Article

Contract Theory and the Limits of Contract Law

Alan Schwartz[†] **and Robert E. Scott**[††]

CONTENTS

† Sterling Professor of Law, Yale Law School; Professor, Yale School of Management.
†† David and Mary Harrison Distinguished Professor of Law, University of Virginia School of Law; Justin W. D'Atri Visiting Professor of Law, Business and Society, Columbia Law School. This Article benefited from comments received at workshops at the Law Faculty, Cambridge, England, at the 2003 American Law and Economics Association meetings, and at the Harvard, Pennsylvania, Texas, Toronto, Virginia, and Yale Law Schools. We also are grateful to Bruce Ackerman, Scott Baker, Richard Brooks, Jules Coleman, Victor Goldberg, Sam Issacharoff, John Jeffries, Jason Johnston, Paul Mahoney, Tom Nachbar, Eric Posner, Andrew Postlewaite, Henry Smith, Paul Stephan, William Stuntz, and George Triantis for helpful comments.

I. INTRODUCTION

Contract law has neither a complete descriptive theory, explaining what the law is, nor a complete normative theory, explaining what the law should be. These gaps are unsurprising given the traditional definition of contract as embracing all promises that the law will enforce. Even a theory of contract law that focuses only on the enforcement of bargains must still consider the entire continuum from standard form contracts between firms and consumers to commercial contracts among businesses. No descriptive theory has yet explained a law of contract that comprehends such a broad domain. Normative theories that are grounded in a single norm—such as autonomy or efficiency—also have foundered over the heterogeneity of contractual contexts to which the theory is to apply.[1] Pluralist theories attempt to respond to the difficulty that unitary normative theories pose by urging courts to pursue efficiency, fairness, good faith, and the protection of individual autonomy. Such theories need, but so far lack, a meta-principle that tells which of these goals should be decisive when they conflict.[2]

1. For a broad discussion of this problem, see MICHAEL J. TREBILCOCK, THE LIMITS OF FREEDOM OF CONTRACT 241-68 (1993). Autonomy theories require elastic notions of consent in order to regulate the full scope of contracting behavior with one norm. *See, e.g.*, Randy E. Barnett, *A Consent Theory of Contract*, 86 COLUM. L. REV. 269 (1986); Randy E. Barnett, *The Sound of Silence: Default Rules and Contractual Consent*, 78 VA. L. REV. 821 (1992); Peter Benson, *Abstract Right and the Possibility of a Nondistributive Conception of Contract: Hegel and Contemporary Contract Theory*, 10 CARDOZO L. REV. 1077 (1989); Peter Benson, *Contract*, *in* A COMPANION TO PHILOSOPHY OF LAW AND LEGAL THEORY 24, 33-43 (Dennis Patterson ed., 1996); Peter Benson, *The Idea of a Public Basis of Justification for Contract*, 33 OSGOODE HALL L.J. 273 (1995). Efficiency theories tend to have a more limited scope. Positive articles analyze broad doctrinal patterns in an attempt to find fundamental consistency between these patterns and the efficiency norm, but the authors do not purport to provide a fully descriptive theory of contract law. *See, e.g.*, Ian Ayres & Robert Gertner, *Filling Gaps in Incomplete Contracts: An Economic Theory of Default Rules*, 99 YALE L.J. 87 (1989); Charles J. Goetz & Robert E. Scott, *Enforcing Promises: An Examination of the Basis of Contract*, 89 YALE L.J. 1261 (1980). Normative economic theories, on the other hand, typically evaluate discrete doctrines by the efficiency norm. *See, e.g.*, Charles J. Goetz & Robert E. Scott, *Liquidated Damages, Penalties and the Just Compensation Principle: Some Notes on an Enforcement Model and a Theory of Efficient Breach*, 77 COLUM. L. REV. 554 (1977) [hereinafter Goetz & Scott, *Liquidated Damages*]; Christine Jolls, *Contracts as Bilateral Commitments: A New Perspective on Contract Modification*, 26 J. LEGAL STUD. 203 (1997); Alan Schwartz, *The Case for Specific Performance*, 89 YALE L.J. 271 (1979); Robert E. Scott, *The Case for Market Damages: Revisiting the Lost Profits Puzzle*, 57 U. CHI. L. REV. 1155 (1990).

2. The problems that pluralist theories without meta-norms pose are nicely illustrated in Melvin Eisenberg's effort, which purports to solve the broad-scope-of-contract problem by proposing overlapping sets of norms. *See* Melvin Aron Eisenberg, *The Bargain Principle and Its Limits*, 95 HARV. L. REV. 741 (1982); Melvin A. Eisenberg, *The Theory of Contracts*, *in* THE THEORY OF CONTRACT LAW: NEW ESSAYS 206 (Peter Benson ed., 2001) [hereinafter Eisenberg, *Theory of Contracts*]. For example, Eisenberg's schema restricts the domain of freedom of contract by norms of reciprocity, trust, and fairness. He recognizes that this multivalue approach can generate conflicting social propositions. When conflicts actually occur, "the lawmaker must make a legal rule that gives a proper weight and role to each of the conflicting values or goals in the context at hand." Eisenberg, *Theory of Contracts*, *supra*, at 244. Further, "when social propositions conflict the Legislator must exercise good judgment concerning the weight and role

We attempt to make progress here with a more modest approach—to set out and defend a normative theory to guide decisionmakers in the regulation of business contracts.[3]

The theory's affirmative claim, in brief, is that contract law should facilitate the efforts of contracting parties to maximize the joint gains (the "contractual surplus") from transactions. The theory's negative claim is that contract law should do nothing else. Both claims follow from the premise that the state should choose the rules that regulate commercial transactions according to the criterion of welfare maximization.

A simple categorization of the universe of bargaining transactions will clarify the domain of our theory. A transaction involves a seller (whether of goods or services) and a buyer. Parties to transactions can be partitioned into individuals and firms. This yields four transactional categories: (1) A firm sells to another firm, (2) an individual sells to another individual, (3) a firm sells to an individual, and (4) an individual sells to a firm. Category 2 contracts, between individuals, are primarily regulated by family law (antenuptial agreements and divorce settlements) and real property law (home sales and some leases). Few litigated contracts between individuals are regulated by the rules of contract law. Category 3 contracts, between a firm as seller and an individual as buyer, are primarily regulated by consumer protection law, real property law (most leases), and the securities laws. Category 4 contracts, between an individual as seller and a firm as buyer, commonly involve the sale of a person's labor, and are regulated by laws governing the employment relation. That leaves Category 1 contracts (those between firms) as the main subject of what is commonly called contract law—namely, the rules in Article 2 of the Uniform Commercial Code (UCC) and the provisions of the *Restatement (Second) of Contracts*. Such provisions are primarily invoked to resolve disputes arising under Category 1 contracts. Our theory applies only to these contracts, and thus has important implications for the content of the UCC and the common law of contracts.

to be given to each proposition in the issue at hand." *Id*. Eisenberg recognizes that his theory lacks a metric that would tell the lawmaker just how to give the proper "weight and role" to each social proposition or value when conflicts occur. *Id*. Since courts or legislatures are likely to be involved when the relevant social propositions or values arguably favor more than one type of litigant or interest group, pluralist theories such as Eisenberg's tend to be least helpful when they are most needed.

3. In a thoughtful critique of autonomy and efficiency theories of contract, Michael Trebilcock concludes that both theory types are "legitimate in their own terms," but that without a "meta-theory that weighs or ranks these various values," both should be pursued in various social contexts according to the relative competence of different legal institutions to perform effectively. TREBILCOCK, *supra* note 1, at 248. This Article takes up Trebilcock's invitation and proposes a normative theory that fits business contracts, the subsidiary category of contractual relationships that the law most affects.

Category 1 contracts, however, can be partitioned into two subcategories. Some parties obviously are sophisticated economic actors (e.g., the General Electric Company). Other parties function in commercial contexts but have many of the characteristics of ordinary persons (e.g., a gift shop owned and run by a retired teacher). Any effort to analyze contracts between "firms" thus confronts a boundary issue—how to define a firm for purposes of the analysis. We draw this boundary here by defining a Category 1 firm as (1) an entity that is organized in the corporate form and that has five or more employees, (2) a limited partnership, or (3) a professional partnership such as a law or accounting firm. These economic entities can be expected to understand how to make business contracts, and the theory we develop applies only to contracts between two such firms. We do not address the extent to which our conclusions hold when one or both of the parties to a commercial contract fall on the other side of our boundary.

Firms that maximize profits face the canonical "contracting problem" of ensuring both efficient ex post trade and efficient ex ante investment in the subject matter of the contract.[4] Parties trade efficiently when, and only when, the value of the exchanged performance to the buyer exceeds the cost of performance to the seller. Parties invest efficiently when their actions maximize a deal's expected surplus. Many observers would agree that contract law should attempt to facilitate efficient trade and investment. The novelty of our theory lies in its systematic development of the implications of this goal and in its claim that contract law should restrict itself to the pursuit of efficiency alone (for Category 1 contracts).

Four objections may be made to the claim that contract law should restrict itself to encouraging efficient trade and investment. First, one could argue that firms sometimes do not maximize profits and, owing to the systematic cognitive errors made by the people who run them, are incapable of doing so should they try. A law that presupposes profit maximization would then be misguided. Second, one might claim that firms that maximize profits sometimes do bad things—pollute the environment, for example—that the law should attempt to deter. Third, one could contend that the state should promote fairness in contracting in addition to efficiency. And, finally, one might maintain that the state should pursue distributional goals, even if they may sometimes conflict with efficiency.

These objections would be troublesome for an efficiency approach that covered all contract types. We will argue, however, that they have little force when Category 1 contracts *alone* are considered. Firms and markets are structured so as to minimize the likelihood of systematic cognitive error

4. Legal scholars commonly refer to investment in the contract's subject matter as "reliance." We use reliance and the economist's term "investment" interchangeably.

by important decisionmakers within the firm. Cognitive error, then, is more likely to afflict Category 2, 3, and 4 contracts than Category 1 contracts. Further, the bad things that firms do commonly entail imposing costs on third parties, such as creating environmental harms or erecting barriers to entry. These behaviors—the creation of negative externalities—are regulated by the environmental and antitrust laws. An analysis of contract law as such therefore can assume the absence of externalities. Finally, it usually is futile to pursue either distributional goals or contractual fairness when firms are permitted a large measure of contractual freedom. This is because firms will contract away from redistributive or fair legal rules that do not maximize joint surplus. In sum, efficiency is the only institutionally feasible and normatively attractive goal for a contract law that regulates deals between firms.[5]

An efficiency theory restricted to contracts between firms (as firms are defined above) has four major implications for contract law. The first implication follows from an important fact: Many contracts would be performed even if there were no legal sanction for breach. Contracts will be "self-enforcing" when parties contemplate making a series of contracts and the gains from breach are lower than the expected profit stream from future contracts that breach would cause to vanish. Moreover, neither party would breach if the gains from breach were less than the reputational sanction the market would exact. When contracts fall outside the self-enforcing range, however, legal enforcement is necessary to ensure performance in two principal cases: in volatile markets, when a party's failure to perform could threaten its contract partner's survival; and when contractual surplus would be maximized if one or both parties made relation-specific investments.[6] "Enforcement" includes more than simply requiring parties to perform, however. It also entails the prevention of fraud and duress, as well as rules to encourage or facilitate performance, such as damages rules. Perhaps a third of the sections in UCC Article 2 are enforcement rules under the definition here. The initial implication of our theory is that enforcement, when needed, is by far the most important thing the state does. Put

5. As another example of the criticism that we sidestep here, Professor Eisenberg has criticized theories holding that contract law should maximize welfare alone on the ground that these theories are "impoverished . . . because they exclude other important policy values, such as the value of keeping intimate and other affective relationships free from the intrusion of state power." Eisenberg, *Theory of Contracts*, *supra* note 2, at 238. This objection may have force as applied to Category 2 contracts, between persons, but seems irrelevant to the Category 1 contracts we analyze. Contracts between General Electric and General Motors do not involve "intimate" or "affective" relationships.

6. A relation-specific investment is not fully "redeployable." As an example, assume that a seller purchases standard steel tubes to make a machine for the buyer. The seller's investment would be "general" if breach occurred before the seller began work on the tubes because the tubes could be resold on the market. The investment would become "relation-specific" if breach occurred after the tubes had been fabricated into shapes that only the buyer could use, for then the transmuted tubes could only be resold as scrap, probably for less than their cost.

more starkly, a modern commercial economy can function well with little more than honest courts and a set of enforcement rules. The rest is of second-order importance.

A court cannot enforce contracts, however, without a theory of interpretation that "maps" from the semantic content of the parties' writing to the writing's legal implications. Our second implication thus holds, in contrast to the UCC and much modern scholarship, that textualist interpretation should be the default theory for Category 1 contracts. Business firms, that is, commonly prefer courts to adhere as closely as possible to the ordinary meanings of words, to apply a "hard" parol evidence rule,[7] and to honor "merger clauses" (which state that the parties intended their writing to be interpreted as if it were complete). A textualist theory of interpretation, however, will not suit all parties all of the time. Therefore, courts should use narrow evidentiary bases when interpreting agreements between firms, but also should comply with party requests to broaden the base that is applicable to them. This implication is at variance with current law, which holds that interpretation is an issue for courts to decide and should be conducted according to rules that parties cannot vary.

Contract law has more rules regulating various aspects of the contracting relationship than are needed solely to perform its enforcement and interpretation functions. Typically, these rules are defaults, controlling only when parties do not contract out of them. Creating good defaults is widely believed to be the principal function of a law of contracts. This belief is misguided. Our theory's third implication holds that the *effective domain* of business-contract law is much smaller than is commonly thought. The state can create defaults that business firms would want only under very stringent conditions. To be useful, a default *rule*[8] must apply in very few possible states of the world, be relatively simple in form, be efficient in a highly heterogeneous set of circumstances, and not rely on information that courts cannot conveniently recover. A default *standard* should be written when parties do not need, or it is too costly to provide them with, concrete guidance regarding the performance obligation. Because standards permit parties much latitude (e.g., the seller must deliver in a "reasonable" time), a good standard will confer discretion only when a party's likely actions under it will maximize joint—rather than individual—gains. Statutory drafters and courts, we argue, often adopt default rules and standards that fail to satisfy these stringent conditions. This is itself inefficient because parties respond to bad rules or standards by

7. A "hard" parol evidence rule treats writings that appear to be complete contracts *as* complete contracts. *See infra* text accompanying notes 96-98.

8. The decisionmaker specifies the content of a rule in advance. For example, drivers cannot exceed a speed limit of fifty-five miles per hour. In contrast, the decisionmaker specifies the content of a standard ex post. Thus, parties must drive "reasonably" under the circumstances.

contracting out of them. The creation of inefficient defaults thus raises business parties' contracting costs but does not otherwise affect their behavior. The lack of good defaults thus makes much of today's contract law irrelevant to commercial life.

