very act of proceeding to the next stage, adjudication costs are incurred. When the deterrence gain exceeds the sum of chilling costs and continuation costs, continuation is optimal. To the extent feasible, legal systems need to heed this lesson if they are to impose liability for harmful acts as often as possible so as to discourage their commission, penalize benign conduct only infrequently so as to avoid significantly chilling such behavior, and accomplish these objectives without undue effort and expense.

Determinants of each component are explored and implications drawn. Some conclusions are straightforward: all else equal, continuation is favored when acts subject to liability are more harmful, when fewer benign acts might be mistaken for harmful ones, and when the available information is more often generated by harmful than by benign acts. Others are more subtle and unexpected. Although higher adjudication costs make continuation more expensive, they also make deterrence more valuable and chilling less detrimental because both behavioral effects reduce the number of underlying acts that might enter the legal system and thereby generate these heightened costs. For this and other reasons, higher average system costs have ambiguous implications for the optimality of continuation. In areas where key information ordinarily is solely in defendants' hands, continuation may be attractive, but when this is so only in the scenario at hand and not more generally, termination tends to be favored. And despite the conventional wisdom that decision rules should be more demanding at later stages of adjudication, some factors complicate this relationship (notably, greater availability of information does not necessarily favor tougher criteria) and others contradict it (at later stages, more costs are sunk, so the cost of continuation is falling).

The Article also considers a number of variations and extensions. In system design, stages are constructed rather than predetermined, so it is important to assess when adjacent stages should be separated or combined, what sequence they should follow, and how they should be composed. Systemic objectives are also affected by changes in substantive law, the allowance of early findings of liability as well as early terminations, and the appropriate calibration of enforcement effort and sanctions. Because each feature influences what decision rules are optimal at every stage, and vice versa, interactions are explored. In opposition to a commonly expressed view, fine-tuning in different areas of substantive law is often best accomplished by adjusting procedure — specifically, decision thresholds at various stages of adjudication — rather than substantive rules. Also, the analysis challenges the coherence of the procedure-substance distinction that is implicit, for example, in the advocacy of transsubstantive legal procedure. In addition, concerns for system error, including that involving the mistaken imposition of sanctions on those who commit benign acts, may well favor setting higher

sanctions, in combination with more stringent rules for case continuation and ultimate assignment of liability. Private parties' and government enforcers' incentives to pursue cases, and to settle them, are also influenced by the decision rules under examination, adding another dimension to the analysis.

Throughout, the exploration is theoretical, preliminary (in light of the substantial absence of previous inquiry), dependent on empirical knowledge that is largely absent, and contingent on features of particular legal systems that vary tremendously. Even so, it is helpful to apply the framework in an attempt to illuminate the operation of existing legal systems and to identify potential reforms.

In this regard, the Article focuses on motions to dismiss and for summary judgment in U.S. federal civil litigation. To an even greater degree than seems to be recognized, current legal standards, as stated in the Federal Rules of Civil Procedure and elaborated by the Supreme Court, are unclear, question-begging in key respects, and at bottom open-ended. The Federal Rules articulate a purposive approach to their construction and administration, and *Twombly* and *Iqbal* refer frequently to consequences and endorse a context-specific analysis that draws on experience and common sense. Accordingly, it is natural to contemplate using this Article's decision criteria to inform the interpretation and application of existing rules. In addition to considering direct implications, the Article addresses a number of allied topics: the nature of facts (the relationship between facts and evidence and the bases for inferences in particular legal settings), informational challenges in applying sensible decision criteria (their character, magnitude, and possible institutional responses), and judicial discretion (why it is so great in this context and factors that influence its exercise).

This Article begins to answer the fundamental question of how to design decision rules for multistage legal proceedings in a manner that best advances the legal system's underlying objectives. It breaks new ground not only in the particular answers it develops but in its very formulation of the problem. Scholars have not performed a ground-up analysis, while courts, practitioners, and legal commentators have not appreciated the extent to which both longstanding and newly minted legal tests substantially obscure extant shortcomings and fail to provide coherent, workable guidance. It is unwelcome news that so much needs to be done and that the task is so daunting. But such is how it is. The time has come to continue the inquiry rather than allow it to remain terminated, or at least truncated, at so early a stage.