In addition to its many defaults, contract law contains a number of mandatory rules that are applied to contracts between firms as well as to contracts between firms and persons. The fourth implication of our efficiency theory is that many of the rules regulating business contracts should not be mandatory. We discuss a number of mandatory rules, including interpretation rules, modification rules, and rules relating to liquidated damages clauses. The only justification for these rules is a form of paternalism: These rules do not override contractual terms because those terms create externalities or are unconscionable, but rather they seek to override terms that appear to conflict with parties' true substantive intentions. We argue, however, that business firms would have good reasons to adopt the terms that today are prohibited. A commitment to party sovereignty requires that those reasons be respected.

The need for a general efficiency theory of business contracts is particularly salient today. The creation of a contract law has become an important priority in many countries that have made a new (or renewed) commitment to markets, for there is a consensus that a good contract law is a necessary condition for a modern commercial economy. It is less well understood just how such a law is supposed to function. Our Article thus addresses concerns that have global implications.

A further reason motivating our decision to develop a theory of business contracts is that the building blocks for such a theory are only now becoming available. Contract theory has become one of the most significant fields in modern microeconomics and industrial organization economics. Three recent Nobel Prizes were awarded to George Akerloff, Michael Spence, and Joseph Stiglitz, largely for work in contract theory, even though the field is less than thirty years old.[9] Moreover, much of the work in the field takes a mathematical form, and thus has not been easily accessible to nonspecialists. We draw heavily on contract theory to construct our normative theory of contracts.[10]

9. The work of these scholars is concisely summarized in Karl-Gustaf Löfgren et al., *Markets with Asymmetric Information: The Contributions of George Akerlof, Michael Spence and Joseph Stiglitz*, 104 SCANDINAVIAN J. ECON. 195 (2002).

10. Law-and-economics scholars such as Aaron Edlin, Ian Ayres, and Jason Johnston have used contract theory in an illuminating fashion when discussing particular legal rules. *See, e.g.*, Ayres & Gertner, *supra* note 1; Aaron S. Edlin, *Cadillac Contracts and Up-Front Payments: Efficient Investment Under Expectation Damages*, 12 J.L. ECON. & ORG. 98 (1996); Jason Scott Johnston, *Strategic Bargaining and the Economic Theory of Contract Default Rules*, 100 YALE L.J. 615 (1990). The genre of model that we and these scholars use has performed well in empirical tests. *See* P.A. CHIAPPORI & B. SALANIE, TESTING CONTRACT THEORY: A SURVEY OF

Finally, as we suggested earlier, the current state of contract law scholarship suffers from the absence of a successful theory of contract. Thirty years ago, Grant Gilmore described what he called the classic Willistonian model. This model was grounded in formalist notions of the centrality of written agreements voluntarily exchanged between contracting parties, and it emphasized the limited role of the law in enforcing and interpreting these agreements.[11] According to Gilmore, this classical model owed more to Holmes's imagination than to a careful reading of the case law.[12] But whether this was so or not, Gilmore believed that modern case law repudiated the model. The disjunction between the dominant scholarly view and the lived doctrine, he thought, produced incoherence.[13] Modern scholars commonly share Gilmore's rejection of Williston but have yet to disprove his incoherence thesis.[14] We, too, lack a theory of everything. Rather, the theory we develop here is Willistonian in spirit, but applies in a limited domain—to contracts between firms that do not create externalities.

This limited scope permits our normative thesis to develop according to a particular logic. The market's social function is to maximize welfare, subject to distributional and fairness constraints. Firms, we show below, have incentives to choose the contracts and contracting strategies that will maximize the surplus from their deals. Further, firms are better able than courts or statutory drafters to choose efficient terms and strategies. It follows that, when externalities are absent, a contract law that regulates firms should be the contract law that firms would prefer generally to apply to their transactions. That is, the state should let the preferences of firms control because firms can better pursue the objective that both the state and firms share. Thus, the central organizing question of this Article is: What contract law would commercial parties want the state to provide?

We proceed as follows: Part II defends the welfare-maximization norm as applied to the contracts of sophisticated actors. In Part III, we describe commercial parties' first-order preference to have the state enforce contracts in order to protect relation-specific investments and to guard against especially disruptive market movements. Part IV argues that firms

SOME RECENT WORK 27 (Institut National de la Statistique et des Études Économiques, Working Paper No. 2002-11, 2002), http://www.crest.fr/doctravail/document/2002-11.pdf.

11. *See* GRANT GILMORE, THE DEATH OF CONTRACT 42-53 (1974).

12. *See id*. at 63 ("[T]he theory of contract, as formulated by Holmes and Williston, seems to have gone into its protracted period of breakdown almost from the moment of its birth."). *But see* Richard E. Speidel, *An Essay on the Reported Death and Continued Vitality of Contract*, 27 STAN. L. REV. 1161, 1182-83 (1975) (reviewing GILMORE, *supra* note 11).

13. *See* GILMORE, *supra* note 11, at 67-68 ("[Consider] the *Restatement*'s definition of consideration (§ 75) taken in connection with its most celebrated section—§ 90 [promissory estoppel]. . . . The one thing that is clear is that these two contradictory propositions cannot live comfortably together: in the end one must swallow the other up.").

14. *See, e.g*., Eric A. Posner, *Economic Analysis of Contract Law After Three Decades: Success or Failure?*, 112 YALE L.J. 829 (2003) (lamenting the absence of successful normative or descriptive theories of contract law).

want the state to supply a theory of interpretation, but not the theory currently advanced by the UCC and the *Restatement (Second) of Contracts*. Rather, we defend a textualist theory of interpretation as the optimal default approach for business contracts. In Part V, we develop the restrictive conditions under which the state can create default terms that satisfy typical party preferences. Part VI analyzes a set of unjustifiable mandatory rules—rules that rest on a misplaced view of the parties' interests. We conclude, in Part VII, that today's contract law is a series of category mistakes. Rules that are appropriate for contracts involving individuals (Categories 2 through 4 above) are too frequently applied to sophisticated parties. Commercial law for centuries has drawn a distinction between mercantile contracts and others.[15] Modern scholars have not systematically pursued the normative implications of this ancient distinction, however. We attempt to cure this neglect by setting out the theoretical foundations of a law merchant for our time.

II. JUSTIFYING AN EFFICIENCY THEORY OF CONTRACT

A. *What Firms Maximize*

It has been traditionally assumed that firms attempt to maximize expected profits.[16] Recently, the accuracy of this assumption has been challenged. There are two reasons why an economic actor may fail to maximize wealth: (1) She is maximizing something other than her own wealth, perhaps because she is concerned with fairness; or (2) she cannot maximize wealth in the context under study, perhaps because she is prone to cognitive error. These reasons apply much less to firms than to persons. A firm is directed by its owners, who often are shareholders. Shareholders prefer their firms to maximize profits, which the shareholders then can

15. Vestiges of this distinction exist in the few UCC sections that regulate deals "between merchants" differently from deals between a merchant and a person. *See, e.g.*, U.C.C. §§ 2-104(3), 2-201(2), 2-207(2), 2-209(2), 2-603, 2-609(2) (2003).

16. Individuals are assumed to be risk-averse while firms are assumed to be risk-neutral. The utility function of a risk-neutral party is linear in money—that is, the party values each additional dollar of wealth it may receive as much as it valued all previous dollars. Because monetary gains are coextensive with utility gains for risk-neutral parties, risk-neutral firms will maximize profits, a monetary measure. The risk-aversion or profit-maximization assumption for firms follows from two facts. First, the amount at stake in any one contract commonly is small in relation to the size of the firm, so firms actually hold contract portfolios. Individual risks tend to offset one another in a portfolio, so the portfolio holder—i.e., the firm—wants only to maximize the monetary value of the portfolio as a whole, which is best done by maximizing the value of each contract in it. Second, as is considered immediately in the text above, firms are owned by shareholders who themselves hold diversified portfolios. The value of a shareholder's portfolio is maximized when each firm in it does as well financially as it can do. The first ground for supposing that firms are risk-averse sometimes does not hold. We pursue the implications of this failure in Section III.C below.

consume or save. Firms thus will choose to maximize profits unless the managers who run them cannot be controlled by the shareholders who own them. In these cases, the managers may be maximizing their own earnings or perks at the expense of profit maximization.

No one doubts that managers sometimes successfully sabotage owners. For two reasons, however, we will assume that managers obey shareholder instructions. First, managers sabotage shareholders either by diverting corporate wealth to themselves or by failing to take appropriate risks on behalf of the firm. Managers, however, have no incentive to degrade the quality of the contracts that they write; after all, these contracts create the wealth that the managers later can divert. Second, the legal rules that attempt to deter bad manager behavior fall in the domains of the criminal, corporate, and securities laws. Contract law should exploit this specialization by assuming that the agreements it regulates reflect the parties' maximizing choices.

Firms that attempt to maximize profits can be expected to do as well as their circumstances permit. This is because the pressure to survive promotes competence. This pressure takes two forms. First, firms that systematically make bad economic decisions lose out in competition with profit-maximizing firms. Hence, surviving firms are generally the ones that can do what they set out to do. Second, employees who systematically make bad economic decisions are unlikely to be promoted to positions of responsibility. Hence, senior managers can generally do what they set out to do.[17] This is not to say that all firms all the time pursue profit-maximizing strategies. But it is to say that owners and the market put systematic pressure on firms to behave optimally; hence, it is a plausible working assumption that firms rationally pursue the objective of maximizing profits.[18]

17. In addition, many corporate executives have attended business school and also attend business school executive programs for working managers. It is a function of business education to teach people to make optimizing (rather than cognitively erroneous) decisions. Studies also show that individual subjects can be trained to perform complex game-theoretic reasoning. *See* Miguel Costa-Gomes et al., *Cognition and Behavior in Normal-Form Games: An Experimental Study*, 69 ECONOMETRICA 1193 (2001); Eric J. Johnson et al., *Detecting Failures of Backward Induction: Monitoring Information Search in Sequential Bargaining*, 104 J. ECON. THEORY 16 (2002).

18. Psychologists and economists have shown that persons make systematic cognitive mistakes in laboratory experiments when asked to solve specified individual decision problems. These experiments do not test a general theory of how people make decisions, and thus they raise an issue of external validity: When will real-world parties behave as did the experimental subjects? Two scholars recently noted a consequence of this "lack of theoretical foundations": "[T]he policy influence of BE [behavioral economics] is limited by its inability to predict circumstances in which anomalous behavior will arise (other than in those sorts of circumstances in which it has been observed before) or how it will respond to policy changes." Jessica L. Cohen & William T. Dickens, *A Foundation for Behavioral Economics*, 92 AM. ECON. REV. 335, 335 (2002). For a recent skeptical view of the relevance of the psychological literature to the law, see

The assumption that each party to a contract wants to maximize its own profit does not itself imply that parties also want to maximize joint gains. Rather, a party may prefer a larger share of a smaller pie. Thus, one might think that parties will behave strategically at the expense of joint welfare maximization. On a deeper view, however, one can see that sophisticated parties at the negotiation stage prefer to write contracts that maximize total benefits.[19] To see why, assume that each party's share of the contractual surplus is set exogenously. This assumption holds that a party cannot affect the size of its share of the parties' bargain by the (nonfraudulent) actions it takes during a negotiation. On this assumption, the parties will want only to maximize the total surplus.[20] To put this point in a contracting context, let parties contemplate making a simple sales contract for goods that the buyer values at $100 and that would cost the seller $80 to produce. Now assume that each party's share in the contracting surplus ($100 – $80) is fixed in advance at one-half each. Then the price would be $90, and each party's profit would be $10. Of more importance, assume that the seller could make a $2 investment in the subject matter of the contract that would lower its production cost to $70. The seller would want to make this investment because then its share of the new $28 contractual surplus ($100 – $2 – $70) would be $14, a share that would be realized by a price reduction to $86.

Gregory Mitchell, *Taking Behavioralism Too Seriously? The Unwarranted Pessimism of the New Behavioral Analysis of Law*, 43 WM. & MARY L. REV. 1907 (2002).

We provisionally view the individual decision experiments as not relevant to our project for three reasons. First, as we are in a world of speculation, we speculate that individuals in laboratories may perform worse than officers of firms because experimental subjects have not been trained to make good decisions and are not subject to the pressures to maximize that are described above. Second, recent evidence suggests that behavioral anomalies can be substantially mitigated or made to disappear when individuals are asked to perform as actors in firms, *see* Jennifer Arlen et al., *Endowment Effects Within Corporate Agency Relationships*, 31 J. LEGAL STUD. 1 (2002), or when the applicable institutions permit communication within a group of actors and require competition between groups, *see* TILMAN SLEMBECK & JEAN-ROBERT TYRAN, DO INSTITUTIONS PROMOTE RATIONALITY? AN EXPERIMENTAL STUDY OF THE THREE-DOOR ANOMALY (Universität of St. Gallen, Working Paper No. 21, 2002), http://www.vwa.unisg.ch/RePEc/usg/dp2002/dp0221slembeck_ganz.pdf. The view that organizations composed of experts make better decisions than individual subjects is coming to be accepted in the psychological community as well. *See* Jeffrey J. Rachlinski, *The Uncertain Psychological Case for Paternalism*, 97 NW. U. L. REV. 1165, 1214-16 (2003). Third, experiments show that when persons are required to function in markets rather than to perform individual tasks, the persons reach equilibria that are consistent with individually optimizing behavior. *See* Vincent P. Crawford, *Introduction to Experimental Game Theory*, 104 J. ECON. THEORY 1, 3 n.8 (2002) ("[R]epeated play of the same game often converges to equilibrium no matter what subjects are thinking."); Dhananjay K. Gode & Shyam Sunder, *Allocative Efficiency of Markets with Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality*, 101 J. POL. ECON. 119 (1993); Dhananjay K. Gode & Shyam Sunder, *What Makes Markets Allocationally Efficient?*, 112 Q.J. ECON. 603 (1997); Vernon L. Smith, *Rational Choice: The Contrast Between Economics and Psychology*, 99 J. POL. ECON. 877 (1991) (review article).

19. Parties need the law's help to deal with postcontractual opportunism. *See infra* Part III.

20. Put simply, if a party is to receive a fixed twenty percent of a joint gain, it would *always* prefer the joint gain to be $200 rather than $100.

Similarly, the buyer has an incentive to make cost-justified value-increasing investments.

It remains for us to show that parties' bargaining shares actually are set exogenously. In standard bargaining theory, bargaining power is a function of two factors. The first factor is the parties' relative patience. The more patient bargainer will reject offers it dislikes to wait for more favorable offers, while the less patient bargainer will accept relatively unfavorable offers just to get a deal.[21] The second factor is each party's disagreement point (or next-best option). Disagreement points affect bargaining power in a subtle way. To see how, suppose initially that the parties are equally patient (they have the same discount rate). They then will divide the surplus from a deal equally, if they do the deal at all. Now let d_s be the seller's payoff from its disagreement point, and let d_b be the buyer's disagreement point. The expected surplus if the parties contract is s. A deal would be efficient if it would generate a surplus that exceeded the sum of the parties' outside options (i.e., if $s > d_s + d_b$). The parties' disagreement points will not determine the split of the surplus s from such a deal if one-half s is greater than *both* d_s and d_b. Suppose, however, that one-half s is less than d_s, the seller's disagreement point. The seller would not contract with the buyer unless the seller would do at least as well in the deal as it could do elsewhere. In this variant of the example, the seller thus would receive a bargaining payoff that is its disagreement point d_s plus one-half of the surplus that remains after subtracting the parties' disagreement payoffs. Hence, disagreement points only affect bargaining power in a deal when one of the parties would do better taking its outside option than accepting the surplus split that the parties' discount rates would otherwise imply.[22]

21. A party's discount rate measures his patience: The higher a party's discount rate, the more highly the party values current dollars than future dollars. Parties with high discount rates thus are impatient bargainers: They want their share of the surplus now. A party for whom current dollars are relatively less important—a party with a low discount rate—suffers less from delay and, as a result, is more willing to reject low current offers. Hence, patient parties do well when bargaining with impatient parties, who will reduce their demands in order to reach agreement quickly.

22. There are two versions of the Nash bargaining game: "split the difference" and "deal me out." In split-the-difference bargaining, if a deal would be profitable, each party receives its disagreement payoff plus one-half of the remaining surplus; in deal-me-out bargaining, each party gets half the surplus (if they are equally patient) unless one party's disagreement payoff exceeds half of an equal split. To illustrate the difference, assume that the seller's disagreement point is \$2; the buyer's disagreement point is \$1; and the surplus from a deal between them would be \$10. The parties' payoffs under the two games (with split-the-difference set out first) are

(1) Seller: \$2 + ½ (\$10 – \$3) = \$5.50
Buyer: \$1 + ½ (\$10 – \$3) = \$4.50

(2) Seller: ½ (\$10) = \$5
Buyer: ½ (\$10) = \$5

Game theorists predict that deal-me-out bargaining is common because a party's threat to exit unless it receives its disagreement payoff plus one-half the remaining surplus seldom is credible. On the assumed payoffs here, the buyer thus would not believe a seller's threat to exit unless the

A business party's patience is a function of its ability to finance its projects. Firms that have capital or convenient access to capital can be more patient than firms that need revenue immediately to survive. Parties ordinarily cannot affect the access of prospective contract partners to the capital market. Moreover, one party to a possible contract ordinarily cannot affect the other party's alternative business opportunities (its disagreement point). Thus, each potential contract partner will realize that its share of the maximum surplus the parties could generate jointly has already been fixed before any contract is signed. Hence, each party will contract so as to maximize the size of the pie.

This result is contrary to the common view. It is widely believed that parties exercise bargaining power by requiring weaker contracting partners to take unfavorable terms. Thus, section 2-302(1) of the UCC authorizes a court to strike "any clause of the contract" if the clause is unconscionable.[23] Terms that superficially appear one-sided are commonly described as the product of "unequal bargaining power." But when bargaining power is determined prior to contract formation, as is common in business contexts, these views are incorrect. Bargaining power instead is exercised in the division of the surplus, which is determined by the price term. Parties jointly choose the contract terms so as to maximize the surplus, which the price may then divide unequally.[24]

seller received $2 plus one-half the residual surplus (the total $5.50 payoff) because the buyer would know that the seller never would reject $5 (its deal-me-out bargaining payoff) in favor of $2, its outside option. Anticipating the buyer's belief, the seller would agree to accept $5 rather than futilely attempt to get $5.50. Now let the seller's disagreement payoff increase to $6. Then the parties would switch from deal-me-out bargaining to split-the-difference bargaining: The seller would receive $6 + ½ ($10 – $7) = $7.50 and the buyer would receive $1 + ½ ($10 – $7) = $2.50. The switch in bargaining games is predicted to occur because when the seller would realize $6 from its outside option, its threat to exit unless it gets at least $6 in the bargain becomes credible. We use the deal-me-out bargaining game throughout this Article (except when the seller has a high outside option) because of the game theorists' logic and because experimental tests of bargaining behavior reject "split the difference" in favor of "deal me out." *See* COLIN F. CAMERER, BEHAVORIAL GAME THEORY 175-82 (2003); Ken Binmore et al., *An Outside Option Experiment*, 104 Q.J. ECON. 753 (1989).

23. U.C.C. § 2-302(1) (2003).

24. This conclusion applies even though parties may be uncertain about the amount of bargaining power they actually have. For example, parties can use a maximin strategy when they know the set of possible disagreement points but do not know which member of the set applies to their case. Parties using this strategy will evaluate uncertain gains by comparing their no-deal result to the smallest payoff possible under the set of potential disagreement points. If this minimum potential payoff exceeds the surplus in a no-deal condition, parties will contract; otherwise, they will not. As a consequence, some efficient deals will not be made, but those that are will be Pareto efficient. *See* Walter Bossert & Hans Peters, *Efficient Solutions to Bargaining Problems with Uncertain Disagreement Points*, 19 SOC. CHOICE & WELFARE 489 (2002). Our conclusion in the text is unaffected by this form of uncertainty because neither party can affect any of the disagreement points if the full set is exogenous. This set of models requires parties to reach agreement promptly on the basis of the possible choices available to them. When parties are optimistic about their bargaining power but can learn the truth by inference from the sequence of offers each of them makes, they will reach efficient bargains, though with delay. *See* Muhamet Yildiz, *Waiting To Persuade*, 119 Q.J. ECON. (forthcoming Feb. 2004). We do not claim that

B. *Why the State Should Help Firms*

We noted at the outset that there are four main objections to the single-minded pursuit of welfare maximization for commercial contract law. Section II.A argued that the cognitive objection is weak, and Section V.A argues against the fairness objection. We discuss here the externality and distributional objections. The externality objection is weak because, as a descriptive matter, most commercial contracts affect only the parties to them. A single sales contract that turns out badly is unlikely to put employees out of work or cause retailers in the firm's locality to suffer. It is a firm's *systematic* decisions that may affect third parties in material ways. For example, a firm may run a factory with disregard for the environment or the rights of its workers. Systematically inefficient or unfair behavior of this kind is subject to legal regulation under the headings of environmental and employment law. Relying on this specialization principle, we assume that the transactions regulated by contract law do not create externalities, unless there is a particular reason to believe that they do.

Commercial contract rules seldom create systematic distributional benefits for particular classes of parties. In the first place, commercial parties commonly occupy both the roles of seller and buyer (or licensor and licensee, etc.). As a consequence, a pro-seller rule would hurt firms when they buy, and a pro-buyer rule would hurt these same firms when they sell. In addition, because most commercial contract law rules are defaults, distributional benefits are hard to create even for firms that primarily buy or sell. Suppose, for example, that a contract rule allocates a risk to the selling side of the market in order to create a distributional benefit for the buying side. Suppose also that contractual surplus would be maximized if buyers bore the risk at issue (because, say, they are the cheapest cost avoiders). A contract allocating the risk to the buyer would make both the seller and buyer better off (because they would split a larger surplus). Consequently, the legal rule's allocation would be unstable. Because business firms attempt to maximize contractual surplus, the default rules that constitute the bulk of commercial law rarely could systematically benefit either side of the market.

Moreover, it is difficult to create distributional benefits for the shareholders who own most business firms. Shareholders typically hold diversified portfolios. A diversified shareholder often will own some firms that buy and sell, some firms that primarily buy, and others that primarily

parties always choose efficient contract terms. The existence of asymmetric information sometimes will cause parties to make constrained efficient contracts; these contracts are not "first-best" but are efficient given the information structure facing the parties. In Part V, we argue that the state seldom can improve on constrained efficient contracts because information that is unavailable to the parties is unlikely to be available to the decisionmaker.

sell. An attempt to benefit either side of the market distributionally is unlikely to create net gains for such a shareholder. Diversification is also normatively relevant. A diversified owner wants the value of his portfolio to increase, not the value of particular firms in his portfolio at the expense of other firms in his portfolio. Indeed, investors diversify precisely to escape firm-specific risk—the risk that a particular firm that the investor owns will have an unusually bad outcome. Satisfying this investor preference thus requires legal rules that maximize surplus across firms.

In sum, cogent reasons exist to justify our principal normative claim that contract law should facilitate the ability of firms to maximize welfare when making commercial contracts. The reasons set out here also imply, for this class of contracting parties, that it is unnecessary or futile for courts or statutory drafters to pursue distributional goals. The contract law of commercial parties is about efficiency.

III. THE ENFORCEMENT FUNCTION

A perennial question in contract law is why the state should enforce a contract against the wishes of a party to it. We exclude answers to this question that take the following form: The state should enforce a party's contractual promises the better to permit persons to enlist other persons in their projects, and thus to increase the sphere of autonomy within which persons can operate;[25] or the state should enforce promises to reinforce the morality of keeping them. These answers are ruled out here because the business firms that make commercial contracts are *artificial* persons whose autonomy the state need not respect on moral grounds, and whose morality is ordinarily required by positive law. The relevant question for a normative theory of commercial contract law is just when, if ever, the goal of welfare maximization requires legal enforcement of the contracts that business entities make.

25. For excellent analyses of the strengths and limitations of the various autonomy-based theories of contract law, see Richard Craswell, *Contract Law, Default Rules, and the Philosophy of Promising*, 88 MICH. L. REV. 489, 514 (1989); and Jody S. Kraus, *Philosophy of Contract Law*, *in* THE OXFORD HANDBOOK OF JURISPRUDENCE AND PHILOSOPHY OF LAW 687 (Jules Coleman & Scott Shapiro eds., 2001). We also are not interested in the question of when individual persons should keep their promises, as we want to identify the cases in which the *state* should enforce the agreements that firms make. For an interesting analysis of the duty-to-keep-promises question as applied to persons, see Niko Kolodny & R. Jay Wallace, *Promises and Practices Revisited*, 31 PHIL. & PUB. AFF. 119 (2003).

A. *Enforcement Often Is Unnecessary*

A contract has an intertemporal aspect: Parties agree today to do something tomorrow.[26] State enforcement of these agreements is unnecessary when the agreements fall within the self-enforcing range or can be enforced with reputational sanctions.[27] An agreement is said to be self-enforcing when the threat by either party no longer to deal with the other is sufficient *in and of itself* to induce performance.[28] Reputation, in turn, will induce performance when a single contract partner's boycott would not.[29] For reputation to work, however, potential future contracting parties must be able conveniently to learn why the original parties' deal broke down. Reputations, therefore, are difficult to establish in large economies in which particular contracting parties often are anonymous. Rather, reputations work best in small trading communities, especially those with ethnically homogenous members, where everything that happens soon becomes common knowledge, and boycotts of bad actors are easy to enforce.[30] Reputational sanctions also can be effective in industries that can establish trade associations; the associations become a form of collective memory regarding the contracting behavior of their members.[31]

This Article nevertheless focuses on contracts that fall outside the self-enforcing range and that cannot be enforced by reputational sanctions. We take this focus for two reasons. First, while informal business networks

26. Agreements often are written even when the parties expect not to enforce them legally. A writing reduces disagreements over what the parties had actually agreed to do. Disagreements as to what the contract directs raise interpretation issues that are discussed in Part IV below.

27. For good, largely informal discussions of these issues, see Benjamin Klein, *Why Hold-Ups Occur: The Self-Enforcing Range of Contractual Relationships*, 34 ECON. INQUIRY 444 (1996); and Robert E. Scott, *Conflict and Cooperation in Long-Term Contracts*, 75 CAL. L. REV. 2005, 2039-50 (1987).

28. Suppose *S* and *B* write a contract in a state that does not legally enforce contracts. *B* later learns that it could make $100 more by breaching the contract than by performing it. If *B* breaches, however, *S* will no longer deal with *B*. Let *B*'s expected profits on these future contracts have a present value of $200. Then *B* will perform the contract, though it could not be sued for breach, because breach would cause it to lose $100, that is $100 – $200 = –$100. The contract is self-enforcing.

29. For example, suppose that *S*'s later refusal to deal would impose only a $10 loss on the breaching buyer, but other sellers also will refuse to deal, raising the buyer's total loss from breach in present value terms to $200 ($10 plus a $190 reputational sanction). Again, *B* would voluntarily perform.

30. *See* Janet T. Landa, *A Theory of the Ethnically Homogenous Middleman Group: An Institutional Alternative to Contract Law*, 10 J. LEGAL STUD. 349 (1981). An excellent survey of early informal enforcement mechanisms is Avner Greif, *Informal Contract Enforcement: Lessons from Medieval Trade*, *in* 2 THE NEW PALGRAVE DICTIONARY OF ECONOMICS AND THE LAW 287 (Peter Newman ed., 1998).

31. For discussion on this point, see Lisa Bernstein, *Merchant Law in a Merchant Court: Rethinking the Code's Search for Immanent Business Norms*, 144 U. PA. L. REV. 1765, 1781-82 (1996) [hereinafter Bernstein, *Merchant Law*]; and Lisa Bernstein, *Private Commercial Law in the Cotton Industry: Creating Cooperation Through Rules, Norms, and Institutions*, 99 MICH. L. REV. 1724 (2001) [hereinafter Bernstein, *Private Commercial Law*].

can be optimal for the parties to them, they often reduce social welfare. Networks absorb the most reliable firms, thereby reducing the quality of exchanges in the anonymous market. Hence, unless a network encompasses much of the economy, which is unlikely, the network's existence reduces efficiency in society generally.[32] Society is therefore better off when it adopts laws that improve market functioning rather than when society eschews legal reform on the ground that private associations will emerge as satisfactory substitutes. Second, market exchange will be common even when business networks exist. Both of these reasons support the utility of asking how the state best facilitates contracting in market environments.[33]

Even so, the efficiency gains from enforcing contractual promises presuppose the existence of contracts, and commercial transactions often are conducted without them. Firms often make simultaneous exchanges of cash for goods or services rather than exchange promises for the later trade

32. For a general analysis of this issue, see, for example, Raja Kali, *Endogenous Business Networks*, 15 J.L. ECON. & ORG. 615 (1999).

33. Recent theoretical analyses and economic experiments suggest that the domain of self-enforcing contracts would expand beyond that set out in the text if market actors behaved fairly toward those who had behaved fairly toward them and punished actors who had behaved unfairly. In the experiments, subjects reciprocate good behavior even though they are not required to do so and punish bad behavior even though it is costly to do so. For a review of some of this evidence, see Ernst Fehr et al., *Reciprocity as a Contract Enforcement Device: Experimental Evidence*, 65 ECONOMETRICA 833 (1997); and ERNST FEHR & KLAUS M. SCHMIDT, THEORIES OF FAIRNESS AND RECIPROCITY—EVIDENCE AND ECONOMIC APPLICATIONS 2-3 (Univ. of Zurich, Inst. for Empirical Research in Econ., Working Paper No. 75, 2001), http://www.iew.unizh.ch/wp/iewwp075.pdf. Analyses of how market contracting would improve if parties engaged in reciprocal fairness can be found in Yongmin Chen, *Promises, Trust and Contracts*, 16 J.L. ECON. & ORG. 209 (2000); and Ernst Fehr & Klaus M. Schmidt, *A Theory of Fairness, Competition, and Cooperation*, 114 Q.J. ECON. 817 (1999).

Whether this literature applies to market contracting among firms remains an open question for two reasons. First, when individuals in the experiments were put in market contexts and deprived of information about the payoffs of other actors, reciprocity tended to diminish. These results caused a leading experimenter to speculate:

> The effects of multiple players and limited information [in experiments] suggest a general conjecture about bargaining and markets. In two-person games with perfect information about how much each side is earning, fairness concerns loom largest. . . . The concern for fairness evident in two-player perfect information games . . . disappears in large markets. This does *not* mean traders in such markets do not care about fairness per se. They may care, but they behave self-interestedly because they aren't sure whether others are being fair and can't easily punish unfairness. A competitive market is simply a place in which it is hard to express your concern for fairness because buying or selling (or refusing to do so) will not generally change your inequality much. This . . . just means that people will then express social preferences about unfair market outcomes through "voice" . . . , regulation, and law.

CAMERER, *supra* note 22, at 115. Second, the subjects in the experiments were individual persons who had not been trained in market behavior. Such subjects may respond differently from officers of firms, who commonly have market experience and who are subject to pressures to maximize profits. These two reasons suggest that it is premature to apply results in the fairness literature to an analysis of contracting among sophisticated firms. For a discussion of the relevance of the fairness literature to the enforcement of deliberately incomplete or indefinite agreements between individual actors, see Robert E. Scott, *A Theory of Self-Enforcing Indefinite Agreements*, 103 COLUM. L. REV. 1641 (2003).

of these goods or services. Under such conditions, property law is a sufficient encouragement to commerce because a party will only part with goods or money if the party values more highly what is offered in exchange. Thus, protecting property supports efficiency. But, in contrast to simultaneous exchanges, a contract is a set of promises regarding future behavior. Such promises are costly to make and to memorialize. In order to understand the role of the state in relation to contracting behavior, it thus is necessary to explain why parties will incur these costs when contracts are legally enforceable but not otherwise.

State enforcement is helpful to contracting parties in a number of contexts but is particularly important in the two cases that Sections III.B and III.C next discuss—when investment is relation-specific and when the realization of a bad state of the world would create serious disruption costs. We analyze these cases both because of their intrinsic economic significance and because of their relation to other aspects of contract law. For example, the need to cushion the effects of realizing an adverse state and the need to avoid an adverse contract interpretation that would create similar disruption costs can cause risk-neutral parties to act as if they were risk-averse. The recognition of this should influence legal doctrine. The next two Sections thus treat paradigmatic cases.

B. *Encouraging Relation-Specific Investment*

We earlier identified the canonical contracting problem as ensuring efficient ex post trade and efficient ex ante investment. We assume here, per the Coase Theorem, that parties can ensure efficient trade on their own. For example, if the parties had agreed to trade ten units, but it turned out that trading twenty units would maximize joint gains, then the parties could modify the contract to provide for delivery of the larger quantity. As we will see, ensuring efficient investment is more difficult. The investments we have in mind would include the production of specialized goods, the development of human capital specific to a particular deal, or research to acquire information about future costs or prices.

We develop a simple model to explain why the state plays an essential role in encouraging investment. In the model, contracts are not legally enforceable. The sellers in our story can function in two distinct markets. A seller can produce a generic version of a particular product and sell the generic in a competitive market at a price that equals cost (including a return on the seller's investment). The seller also can produce a specialized version of the product for buyers who are willing to pay the extra cost. To be precise, a buyer's valuation for the generic product is denoted v_g and the cost of the product is simply g. Thus, the generic product will sell at the price g (because price equals cost in competitive markets) and generate a

contractual surplus of $v_g - g$. Suppose a particular buyer values the specialized version of the product at v_s and that the product costs s to produce (where $s > g$). Importantly, we assume that the seller's investment to make the specialized product would not be redeployable; if the seller were to spend s but the deal were to break up, the seller would lose all of s. The parties prefer to produce the specialized product when that would maximize the contractual surplus. This leads to the efficient decision rule: Produce the specialized version when $v_s - s > v_g - g$.

To make our example more concrete, suppose the buyer values the specialized product at \$80 (v_s = \$80); its cost is \$50 (s = \$50); the buyer values the generic version at \$50 (v_g = \$50); and its cost is \$40 (g = \$40). On these values, the parties' efficient decision rule requires the seller to produce the specialized product: It would generate a surplus of \$30 while the generic would generate a surplus of \$10.

If the parties contract as the efficient decision rule requires, they will bargain over how to divide the expected \$30 surplus. The division of the surplus will be determined by the bargaining game the parties play, and we assume for the reasons given above that they engage in deal-me-out bargaining.[34] Deal-me-out bargaining will generate an equal split in our illustration if the parties are equally patient bargainers because each party's disagreement payoff (\$10 for the buyer and \$0 for the seller) is less than one-half the \$30 surplus from producing the specialized product. We suppose that the parties are equally patient, because firms commonly have similar costs of capital, so that a contract to produce the specialized product would split the bargaining surplus equally. An equal split is achieved by a \$65 price.[35]

The price at which the parties will ultimately transact, however, would not be \$65, because the buyer's incentive to cooperate vanishes after the seller invests s in the subject matter of the deal. To see why, assume that the contract was made as described. After the seller had made its investment, the buyer would have an incentive to demand renegotiation of the price. At that point, the investment cost s would have been sunk and so would be ignored in the new bargain: The only issue for the parties would be whether to trade the specialized product at some price or not to trade. Trading would produce a gross gain of \$80, the buyer's valuation, while not trading would produce no value. The parties thus would trade, dividing the \$80 gain

34. *See supra* note 22.

35. The buyer earns its valuation less the price, which must equal its payoff from bargaining. Since its valuation is \$80 and its bargaining payoff is \$15, the price must be \$65. The seller earns the difference between the price and its cost, which must equal its bargaining payoff. Since the specialized product costs \$50 to produce and the seller's bargaining payoff is \$15, again the price must be \$65. Notice that a buyer's threat to exit unless it receives its disagreement payoff (\$10) plus one-half the surplus after this payoff is deducted would not be credible: The seller would know that the buyer would not pass up \$15, its payoff from a deal, to take \$10, its outside option.

equally. If the price were reduced to $40, the buyer would receive a $40 payoff and the seller would lose $10 ($40 less its cost of $50). Since the seller would lose its entire $50 investment if the parties failed to trade, it would agree to the new price.

The lesson that this example teaches is not that the parties' ultimate transaction prices would differ from their initial contract prices; the point, rather, is that when contracts are unenforceable a sophisticated seller will refuse to produce the specialized product, even though producing it would maximize expected surplus. The seller would anticipate losing $10 under a renegotiated contract to produce the product rather than earning $15 under the initial contract. Therefore, the seller would produce the generic product. This result is inefficient. The generic product generates a social surplus of $10, while the specialized product would have generated a surplus of $30.[36]

The parties would cooperate to produce the specialized product, however, if the buyer's promise to pay the contract price were legally enforceable. Under the UCC (and the common law of contracts) the seller could treat the buyer's demand to renegotiate rather than perform as an anticipatory breach.[37] The seller then would be entitled to recover the price if the goods contracted for could not be resold at a reasonable price.[38] Since the seller could not resell the specialized product for a positive price (its investment, recall, is assumed not to be redeployable), it thus could recover the original $65 price from the buyer. Knowing this, the parties would write the contract to produce the specialized product and trade it for $65. The seller would anticipate being compensated for its investment, and the buyer would prefer to have the specialized product and realize a $15 payoff rather than have the generic product and realize a $10 payoff.

36. The Nash bargaining that the text describes is ahistorical: Only the total surplus, the parties' discount rates, and (sometimes) their disagreement points determine the bargaining outcome. Hence, the parties in the examples in the text are assumed to ignore the seller's sunk $50 investment cost when renegotiating the contract. Individual persons in experiments and surveys sometimes take sunk costs into account, however, so that a party's payoff in a bargain will increase if the other party knows that its partner has spent money to prepare. This will increase the party's incentive to invest. *See* CAMERER, *supra* note 22, at 85-90; Lorne Carmichael & W. Bentley MacLeod, *Caring About Sunk Costs: A Behavorial Solution to Holdup Problems with Small Stakes*, 19 J.L. ECON. & ORG. 106 (2003); Joep Sonnemans et al., *On the Relation Between Asset Ownership and Specific Investments*, 111 ECON. J. 791 (2001). Whether corporate entities are motivated by the fairness concerns regarding sunk costs that individual experimental subjects act upon is unknown. When parties can stage investments across periods, there is an equilibrium of the dynamic bargaining game they play in which parties invest more than the parties in the model we use, even when those parties are unable to enforce their contracts. *See* Yeon-Koo Che & József Sákovics, A Dynamic Theory of Holdup (Apr. 25, 2003) (unpublished manuscript, on file with authors). Whether the stage-investment assumption applies in many cases is also unknown. Thus, we do not claim that the ability of a party to hold up its contract partner will always cause underinvestment—i.e., a refusal to make the specialized product—when contracts are not enforceable; rather, our claim is that underinvestment would occur often enough to make legal enforcement worthwhile.

37. *See* U.C.C. §§ 2-610(b), 2-703(e) (2003).

38. *See id*. § 2-709.

This analysis supports two conclusions. First, contract remedies are thought to protect injured promisees—the seller here—by awarding the expectation interest. This view is true but shallow. If contracts were not enforceable, sophisticated commercial parties seldom would put themselves in positions where they needed the law's aid. They would instead act as would the seller here, who would produce the generic product and sell it on the market rather than subject itself to exploitation. Enforcement actually empowers promisors by enabling them to make credible promises to perform or to pay. The buyer in our example, when a contract is contemplated, thus *wants* the power to make a legally enforceable—that is, a credible—promise to pay the seller the $65 contract price. Enforcement, in sum, permits parties to make believable promises to each other when reputational or self-enforcement sanctions will not avail.

Second, and relatedly, our example helps to explain the very small amount of foreign direct investment that private parties have made in the former Soviet states and in many Third World countries. Much of this investment would have been relation-specific (e.g., building a factory far from the home country, developing a mine or an oil field). Potential investors would not deal unless the host country or local firm could make credible promises to adhere to the terms originally agreed upon rather than renegotiate those terms after investments had been made. The lack of enforcement rules and honest courts in many of these countries, however, prevents the local parties from making promises that are more believable than was the buyer's promise in the example above. In response, foreign parties reduce investment. The ability of a firm to make a credible promise, which lawyers in developed countries take for granted, is a regrettably rare power in many parts of the world.[39]

C. *Contracting To Avoid Disruption: The Case of Volatile Markets*

Parties that function in "thick" markets[40] have a choice of making a contract for future delivery or making a spot purchase—that is, a simultaneous exchange of cash for goods. A fixed-price contract for future delivery is understood to allocate the risk of price declines to the buyer and

39. For a vivid example of the second-best strategies used when attempting contract enforcement in weak states, see Timothy Frye & Ekaterina Zhuravskaya, *Rackets, Regulation, and the Rule of Law*, 16 J.L. ECON. & ORG. 478 (2000). As the authors explain,

> [P]rivate protection rackets primarily provide two services. First, they provide basic protection from other rackets and from criminals. Second, and to a lesser extent, they help enforce agreements. These results suggest that private protection serves first as a substitute for the notoriously ineffective Russian police forces. To a lesser extent, it also serves as a substitute for the notoriously ineffective Russian courts.

Id. at 491-92.

40. A thick market exists when there are many sellers and buyers trading a roughly homogenous product.

of price increases to the seller. It is less well understood why parties write these contracts when contracting is costly and the spot option is available. We set out another simple model that states this question formally and we then attempt to give a plausible answer.

In the model, the parties can make an enforceable contract at time T_0 for delivery of goods at time T_1, or the buyer can wait until T_1 to make a spot purchase. The parties believe at T_0 that the T_1 market price could take one of three values: (1) p_k, (2) $p_k + z$, or (3) $p_k - z$. Each of these outcomes is thought to be equally likely; hence, viewed from T_0, the expected T_1 price is p_k.[41] This also will be the price paid at T_0 for a promise to deliver the goods at T_1. To see why, realize that if the T_0 price exceeded p_k, sellers at T_0 would enter the market to sell contracts for future delivery at the high T_0 price. The resulting increase in supply would cause the T_0 price to fall. On the other hand, if the T_0 price were below p_k, buyers would enter to buy contracts for future delivery at the low T_0 price. The resulting increase in demand would cause the T_0 price to rise. Hence, the unique equilibrium T_0 price must be p_k. As a consequence, if the parties did write a contract at T_0 for T_1 delivery, the contract price would be p_k. To complete the model, we assume that the seller's cost to produce or buy the goods is $c \leq p_k$ and that the buyer's valuation for the goods is $v = p_k + y$. On these assumptions, trading the goods would be efficient because the buyer's valuation exceeds the seller's cost. The total cost of writing a contract (instead of the buyer waiting until T_1 to buy) is $w > 0$. Since price equals cost in a competitive market, the seller's fraction of w will be included in the price; the buyer thus will bear all of w—the seller's fraction of the price, plus the fraction which the buyer pays out of pocket.

Actual numbers might make this story a little more concrete. We let p_k = \$100 and w = \$2. We assume further that, for simplification, the buyer's valuation is equal to the high market price for the good in question, so that $z = y$ = \$20 (and thus $v = p_k + y$ = \$120). The question is whether the buyer will make the contract at T_0 or make a spot purchase at T_1.

In this model, the buyer would wait. If the buyer did contract, and contracts were enforceable, the buyer would pay the price p_k and realize its net valuation y with certainty, either because the seller performed or through a damage recovery. The buyer's gain from contracting thus would be its net valuation less contracting costs, or $y - w$. On our assumed numbers, this gain would be \$20 – \$2 = \$18. The buyer's *expected gain* as of T_0 if the buyer did not contract, however, would be the larger sum y, or \$20. Without a contract, the buyer would have to purchase at the T_1 market price, whatever that price turned out to be. Since there are three equally likely future prices, the buyer's expected return from waiting until T_1 to buy

41. The expected T_1 price is $\frac{1}{3}(p_k) + \frac{1}{3}(p_k + z) + \frac{1}{3}(p_k - z) = p_k$.

(measured as of T_0), is

$$E(G_{nk}) = \tfrac{1}{3}\,[(p_k + y) - (p_k + z)] + \tfrac{1}{3}\,[(p_k + y) - p_k] + \tfrac{1}{3}\,[(p_k + y) - (p_k - z)] = y.$$

The buyer's spot purchase would either be made at the high T_1 market price ($p_k + z$), the average T_1 market price (p_k), or the low T_1 market price ($p_k - z$). Given a buyer valuation of $p_k + y$, and recalling that y is assumed to equal z, the buyer's expected gain from not contracting, and thus saving the contracting cost w, is just the buyer's net valuation y, or \$20.[42]

A risk-neutral buyer would not pay a premium to ensure a certain gain; rather, the party would choose a higher expected but risky gain. The "premium" here would be the contracting cost (w = \$2). Paying that cost would guarantee the buyer a sure \$18 gain, which is less than the preferred expected gain of \$20. A risk-neutral seller would reject a contract for the same reason. In fact, spot purchases are very common. And if parties would not write contracts for future delivery in thick markets even if these contracts were legally enforceable, there is no need to make them legally enforceable. Thus, the question why risk-neutral firms sometimes write contracts for future delivery is not trivial.

We argue that parties write these contracts (when they are enforceable) to reduce the risk of potentially disastrous outcomes that would impose additional costs on the buyer. If the buyer in the example did not contract, then one-third of the time it would have to make a spot purchase at the high T_1 price of $p_k + z$, or \$120. Since the buyer's gross valuation is also \$120 (that is, $p_k + y$), one-third of the time a purchase of the goods would contribute nothing toward the buyer's fixed obligations. A buyer who failed to pay rent or interest, however, would incur serious disruption costs. Once we incorporate disruption costs (denoted as f) and realize that these would be incurred one-third of the time, the comparison in the model must be modified. The buyer's actual decision rule would be to contract when the net gain from contracting exceeded the net gain from not contracting, or when $y - w > y - \tfrac{1}{3} f$, or when $w < \tfrac{1}{3} f$. In this example, if disruption costs exceeded \$6, the buyer would write a contract for future delivery.

This analysis predicts that a buyer who contemplates making many transactions, none of which will be large in relation to the buyer's need for cash, will act as would a risk-neutral person, purchasing goods on the spot market. When a bad realization on a single contract could seriously endanger the buyer's business, however, the buyer will act as if it were risk-averse, purchasing the assurance of performance at the cost of writing a contract for future delivery. To be sure, some such buyers could guard against disruption by maintaining sufficient cash reserves, but this strategy

42. Using our illustrative numbers, the buyer's expected gain without a contract is $E(G_{nk})$ = ⅓ (\$120 – \$120) + ⅓ (\$120 – \$100) + ⅓ (\$120 – \$80) = \$20.

is difficult for all businesses to follow. Also, if the buyer has good business prospects, the opportunity cost of hoarding cash likely would exceed contracting costs, especially for the simple fixed-price contracts we consider. The result of this analysis is that the state reduces social costs by giving parties that function in volatile markets the opportunity to make enforceable contracts for future delivery.[43]

In sum, when exchange is intertemporal rather than simultaneous, efficiency is enhanced when parties can make enforceable contracts in two principal situations: when at least one of the parties is required under the contract to make an investment that is more profitable in the relationship than elsewhere, and when market prices are volatile and an adverse market movement can have spillover effects.[44] In the first case, in the absence of legal enforcement the noninvesting party has an incentive to renegotiate the contract price downward rather than to perform under the original contract. In the second case, the party whom a market movement disadvantages may suffer disruption costs that would much exceed its expectation interest (as conventionally measured). Contracts do sometimes fall within the self-enforcing range and in some subeconomies reputation can make promises to perform credible. But nonlegal incentives can be ineffective in larger markets and in countries where social trust is low. Thus, without legal enforcement, private contracting parties cannot be expected often to create deals that maximize social surplus.

D. *Enforcement and Duress*

Enforcement entails more than simply ordering a recalcitrant party to perform. As an illustration, suppose that a seller sold its goods to a third party instead of delivering them to the contract buyer. A specific performance order would thus be futile. Should the buyer be able to sue the seller for damages or to impose a constructive trust on the proceeds of the sale? If the buyer makes a substitute purchase, is the seller's obligation discharged? Suppose instead that the buyer takes delivery of the goods but

43. We have argued that contract enforcement serves an insurance function in the volatile markets case. Parties also can purchase business-interruption insurance, but the transaction costs of this alternative would ordinarily exceed the costs of the simple fixed-price contracts we consider. Companies face pressure to settle "bet the ranch" lawsuits for reasons similar to those developed here; settlement is a form of insurance against being put out of business. *See* J.B. Heaton, Settlement Pressure 30-36 (Feb. 2002) (unpublished manuscript, on file with authors).

44. Recall our assumption that parties to business contracts are risk-neutral. A third motive to contract is to transfer risk from more to less risk-averse parties. The legal enforcement of these contracts sometimes is necessary because the transferee of risk has an incentive to breach when large risks materialize. Risk-shifting contracts are not considered here, in part because one of the parties to them commonly is an insurer, and insurance contracts are the subject of a distinct and heavily regulated legal field. Moreover, although many contracts have an insurance component (e.g., commodities contracts, currency hedging), these contracts tend not to give rise to litigation.

claims that they do not conform to the quality the seller promised to deliver. Must the buyer still pay the price and sue for damages or can the buyer cancel the sale? Finally, suppose that, after the contract is made but before delivery, a federal agency passes a regulation that prohibits production of the product the buyer purchased the goods to make. The buyer no longer needs the goods. Must the buyer still pay the price? If not, does the seller have a remedy?

These questions illustrate the complexity of the concept of "enforcing a contract." It is tempting to suggest, therefore, that a supplementary set of publicly supplied "enforcement rules" also is needed. This suggestion would be premature, however, because parties can answer these questions in their own contracts. For example, parties can write a force majeure clause specifying the events that would excuse the seller's obligation to deliver or the buyer's obligation to pay; these events could include the possible passage of an unfavorable administrative regulation. That the contract laws of advanced nations commonly contain sets of enforcement rules thus requires explanation, a task that we address in Part V of this Article.

The duress doctrine, however, is an enforcement rule that parties cannot create on their own. The law of duress applies in two contexts. Ex ante duress occurs when a party is wrongfully coerced to make a contract.[45] Ex post duress occurs when a party is wrongfully coerced to modify an existing contract.[46] Contract law applies the same legal standard in *both* cases: A contract or a modification is unenforceable if a party's consent thereto was obtained by an improper threat that left the party no reasonable alternative but to submit.[47] Our focus here is on ex post duress, and we suggest that courts should ask a different question from those asked in ex ante duress cases. In an ex post duress case, the contact was fairly obtained and the parties could have provided for the situation that later arose had they thought about the issue. The court thus should ask whether parties with

45. An ex ante duress argument succeeds if the party proves that he would not have made the challenged contract absent the improperly coercive behavior of the other party. The key is the law's focus on behavior rather than circumstances. Thus, it is not duress on the part of an employer when a poor person accepts an offer to work at a low wage, nor is it duress on the part of a seller to charge a high price for gas when it is the only seller for many miles. The employer does not create the employee's low wealth, nor does the seller lure the buyer to the solitary spot. Contract law thus requires a duress claim to rest on the behavior of the promisee, not on the preexisting circumstances of the promisor. Chouinard v. Chouinard, 568 F.2d 430, 434 (5th Cir. 1978). This is because absent coercion by the promisee, the promisor does better by contracting than by not contracting. *See* Robert E. Scott & William J. Stuntz, *Plea Bargaining as Contract*, 101 YALE L.J. 1909, 1919-20 (1992) (discussing the general proposition that "the wrongful acts that constitute duress may be either physical force or an improper threat, but in any case the compulsion must be produced by the promisee and not by exigent circumstances facing the promisor").

46. *See, e.g.*, Austin Instrument, Inc. v. Loral Corp., 272 N.E.2d 533 (N.Y. 1971).

47. RESTATEMENT (SECOND) OF CONTRACTS § 175 (1981).

sufficient foresight would have wanted the later modification agreement to be enforceable.

Two examples show how this test should be applied. For the first, assume that the parties initially agreed to trade thirty units at a price of $10 each. Demand for the product turns out to be higher than the parties initially believed; it becomes efficient to trade fifty units. The seller offers to transfer an additional twenty units at a unit price, for all fifty, of $12. The buyer's profit would be higher under the contract as modified, and it agrees to the new terms. The buyer cannot later claim that it was coerced to accept a price increase because, ex ante, the buyer would have wanted the court to enforce a modification that would leave it better off than performance under the original contract would have done.

For our second example, return to the relation-specific investment model set out above. There, we claimed that permitting the seller to sue for the price would deter the buyer's threat to renegotiate after the seller had invested. This claim is too strong because sellers in some cases could not make a credible threat to sue. The seller's threat would be credible only if it had, or had convenient access to, the capital needed to sustain it until the buyer's performance could be replaced. The buyer then would know that it would have to pay voluntarily or involuntarily. In contrast, a seller in a short-term bind (perhaps it had purchased materials on credit and is facing demands for payment) may be better off accepting a low renegotiation price than bringing a lawsuit. Recognizing this, the buyer may demand renegotiation even though the contract is enforceable. The renegotiated contract, however, would be a product of ex post duress. When the initial contract was made, both parties would have wanted a court not to enforce a purely redistributional modification—that is, a modification that would create no new wealth but rather would only redistribute the contractual surplus differently from the original contract. Parties dislike purely redistributional modifications for two reasons. As Section III.B showed, the anticipation of such modifications can destroy the parties' incentive scheme for producing efficient specialized products. Further, the resources involved in negotiating the modification or guarding against it constitute a deadweight loss that reduces the parties' joint gain from the contract.[48]

The duress doctrine thus permits the seller to perform under the renegotiated contract but later to reinstate the price term in the original contract. This is because modifications made under duress are not enforceable. As a consequence, the seller in our example could accept the

48. Ex post duress cases are largely consistent with the test we propose, though the courts formally apply the *Restatement*'s standards. *See, e.g.*, Wolf v. Marlton Corp., 154 A.2d 625 (N.J. 1959); *Austin Instrument*, 272 N.E.2d 533. A more extensive analysis of the ex post duress case is in Alan Schwartz, *Relational Contracts in the Courts: An Analysis of Incomplete Agreements and Judicial Strategies*, 21 J. LEGAL STUD. 271, 308-13 (1992).

$40 renegotiation price, deliver the specialized goods, and later recover the difference between the $65 contract price and the lower renegotiation price. Sellers who have easier access to cash in the long run than in the short run, or who can sell their legal claim, will use the duress doctrine to recover their original expectation. Buyers, therefore, will know that ex post promises by sellers that do not move the parties to Pareto-superior states themselves are not credible. The law will permit a seller to renege on such a promise and sue for full payment. The unreliability of renegotiation promises coerced by duress reduces the incentive to extract them (that is, to behave as the buyer in our example did). The ex post duress doctrine thus is an important aspect of a publicly supplied enforcement function.[49]

In this Part, we have argued that firms need state enforcement in order to permit them to make credible commitments when their promises are not self-enforcing. A court cannot enforce a contract, however, without first determining what the contract says. Thus, the parties' preference for state enforcement entails a further preference over the set of interpretive theories that courts could use to interpret their agreements. We next attempt to identify the interpretive theory in this set that typical firms will most commonly prefer.

IV. THE INTERPRETATION FUNCTION[50]

A. *The Relevant Interpretive Question*

There is a consensus among courts and commentators that the appropriate goal of contract interpretation is to have the enforcing court find the "correct answer." The "correct answer" is the solution to a

49. As may be obvious, the doctrine is symmetrical: It also applies to protect buyers who have made relation-specific investments from overreaching sellers. The doctrine would be unnecessary if capital markets were perfect, or if damage suits were perfectly compensatory. Then parties could finance all good lawsuits or recover all losses. Capital markets, however, are imperfect and damage awards are not always fully compensatory.

50. Part IV concerns cases in which the parties have attempted to solve their problem with written words. The court, that is, is not called upon to fill gaps, but rather is asked to discover what the parties intended their written words to do. A theory of interpretation has two aspects: a set of rules for determining the semantic content of a party's utterances, and a set of rules for determining the legal significance that should attach to the semantic content. We focus on the former set of rules here because we have already developed a normative theory: Courts should enforce business contracts as the parties to them would want the contracts enforced. Our view regarding the separability of the rules determining legal significance from the rules governing interpretation is not free from difficulty, however. Courts doing interpretation may be influenced by their view of what a good substantive outcome is when the parties' meaning is not apparent. *See* Richard Craswell, *Offer, Acceptance and Efficient Reliance*, 48 STAN. L. REV. 481, 551-52 (1996). We ignore this complication on the assumption that if the semantic content rules improve, the courts' need to give legal answers to factual intent questions will lessen. A full theory of contract interpretation would explore more seriously the actual and ideal relation between a court's substantive preferences and its choice of an interpretive style.

contracting problem that the parties intended to enact. Intention, however, is determined objectively and prospectively: A party is taken to mean what its contract partner could plausibly believe it meant when the parties contracted.

There are two justifications for the goal of finding the correct answer. The first follows from an autonomy-based view of contract law. This justification holds that the exercise of state coercion against a person must be justified. A sufficient justification is that the court is making the person do what he had agreed to do.[51] Hence, the court must ascertain just what the person had agreed to do. The second justification is consistent with an efficiency-based view of contract law. On this view, parties contract to maximize the surplus that their deal can create. This goal is unattainable if courts fail to enforce the parties' solution but rather impose some other solution. Thus, the court must ascertain the solution that the parties actually adopted. There is a dispute in the literature as to whether the rules that courts use when attempting to find the correct answer are mandatory in the formal sense, but there is a consensus that the rules are difficult for parties to escape.[52]

In our view, the current consensus asks the wrong question. A commitment to party sovereignty regarding the contract's substantive terms implies a further commitment to party sovereignty regarding the interpretive style an adjudicator should use to find the substantive terms. Party preferences regarding judicial interpretive styles can differ. Therefore, interpretive styles should be defaults. The relevant question, then, is what should be the majoritarian default. Put another way, the issue is not what interpretive style is best calculated to yield the correct answer. Rather, the issue is what interpretive style would typical parties want courts to use when attempting to find the correct answer. We will argue here that the majoritarian default is Willistonian: Typical firms prefer courts to make interpretations on a narrow evidentiary base whose most significant component is the written contract.[53] This proposed rule would both reverse the UCC's interpretive style and make the new interpretive style a default.

51. Most autonomy-based theories are premised on either a notion of "consent" or the exercise of will, such as the making of a promise. *See* sources cited *supra* note 25.

52. A recent commentator referred to the UCC's interpretation rules as "quasi-mandatory," the idea being that the rules are and should be very difficult for parties to avoid. David V. Snyder, *Language and Formalities in Commercial Contracts: A Defense of Custom and Conduct*, 54 SMU L. REV. 617, 648 (2001). Because contracting has positive costs, a quasi-mandatory rule will be mandatory in practice for many parties, who will be unwilling to bear the additional costs of specifying alternative regimes. Other commentators believe that the interpretation rules are mandatory. *See, e.g.*, Omri Ben-Shahar, *The Tentative Case Against Flexibility in Commercial Law*, 66 U. CHI. L. REV. 781, 792 (1999).

53. *See* 4 SAMUEL WILLISTON, CONTRACTS § 631, at 948-49 (Walter H.E. Jaeger ed., Baker, Voorhis & Co., Inc. 3d ed. 1961) (1920) (explaining that the parol evidence rule "requires, in the absence of fraud, duress, mutual mistake, or something of the kind, the exclusion of extrinsic evidence, oral or written, where the parties have reduced their agreement to an integrated

B. *Two Interpretive Issues: Problems of Meaning and of Language*

We begin by clarifying two interpretive issues that are commonly commingled: What does the language of the contract mean? And in what language was the contract written? To see why these issues are distinct, suppose that there are two sets of linguistic communities. The first set, called *M*, consists of a single linguistic community. This *M* community is composed of judges, lawyers, business persons, and potential jurors. The members of *M* read and write a language that we call "majority talk," since it is the language that people typically use when communicating with each other. The second set is called *P*, and it has many linguistic communities. A community in the set *P* may be as small as the parties to a particular contract, or as large as the entire trade in which the parties function. The set *P* has many communities because there are many party dyads and many trades. The members of each community in *P* may write contracts in their own "party talk" or write them in majority talk.[54]

The existence of multiple linguistic communities raises the two interpretive issues just noted. Imagine that parties are engaged in a dispute regarding the meaning of their written agreement. Words can be vague or ambiguous.[55] If the parties agree on the language in which their contract was written, the court's interpretive task is limited to finding what the parties intended that language to say. If the parties divide on the question of what language they used, the court's interpretive task expands: The court

writing"). Willistonian formalism rests on two basic claims: (1) that contract terms can be interpreted according to their plain meanings, and (2) that written terms have priority over unwritten expressions of agreement. *See* Dennis M. Patterson, *Good Faith, Lender Liability, and Discretionary Acceleration: Of Llewellyn, Wittgenstein, and the Uniform Commercial Code*, 68 TEX. L. REV. 169, 187-88 (1989).

54. Our use of the word "language" is a little loose. Linguists would say that when everyone uses English words, majority talk and the various types of party talk are different dialects of the same language. We use a multilanguage descriptor because it seems to us to be a more convenient system of classification.

55. Courts seldom distinguish between "vague" and "ambiguous" terms. A typical judicial definition of ambiguity, for example, includes any term or word that "has no definite significance or . . . is capable of more than one sensible and reasonable interpretation." Ross Bros. Constr. Co. v. State, 650 P.2d 1080, 1082 (Or. Ct. App. 1982). More narrowly, however, a word is vague to the extent that it can apply to a wide spectrum of referents, or to referents that cluster around a modal "best instance," or to somewhat different referents in different people. *See, e.g.*, Frigaliment Importing Co. v. B.N.S. Int'l Sales Corp., 190 F. Supp. 116 (S.D.N.Y. 1960) (analyzing whether "chicken" includes all types of chicken or only a subset); Highley v. Phillips, 5 A.2d 824, 826 (Md. 1939) (determining whether the sale of "all the dirt" from a tract refers also to subsurface sand). In contrast, "ambiguity" requires at least two distinct, usually inconsistent meanings. *See, e.g.*, Petroleum Fin. Corp. v. Cockburn, 241 F.2d 312, 314-16 (5th Cir. 1957) (analyzing missing punctuation in a telegram that supported two different readings); Raffles v. Wichelhaus, 159 Eng. Rep. 375, 375 (Ex. 1864) (determining the seller's obligation when two ships named *Peerless* sailed from Bombay but the seller was referring to one of them and the buyer understood it as the other). Language commonly is vague in the sense that the set of objects to which a word applies is rarely delineated with absolute precision. True lexical ambiguity occurs infrequently.

now must ask initially whether the parties wrote their contract in majority talk or in a particular private language. This question raises a separate interpretive issue because a contract's language could plainly mean m_M in community M but also plainly mean m_P in a community within P, where m_M and m_P differ. For example, the word "wife" in a sentence in John's will reciting "I leave my money to my wife" would mean in the M community that John left his money to the woman to whom he was legally married when he died. In a well-known case, however, John and another party wrote a contract in order to dispose of part of John's estate upon his death.[56] This contract used the word "wife" to refer to the woman with whom John was living when he died but whom he had never legally married; John had deserted his legal wife years before without divorcing her. The word "wife" in this example would be clear to members of the single community within set M and would also be clear to the particular community within P to which the contracting parties belonged. The two meanings differed, however.[57] The same word or phrase can thus have different meanings in different linguistic communities, which requires courts to make a choice: Should courts permit parties to write contracts in the language of the parties' choosing, or should courts create an incentive for parties to use majority talk by interpreting their agreements as if the agreements were written in that language?[58]

There are two traditional approaches to finding the correct answer to questions of contractual meaning and contractual language. They differ in

56. *In re* Soper's Estate, 264 N.W. 427 (Minn. 1935).

57. For a modern example, see Columbia Nitrogen Corp. v. Royster Co., 451 F.2d 3 (4th Cir. 1971). In that case, the contract specified a "Minimum Tonnage Per Year" of 31,000 tons. The buyer took less than that amount, but the court permitted the buyer to introduce evidence that, in the parties' trade, stated minimum tonnages were "mere projections to be adjusted according to market forces." *Id*. at 7.

58. The distinction between party and majority talk can blur at the edges, but remains easy to draw in most cases. When party talk closely correlates with majority talk—e.g., when a contract's terms are used to convey a subset of the general meanings that can be attributed to the same terms in majority talk—then the line between majority and party talk may be difficult to draw. Consider a contract that requires the seller to deliver "red sweaters." Given the breadth of the category "red" in majority talk, evidence regarding usage of trade may be helpful in delimiting the shades of red that the parties considered permissible. It remains unclear, however, whether the parties intended to convey a more limited subset of meaning for the term "red," specific to the parties, or sought instead to convey the broader meaning associated with the majority definition: The distinction between majority and party talk can be difficult to draw in such an instance.

By contrast, contractual terms that obviously contradict majority understandings provide a clear line between majority talk and party talk. For example, in *Hunt Foods & Industries v. Doliner*, Doliner provided Hunt Foods with an "unconditional" written option to buy all of the stock in his company at a stated price by a stated date. 270 N.Y.S.2d 937, *aff'd*, 272 N.Y.S.2d 686 (App. Div. 1966). The parties, however, intended "unconditional" to mean conditional. When a contract's terms so patently contradict majority meaning, the contract is clearly written in party talk. The New York court's decision to permit Doliner to testify as to the contract's actual meaning can be seen as a permission for the parties to write in their own private language. *Id*. We return to the two-language distinction in a moment. It is also discussed further in Section IV.D below.

the scope of the evidentiary base each requires to make interpretations. To be sure, a minimum evidentiary base is required for any coherent interpretation. This minimum base is denoted here by B_{min} and is composed of the parties' contract, a narrative concerning whether the parties performed the obligations that the contract appears to require, a standard English language dictionary, and the interpreter's experience and understanding of the world. A Willistonian, or "textualist," theory of interpretation assumes that contracts often have "plain meanings" that are apparent to judicial interpreters. Put formally, this view asserts that a court can find the correct answer on the evidentiary base B_{min}. Courts have added five evidentiary categories to this minimum base: (1) the parties' practice under prior agreements; (2) the parties' practice under the current agreement; (3) testimony as to what was said during the negotiations; (4) written precontractual documents (memoranda, prior drafts, letters); and (5) industry custom relevant to determining what the agreement's words meant to the contracting parties. When an adjudicator admits evidence in all five additional evidentiary categories, we denote the evidentiary base as B_{max}. A "contextualist" theory of interpretation holds that a court is more likely to find the correct answer if the evidentiary base expands toward B_{max}.[59]

A Willistonian theory of interpretation has the obvious effect of truncating the evidentiary base that a court uses to make interpretations, and it also has the effect of creating an incentive for parties to write contracts in majority talk. Because the same word can have plain but different meanings in the single community within set *M* and in a particular community within set *P*, a party seldom could establish that its contract was written in party talk unless it could introduce extrinsic evidence.[60] The Willistonian theory bars this evidence.[61] In addition, on this theory the court primarily considers the text of the contract; hence, the parties likely can reduce interpretive errors by writing the contract in the court's language. Thus, the Willistonian theory actually resolves two distinct issues: what language the parties should speak and what evidence is admissible to show what the parties meant in the permitted language. Contextualists resolve the same issues, though differently. Their theory lets courts consider all material evidence to resolve interpretive issues; the practical effect of so widening the

59. Contextualist scholars do not necessarily insist on B_{max}. It is recognized, for example, that evidence in category (3) is less reliable than evidence in category (4).

60. Extrinsic evidence refers here to evidence in the five additional evidentiary categories listed in the text.

61. A court applying the Willistonian theory will admit extrinsic evidence only when the contract's language is vague or ambiguous on its face. *See* Pysell v. Keck, 559 S.E.2d 677, 678-79 (Va. 2002).

evidentiary base has been to permit parties to write in whatever language they choose.[62]

Any analysis of contractual interpretation thus should answer two questions. The first is the question of meaning: Should courts use broad or narrow evidentiary bases in determining the meaning of the contract's language? The second is the question of language: Should the "linguistic default" assume that parties wrote in majority talk unless their contract recites otherwise, or should courts always admit the possibility that the parties wrote in a private language? The practical effect of admitting this possibility is to permit a party to introduce evidence in all of the evidentiary categories to show what language the parties actually used. Section C next argues that when the issue of the contract's language is settled but the meaning of that language is arguably unclear, the majoritarian default should require courts to make interpretations using the evidentiary base $B_{\min}$. Section D then argues that the linguistic default should suppose parties to have used majority talk.

C. *The Parties' Preferences Regarding Interpretive Styles*

An interpretive style can be assessed along two dimensions: (1) the likelihood that the style will generate the correct answer (as defined above); and (2) the costs that the style imposes on courts and parties. We model the performance of judicial interpretive styles on these dimensions in two ways that, in combination, capture most of the relevant cases. Both models suppose that the contract is complete in the sense that the writing expresses the parties' solution to the contracting problem at issue. The parties are aware, however, that their meaning may not always be transparent to a later interpreter. They thus knowingly face the possibility that a court's interpretation may deviate from the correct answer. Our initial model considers the set of cases in which the parties' payoffs under their contract are monotonic and continuous in the space of possible judicial interpretations. A party's payoff, that is, increases as interpretations of the contract become more favorable to it and decreases as interpretations become less favorable. In this set of cases, firms prefer courts to make interpretations on the minimum evidentiary base $B_{\min}$ except in unusual circumstances. We then consider a set of cases in which payoffs are

62. In a well-known case, Chief Justice Traynor stated:

> The fact that the terms of an instrument appear clear to a judge does not preclude the possibility that the parties chose the language of the instrument to express different terms. That possibility . . . exists whenever the parties' understanding of the words used may have differed from the judge's understanding. Accordingly, rational interpretation requires at least a preliminary consideration of all credible evidence offered to prove the intention of the parties.

Pac. Gas & Elec. Co. v. G.W. Thomas Drayage & Rigging Co., 442 P.2d 641, 645 (Cal. 1968).

invariant to the extent of judicial error. Here, every unfavorable interpretation that deviates from the correct answer has the same negative consequences for a firm, regardless of whether the interpretation is close to the correct answer or far from it. In this set of cases, firms have reasons to prefer a broader evidentiary base than $B_{\min}$, but, we argue, litigation cost considerations will still incline a majority of these firms to prefer courts to use a narrow evidentiary base. Taken together, the two models suggest that when the question of contractual language has been settled but the issue regarding what that language means remains, the default interpretive style for courts should be textualist.

In both models, parties negotiate a contract whose object is to maximize the surplus the deal could create, and they divide that surplus through the price term. The parties want a court, should a dispute arise, to give the correct answer to an interpretive question. Anticipating that the answer will be correct permits parties to take surplus-maximizing actions. Parties sign the contract when it is optimally clear, in a sense to be made more precise below.

1. *The Continuous-Payoff Case*

In this case, we focus for convenience on a buyer. We denote the surplus under a deal as s^*; the buyer's share is the difference between its valuation v^* and the price: $s_b^* = v^* - p$. The judicial interpretation that gives the correct answer is denoted i^*. Thus, if the court does make the interpretation i^*, the contractual surplus will be s^* and the buyer will realize s_b^*. The illustrative contract we consider requires the seller to prepare machines prior to delivery so as to minimize the buyer's costs of adjusting the machines for their intended use. Subsequently, the parties disagree over whether the seller fully complied with its duty to prepare the machines. If the court imposes a lesser preparation obligation on the seller than the obligation that the contract, correctly interpreted, would require, the deal will be less profitable to the buyer, and its gain will fall to $s_b^{(-)} < s_b^*$. If the court imposes a greater obligation on the seller than i^*, the buyer's gain will rise to $s_b^{(+)} > s_b^*$. Each interpretation in the space of possible judicial interpretations thus generates a particular payoff for the buyer. In this product preparation example, the buyer's possible payoffs are monotonic and continuous in the space of possible interpretations; that is, as interpretations become more favorable to the buyer, its payoffs increase, and vice versa. The seller's possible payoffs also are monotonic and continuous, but they decline as the buyer's increase.

When analyzing the parties' preferences regarding interpretive styles, we begin at the litigation stage and make three assumptions: that (1) a court

that relies solely on the minimum evidentiary base $B_{\min}$[63] will find the correct answer i^* with positive probability, (2) that the likelihood that a court will find i^* sometimes can be increased if the court considers evidence in categories additional to $B_{\min}$, and (3) that courts are unbiased. We initially support the first assumption with the following reasoning. Since the written contract trumps evidence in the other evidentiary categories,[64] and disputes are expensive, parties benefit from producing a writing that makes clear to a later court what was promised. Therefore, it is plausible to suppose that courts sometimes can discern the parties' exact intentions from the evidentiary base $B_{\min}$ alone. For the second assumption, assume that the parties specified the seller's preparation obligations in considerable detail in a preagreement memorandum that both parties signed. Admitting the memorandum into evidence will increase the likelihood that a court would give the correct answer to interpretation questions about product preparation. Regarding the third assumption, there is no reason to believe that courts will systematically deviate from the correct answer i^* in ways that are more or less favorable to particular parties or classes of firms.

These three assumptions permit us to state formally a party's expectation at contracting time regarding the possible payoffs that a judicial interpretation could induce. If courts are unbiased and can find i^* on the base $B_{\min}$, but can also err, then the buyer's expected payoff in our illustration, given what a court later will do, can be expressed as:

$$E[s_b(i) \mid B_{\min}] = s_b(i^*) + \varepsilon.$$

The left-hand side of this expression is the buyer's expected payoff given a judicial interpretation i made on the evidentiary base $B_{\min}$. The first term on the right-hand side is the buyer's payoff given the correct interpretation i^*; the second term is an error term with mean zero and positive variance. That ε has mean zero means that the court is unbiased. In other words, the court is as likely to make an interpretation that is more favorable to the buyer (less favorable to the seller) than the correct answer as the court is likely to make a less favorable interpretation. Judicial errors therefore cancel, in expectation. The error term has positive variance, however, because in an actual case a court's interpretation can deviate from the correct answer.[65]

63. Recall that $B_{\min}$ is composed of the written contract, a performance narrative, a dictionary, and the interpreter's experience and understanding of the world.

64. Section 1-303(e)(1) (formerly section 1-205(4)) of the UCC provides that when there is conflict, "express terms prevail over course of performance, course of dealing and usage of trade." U.C.C. § 1-303(e)(1) (2003).

65. For readers unversed in statistics, variance is a measure of how far an outcome can deviate from the mean of a distribution. Thus, if a court is unbiased but the variance is large, the interpretation that the court makes in an actual case may well be very far from the correct answer; conversely, when the variance is small, the court is likely to be close to the correct answer. If the

The expression thus says (in English) that when a court restricts itself to the evidentiary base B_{min}, the buyer's expected payoff equals the payoff that the buyer expected to get when it agreed to the contract.

When the variance in the error term is large, the parties know that a court's answer is likely to deviate widely from the correct answer. Our example of the preagreement memorandum regarding the seller's performance obligation shows that adding evidence to B_{min} can shrink the variance. Simply put, if the parties here know that the evidentiary base in a later lawsuit will include preagreement memoranda, they will expect later interpretations to be close to the correct answer.

This specification of the relation between judicial error and party expectations implies that firms commonly prefer courts to be restricted to the minimum evidentiary base B_{min} when payoffs are continuous in the space of possible interpretations. To see why, recall our second assumption: As the evidentiary base approaches B_{max}, the variance in the error term approaches zero. A risk-neutral party cares about the mean of the interpretation distribution but not the variance. This is because the variance term measures risk while risk-neutral parties are indifferent to risk. Therefore, it is enough for a risk-neutral firm that the expected interpretation $E(i)$ equals the correct interpretation i^*. Put another way, a firm's preference at contract time is to have courts make interpretations on the minimum evidentiary base unless it would be costless to widen the base. But it is not costless. As the permissible evidentiary base widens, each party has incentives to introduce more evidence and, in turn, will need to contest more evidence. Since trials are expensive, risk-neutral firms are Willistonians.[66]

This view should not be overstated. Some firms will have a different preference. Recall the volatile markets example in Section III.C above. There, we argued that firms functioning in volatile markets would make spot purchases unless a particular transaction was important relative to the size of the firm. In that event, a firm often would prefer to incur the costs of making a contract in order to avoid the costs of disruption that a bad market realization could cause. A similar preference could obtain here. Thus, when performance of a particular contract is important to the survival of the firm—say, a contract with a major supplier—or when the contract is new and is expected to be widely used, the firm may be unwilling to risk a seriously adverse interpretation. If contextualists are correct that larger

contextualist claim is correct that broader evidentiary bases generate more accurate interpretations (i.e., the variance shrinks more), it follows that litigation where the court sees only B_{min} is more risky for parties than litigation where the court sees additional evidence; when the court sees only B_{min}, its interpretation is more likely to deviate further from the correct answer.

66. There is considerable evidence that firms prefer a formalist adjudicatory style. *See* Bernstein, *Merchant Law*, *supra* note 31; Bernstein, *Private Commercial Law*, *supra* note 31, at 1735-37.

evidentiary bases do shrink variance, then parties concerned with variance will likely prefer that courts use a contextualist adjudicatory style. In the example above, if the machines were crucial to a start-up venture and the buyer had little cash, the buyer ex ante would want the evidentiary base in a later suit to be sufficiently broad so that the memorandum could be admitted into evidence. On the other hand, only unusual contracts have this "bet the ranch" quality. In the typical case, it is good enough that courts get things right on average.

We now return to the contract-creation stage to consider a deeper justification for our first assumption. It is optimal for risk-neutral firms to invest resources in drafting until the writing is sufficiently clear, in an objective sense, so that the mean of the distribution of possible judicial interpretations is the correct interpretation i^*. Contracts sketched out in less detail than this would generate interpretation distributions whose mean could be anywhere. As a consequence, parties could not expect courts to protect their expectation interest in case of a dispute. As this would be inefficient, firms will attempt to write contracts with sufficient clarity to permit courts to find correct answers, though with error.

The current debate between textualists and contextualists is irrelevant to issues of interpretation for cases that resemble the performance-preparation illustration. In that illustration, there was consensus regarding the language in which the contract was written, but that language was arguably unclear. The contextualist in effect asserts that a larger evidentiary base shrinks variance. Indeed, if the base is large enough, the variance approaches zero, so parties will believe when they contract that a court will find the correct answer with certainty. As applied to our illustration, taking into account the memorandum, perhaps together with an industry custom as to a seller's preparation obligations, would leave little doubt regarding what the seller was supposed to do. The textualist, in contrast, claims that variance does not shrink materially with a broader evidentiary base because contracts often have plain meanings. Hence, permitting parties to introduce additional evidence as to intent would generate costs in excess of gains. The proponents in this debate thus disagree on the relation between the width of the evidentiary base and the accuracy of a court's interpretation. But it is unnecessary for courts to decide which side of the textualist-contextualist debate is correct. Greater accuracy is lower variance, and business parties commonly are indifferent to variance. Thus, courts that interpret contracts as typical parties prefer would be indifferent to variance as well, and sensitive only to the costs of administering their evidentiary standard. These courts would thus make interpretations on the evidentiary base $B_{\min}$ unless parties instructed them to the contrary.

2. *The Discontinuous-Payoff Case*

We illustrate this case by recalling the specialized-product example in Section III.B. We denote the quality that the contract required the seller to produce as q_k. The seller produces q_k but the buyer, possibly sincerely, claims that the contract required a higher quality level. The product's actual quality is not in dispute; the issue is what the contract said. If a court decides that the contract required a lower quality level than what was actually delivered, we denote the interpretation as $q^{(-)} < q_k$. If the court decides that the contract required a higher quality level than what the seller delivered, we denote the interpretation as $q^{(+)} > q_k$. Just as before, we assume that the court is unbiased. Therefore, the mean of this three-point distribution is q_k: The seller's expectation regarding a later judicial interpretation of the contract can be written as $E(i) = q_k$.

The seller's payoffs here, however, are not continuous in the space of the court's possible interpretations, because of the perfect tender rule.[67] If a court finds that the contract required a quality level that was less than or equal to the quality the seller delivered, the seller can recover the price; if the court finds that the contract required a higher quality level, the buyer can reject and credibly threaten a suit for damages. The seller then will realize a renegotiation payoff that likely is less than the price because the damage threat reduces its bargaining power. The fact that judicial interpretations are unbiased does not matter for this seller, because the seller does not gain from judicial errors on the low side of q (it cannot recover an amount greater than the price if it has delivered a quality level that is higher than the level the court requires), while the seller is harmed by judicial errors on the high side of q, however small (the buyer can reject if the court interprets the contract to require even a slightly higher quality than the contract actually required). The seller thus wants the court to find the correct answer with certainty, even though the seller is risk-neutral.

The seller's preference has an efficiency implication. In the continuous-payoff case, the possibility of judicial error does not create an ex ante inefficiency: There, when the court is right on average, a party's expected payoff under the contract equals its expectation interest. In the discontinuous-payoff case, we will see, when a court is only right on average, the seller's expected payoff under the contract is less than its expectation interest. An efficiency-minded court thus could prefer to widen the evidentiary base in order to shrink the possibility of error, and thereby better protect a party's expectation.

67. *See* U.C.C. § 2-601 ("[I]f the goods or the tender of delivery fail in any respect to conform to the contract, the buyer may . . . reject the whole . . . .").

To see how the possibility of error in the discontinuous-payoff case can create an inefficiency, and to see why we nevertheless resist the view that the default for this case should be a wider evidentiary base, we generalize the formalization used above. We now suppose that the parties face two possibilities regarding a court's interpretation of their agreement. First, there is a positive probability that a court may find the correct answer with certainty on the evidentiary base $B_{\min}$. For example, the interpreting judge may be commercially sophisticated or the breach narrative may be quite revealing. Second, there is a positive probability a court may err. When error of some kind is possible, the parties must specify an "error distribution": how errors in particular cases may be manifested. As in the initial model, we let parties assume that courts that can err will be unbiased; in expectation, that is, courts are assumed to be correct on average, but parties know that any actual interpretation may be off the mark. We formalize this scenario by letting β be the probability that an interpreting court will find the correct answer with certainty, and letting $(1 - \beta)$ be the probability that the court will make only an unbiased interpretation of the contract terms regulating quality. To simplify this presentation, we continue with our assumption that an unbiased court will interpret the contract either correctly, or to require a higher quality level than the parties actually intended, or to require a lower quality level than was intended. Each of these possibilities is assumed to occur with the same probability, that is, one-third. We complete this story by denoting the contract price—the seller's expectation—as k and the seller's payoff in a possible renegotiation as $r \leq k$. This permits us to write the seller's expected payoff under the contract as of the time the parties make the deal:

$$E(s_s) = \beta k + (1 - \beta)(\tfrac{2}{3}\, k + \tfrac{1}{3}\, r).$$

With probability β, the seller receives the payoff that the contract directs, which is the price k, because the court will interpret the contract correctly. With probability $(1 - \beta)$, the court will make an unbiased interpretation of the quality level the contract required. Hence, when the seller is found to have complied or overcomplied, which in total occurs two-thirds of the time, the seller receives the contractual payoff of k, and when the seller is found to have undercomplied, which occurs one-third of the time, the seller receives the renegotiation payoff of r.[68]

68. To see why the new formalization generalizes the original treatment, we can rewrite the initial model so that courts will sometimes find the correct answer with certainty. The buyer's expected payoff in the continuous-payoff case then is expressed as

$$E(s_b \mid B_{\min}) = \beta s_b(i^*) + (1 - \beta)[s_b(i^*) + \varepsilon] = s_b(i^*)$$

because ε has mean zero: The buyer's expected payoff equals in expectation the correct payoff. When this more general formalization is used, the second term in brackets will equal the expectation interest in the continuous-payoff case, but this term can be less than the expectation in

The seller's expected payoff under the contract—$E(s_s)$—ordinarily will be lower than the seller's expectation interest, which is the contract price of k. This is because β commonly is less than one (the court is expected sometimes not to find the correct answer with certainty) and r commonly is less than k (the seller expects to receive less than the contract price in a renegotiation). Since protecting the expectation interest is efficient, the parties can increase surplus by raising the seller's expected payoff. Our assumptions regarding the relationship between the width of the evidentiary base and the accuracy of a court's interpretation imply that the probability β of a court's finding the correct answer rises as the evidentiary base expands beyond $B_{\min}$. Because the seller's expected payoff under the contract increases as β increases—as the court becomes more accurate—the parties, though risk-neutral, have reason to prefer a broader evidentiary base than $B_{\min}$.[69] Expanding the evidentiary base is not costless, however. The parties, therefore, face a tradeoff between the efficiency of increased accuracy and the inefficiency of increased contract-enforcement costs.

Two factors influence how parties will make this tradeoff. The first factor is the bargaining power the seller will have in a renegotiation; the greater the seller's bargaining power, the more closely the renegotiation payoff r will approach the contractual payoff k, and the closer the seller's expected return will be to the contract price. Hence, when the seller expects to have bargaining power, the parties ex ante are less likely to prefer the possibility of extensive discovery and trials conducted just in order to increase β—the likelihood that the court will make a correct interpretation. Recall that bargaining power is a function of the parties' relative discount rates, which commonly are the same, and the parties' disagreement points. The seller's disagreement point is largely a function of how redeployable its investment is. The seller's bargaining power is minimized when, as in the illustration in Section III.B, its investment is not redeployable at all because then its disagreement payoff will be negative.

In making the tradeoff between accuracy and cost, parties will also consider a second factor: how much a given piece or category of evidence will increase the likelihood that a court will make a correct interpretation. In the example above, the introduction of a single piece of evidence—the preagreement memorandum—was assumed to increase materially the probability that the court would find the correct answer. The memorandum thus was very productive, generating a large increase in accuracy at a low

the discontinuous-payoff case, as it is in the textual treatment above. This is because when courts err in the latter case, the seller is exposed to the possible downside of receiving less than its expectation in a renegotiation, without any concomitant upside.

69. In addition to expanding the evidentiary base, the parties also could increase surplus by changing the contract's payoff structure. *See infra* text accompanying notes 72-77. Since the buyer would benefit through the price term from actions that increased surplus, the buyer would agree to either method when the method would be cost-justified.

cost. When evidence is expected to be less productive, parties will be less inclined to have courts make interpretations on a broad evidentiary base.

This analysis of the relevant tradeoff suggests that parties in the discontinuous-payoff case would more commonly prefer a narrow evidentiary base for interpretations. Many more deals are for relatively standard goods that can be improved by a party's specialized investment than for goods that would be worthless in uses other than those the parties contemplate. Thus, parties such as the seller here ordinarily could redeploy a significant fraction of their investment, and so would not be seriously disadvantaged in a renegotiation. Further, the typical choice that parties expect later to face in a contract action is not whether a particular piece of evidence will be admitted or not; the choice is whether a court will make interpretations on a motion for summary judgment—i.e., on the base $B_{\min}$—or after a trial, often before a jury.[70] Trials can be very costly. Thus, parties would prefer broad evidentiary bases either when their performance would be highly specialized or when an evidentiary category in addition to $B_{\min}$ would be very productive—a clear custom, for example.[71] Since these illustrations appear to capture only a minority of cases, we suggest that the majoritarian preference in the discontinuous-payoff case also is for courts to use narrow evidentiary bases when making interpretations. Turning again to the contract-creation stage, typical parties would invest sufficient resources in drafting just to create the evidentiary base $B_{\min}$ that would permit a court to make an unbiased interpretation.

In addition, parties have the ability to protect the seller's expectation in the discontinuous-payoff case without expanding the evidentiary base. Their method is to change the contract's payoff structure. The structure of the discontinuous-payoff case creates the possibility that a party will incur a large loss because of a small judicial error. The seller in the example here thus could suffer substantially if the court held that it to had to produce only

70. Courts frequently consider extrinsic evidence when interpreting the written contract. A commentator recently explained:

> Extrinsic evidence includes both evidence about trade customs and evidence about interchanges between the parties—concerning the course of performance of the current contract, the course of dealing in prior transactions, or the bargaining history of the current contract. Very often this extrinsic evidence will not be solely documentary and will require evaluation of oral testimony about conversations between the parties.

William C. Whitford, *The Role of the Jury (and the Fact/Law Distinction) in the Interpretation of Written Contracts*, 2001 WIS. L. REV. 931, 937. Over 50% of contract cases tried in federal courts are to juries, and between 25% and 30% are tried to juries in state courts. *See* Marc Galanter, *Contract in Court; or Almost Everything You May or May Not Want To Know About Contract Litigation*, 2001 WIS. L. REV. 577, 591, 598 tbl.3, 602 tbl.5, 605 tbl.7, 625 tbl.15.

71. Richard Craswell shows, however, that courts evaluate the probativeness of customs by evaluating the purposes that the customs are meant to serve. *See* Richard Craswell, *Do Trade Customs Exist?*, *in* THE JURISPRUDENTIAL FOUNDATIONS OF CORPORATE AND COMMERCIAL LAW 118, 138-42 (Jody S. Kraus & Steven D. Walt eds., 2000) [hereinafter JURISPRUDENTIAL FOUNDATIONS]. This suggests that custom evidence is costly to admit and to contest.

a slightly higher level of quality than the correctly interpreted contract would have required. It is this "knife-edge" property of the perfect tender rule that reduces the seller's expected payoff below the price. Parties can respond to the need to increase the seller's expected payoff, however, by contracting out of the rule. The common way to do this is to use the customary warranty term, which contains a repair-and-replacement clause that eliminates the buyer's right to reject but requires the seller to repair or replace defective tenders or parts of tenders. When the buyer cannot reject, the seller is entitled to the price, with damages deducted from it.[72] The customary warranty thus eliminates the knife-edge feature of the perfect tender rule, thereby ensuring that both parties' expectations are protected so long as they believe that their payoffs will be determined by the correct interpretation of the repair-and-replacement clause.

This condition will be satisfied because the repair-and-replacement clause is analytically similar to the product-preparation term considered in Subsection IV.C.1. Under that term, the buyer's expected payoff, $E[s_b(i)]$, equaled the payoff it would receive under the correct interpretation, s_b*. To see why the repair-and-replacement clause has the same property, recall that section 2-719 of the UCC permits a seller to limit the buyer's remedies to repair and replacement,[73] while section 2-608(1) permits the buyer to "revoke his acceptance" only if a "non-conformity substantially impairs [the good's] value to him."[74] Section 2-719(2) authorizes a court to refuse to enforce a repair-and-replacement clause if "circumstances cause . . . [the] limited remedy to fail of its essential purpose . . . ."[75] Comment 1 to section 2-719 explains that this authorization exists because "there [must] be at least a fair quantum of remedy for breach of the obligations or duties outlined in the contract."[76]

In light of these sections, a court's interpretive task under a repair-and-replacement clause is not to decide, as per the perfect tender rule, if the goods "fail *in any respect* to conform to the contract."[77] Instead, the initial question for a court is whether the degree of success the seller achieved in the repair or replacement task was such as to give the buyer "a fair quantum of remedy." If not, the clause becomes inoperative and the question shifts to whether the goods are "substantially impaired" in value. The buyer's possible payoffs under a repair-and-replacement clause thus increase as the court requires the seller to exceed the correct interpretation of the clause and decrease as the seller is permitted to fall short, just as the payoffs in the

72. *See* U.C.C. § 2-717.
73. *See id*. § 2-719.
74. *Id*. § 2-608(1).
75. *Id*. § 2-719(2).
76. *Id*. § 2-719 cmt. 1.
77. *Id*. § 2-601 (emphasis added).

product-preparation case. And since the seller's payoffs here are a function of the buyer's payoffs, the seller's possible payoffs also are continuous in the same way. Both parties thus will expect the payoff under a repair-and-replacement clause to equal in expectation the payoff the parties intended. This suggests that parties choosing a repair-and-replacement warranty will want courts to interpret the warranty using the minimum interpretive base $B_{\min}$. This is because, as we have argued above, it is ordinarily enough for business parties that courts are correct in expectation, not that they are always correct.

3. *Summary*

The case in which the parties' payoffs are continuous in the space of a court's possible interpretations covers a lot of the ground. This is because the case often arises "naturally," as in the product-preparation example, and can also arise "artificially," as when parties contract to create continuous payoffs in order to increase the efficiency of their payoff structure. Firms in the continuous-payoff case ordinarily prefer courts to follow a textualist interpretive style. The case in which the parties' payoffs are invariant to the degree of judicial error probably is less common. Even in this case, however, litigation cost considerations will cause a majority of parties to prefer courts to use a narrow evidentiary base. Therefore, the best interpretive default for firms is textualist when the issue is what their contract language meant.

We conclude this Section with a comment concerning judicial practice. Courts making common law adjudications commonly take a Willistonian approach,[78] while the UCC strongly urges a contextualist interpretive style.[79] Courts in general, however, treat interpretation rules as mandatory. Judges are reluctant to invoke the coercive machinery of the state to require a party to perform a contract (or to pay damages) unless the judge is satisfied that the contract actually directed what the party failed to do. It seemingly follows that courts, not parties, should choose the rules that determine how contracts are read. This view is understandable but misguided. The law in general permits persons and firms to make choices in litigation that may lead courts to act on less than full information. For example, persons and firms may waive the right to counsel, agree to

78. *See* Peter Linzer, *The Comfort of Certainty: Plain Meaning and the Parol Evidence Rule*, 71 FORDHAM L. REV. 799, 805-06 (2002); Robert E. Scott, *The Uniformity Norm in Commercial Law: A Comparative Analysis of Common Law and Code Methodologies*, *in* JURISPRUDENTIAL FOUNDATIONS, *supra* note 71, at 149, 167-69.

79. *See* § 1-205 cmt. 1 ("This Act rejects both the 'lay-dictionary' and the 'conveyancer's' reading of a commercial agreement. Instead the meaning of the agreement of the parties is to be determined by the language used by them and by their action, read and interpreted in the light of commercial practices and other surrounding circumstances.").

stipulated findings of fact, and use summary arbitration procedures whose results courts are required to enforce. In sum, the law generally sacrifices accuracy in adjudication to parties' self-interested choices. Similarly, parties should be permitted to realize the cost savings from contract interpretations on minimal evidentiary bases even if, in any given case, the odds of an accurate interpretation would be higher with a broader base.[80]

D. *Private Languages, Linguistic Defaults, and the Parol Evidence Rule*

1. *The Preferred Linguistic Default*

We now consider the case in which the parties have written a complete contract in some language. The issue is whether, if the contract is silent on the matter, a court should take the parties to have written in majority talk. The alternative judicial assumption would hold that, in case of a dispute, the parties prefer to have the opportunity to introduce extrinsic evidence that relevant parts of the contract were written in the parties' private language. We next set out four reasons why the "linguistic default" should hold that the contract was written in majority talk. The practical implication of this proposal is that, when a contract does not speak to the issue, the court should not go beyond the evidentiary base $B_{\min}$ when attempting to identify the language of the contract.[81]

Many parties would prefer the linguistic default we propose because the default would (1) reduce contracting costs, (2) minimize the opportunities for strategic behavior, (3) reduce the risk of judicial error, and (4) expand the set of efficient contracts that parties could write. Let us begin with the issue of contracting costs. The comments to section 2-202 of the UCC

80. If parties preferred courts to use evidentiary bases that were so broad as to foreclose judicial time for deciding other categories of cases, then courts should override the parties' preference. Such a danger of conforming to parties' interpretive preferences would seem unlikely, however, given that modern-day parties generally seem to want adjudications to be briefer than they now are. As a result, so long as there are no other negative third-party effects from judicial deference to the parties regarding interpretive styles, deferring to parties' interpretive preferences would be efficient in the majority of modern contract cases.

81. The position that a party always should be able to show that its contract was written in party talk sometimes is justified on autonomy grounds. Thus, Chief Justice Traynor explained in *Pacific Gas & Electric Co. v. G.W. Thomas Drayage & Rigging Co.*: "In this state, . . . the intention of the parties as expressed in the contract is the source of contractual rights and duties. A court must ascertain and give effect to this intention by determining what the parties meant by the words they used." 442 P.2d 641, 644 (Cal. 1968). This argument is a non sequitur as stated. No one would claim that the Statute of Frauds interferes with party autonomy in a normatively serious way, though the Statute sometimes requires parties to cast their agreements in written form. *See* U.C.C. § 2-201. It would not be a serious interference with autonomy for courts also to assume that parties cast their agreements in majority talk if good reasons exist to create the incentives that this assumption implies. Thus, it takes more argument than now exists to show that an autonomy-based view of interpretation would justify the rule Chief Justice Traynor stated, even if such a view were appropriately applied to firms.

adopt a strong contextualist linguistic default. Comment 2 thus recites: "[W]ritings are to be read on the assumption that the course of prior dealings . . . and the usages of trade were taken for granted [that is, were meant to be aids in interpretation when the document was phrased] . . . . [*u*]*nless carefully negated* . . . ."[82] But if the default were reversed, parties could contract out of a plain-meaning linguistic default at the same or lower cost. For an example, consider: "This agreement is to be read in light of the customs of the widget trade." Parties seldom would describe the actual customs in the agreement because contracting costs are incurred today with certainty while dispute resolution costs are incurred tomorrow and probabilistically. Thus, it ordinarily will be cheaper just to tell courts to consider custom should an irreconcilable difference later arise.[83] In addition, a minority of contracts are written largely in private languages, and the parties that write them frequently enforce them with expert arbitrators, not lay judges. Thus, fewer parties would have to contract out of a default that supposed them to be writing in majority talk than would have to contract out of the contrary default.[84]

A plain-meaning linguistic default—that is, a default that restricts the court to the interpretive base $B_{\min}$—also would reduce strategic behavior. To see why, consider a contract between party *A* and party *B*, the relevant provisions of which were written in party talk. This contract, suppose, turns out badly for *B*. Therefore, *B* would like to raise an interpretation issue strategically, claiming that the contract was written in majority talk, in

82. § 2-202 cmt. 2 (emphasis added).

83. Some scholars argue that if courts are Willistonian, and thus implicitly adopt majority talk as the linguistic default, parties will incur additional drafting costs translating their private language into the majority language in order to make their intentions clear to judicial interpreters. *See* Jody S. Kraus & Steven D. Walt, *In Defense of the Incorporation Strategy*, *in* JURISPRUDENTIAL FOUNDATIONS, *supra* note 71, at 193, 197-200. This argument assumes that parties can only opt out of the plain-meaning linguistic default with costly translations. But as the discussion in the text has shown, parties can cheaply opt into a private language by agreeing in their contract that, should a dispute arise, evidence should be admitted regarding that language. Further, even when contracts contain technical party talk, most of their words will be written in the majority language. For example, parties may attribute a private meaning to the phrase "two-by-four" (wood supports so described in construction contracts are usually meant to state dimensions of 1⅝ by 3⅜ inches), but such parties seldom would use a private language to describe the delivery date, the place of delivery, the price, and so forth. These parties would want words with trade-language meanings to be read with trade understanding, but would not want words written in the majority language to be read as if they were special. Thus, if the linguistic default is party talk, then parties who wish to exclude party-talk interpretations of majority talk would be required to identify all of the majority terms and explicitly negate the use of extrinsic evidence for interpreting them. On the other hand, by opting into only the technical party talk they wish to incorporate, parties can more readily unbundle the two types of language—e.g., "All measures and specifications in this contract are to be interpreted in light of the customs of the construction industry." Hence, if courts are Willistonian just when parties want them to be, parties will not incur unnecessary writing costs.

84. The current interpretive rules are mandatory (or "quasi-mandatory"). For the purpose of our discussion here, which focuses on party preferences, it is clarifying to treat the rules as defaults.

order to improve *B*'s bargaining position. *B* could raise such an issue only if *B* could plausibly show that words in the parties' private language had a clear but different meaning in the majority linguistic community. *B would* raise such an issue only if the different plain meaning would also relieve *B* of performance. These two conditions for avoiding the effect of a private language are difficult to satisfy because while there are many private languages there is only one majority language. Hence, it would only be coincidence for words in a particular private language to have a clear but different meaning in the majority language that also favored a litigant such as *B*. Now assume that the contract is written in majority talk. The multiplicity of possible private languages would permit party *B* more easily to assert a helpful private meaning. A common move is to claim that stated prices or quantities are only "estimates" or "projections" in the private language the parties used.[85] The linguistic default we favor would reduce the likelihood that this move would succeed, for a party would have to decide before a dispute arose to write in a private language, and then propose to its contract partner that the contract be written and interpreted in that language. The partner would not agree unless such a language actually existed and had meanings that were accessible to courts. Thus, requiring parties ex ante to *say* they are writing in a private language would largely ameliorate the concern that a party would attempt to rescue itself from a bad deal by claiming that its contract was written in a mythical private language.

The two linguistic defaults at issue can now be reconsidered in light of this analysis. If a court always permitted parties to offer proof that they wrote in a private language, then a disappointed party would have a strong incentive to attribute a fictional favorable private meaning to a majority-talk contract. A default that supposed parties to be writing majority talk unless their contract recited otherwise would prevent such manipulation. And then if parties actually did write in a private language and contracted out of the Willistonian default, it would be difficult for one of them later to behave strategically regarding interpretation because, as we have just shown, the conditions for avoiding particular private languages are so stringent. The linguistic default that supposes parties to be writing majority talk thus would reduce strategic behavior more effectively than the rules in the comments to today's UCC.

85. In one case, a contract required the buyer to take "'approximately 70,000 cubic yards' of concrete" and also recited that "'[n]o conditions which are not incorporated in this contract will be recognized.'" S. Concrete Servs., Inc. v. Mableton Contractors, Inc., 407 F. Supp. 581, 582 (N.D. Ga. 1975). The buyer took a little over 12,500 yards in a falling market. *Id*. In the litigation, his claim that parties in the trade understood explicitly specified quantities to be estimates was rejected. *See id*. at 585-86. On the other hand, in *Columbia Nitrogen Corp. v. Royster Co*., 451 F.2d 3 (4th Cir. 1971), course-of-dealing and usage-of-trade evidence was admitted to demonstrate that express price and quantity terms were only "fair estimate[s]." *Id*. at 7 n.3.

A plain-meaning linguistic default that assumes parties to be writing in majority talk also would reduce the risk of judicial error. There is one majority linguistic community but many possible private communities. Hence, when private languages are permissible, as they often are today, a court has two interpretive tasks: to ascertain the parties' language and to ascertain what the parties said in that language. The danger that a court will pick the wrong language is real because a private linguistic community can be as small as the particular parties to the contract. Hence, a disappointed party may plausibly claim that the parties' course of dealing or their oral negotiations showed that, in the parties' language, "all" meant "some,"[86] that agreeing to take a "minimum" quantity meant that the buyer could take much less,[87] or that an unconditional option was really conditional.[88] When such a claim is false but found to be true, the court necessarily will misinterpret the contract.

To see why, recall our argument in Section IV.C that firms are content to have courts be right on average, not right every time. The error here is different from the error described there. For example, assume a contract uses the word "red," and a disappointed party persuades the court, wrongly, that the contract was written in a private language in which the word "red" meant "green." Both red and green are vague. In this example, the space of possible judicial interpretations would center around some instance of the concept "green," but the court here could not be right on average. It would be attempting to find the correct shade of green while the parties, ex ante, wanted a court to find the correct shade of red. When courts are mistaken regarding the contract's language, their constructions must be inefficient because it is only efficient to protect a party's expectation interest under the contract it actually wrote. Hence, parties face a heightened risk of inefficient interpretations when courts always entertain claims that a contract was written in a private language. Typical parties thus would prefer courts to assume that they wrote in majority talk.

Finally, contextualist interpretation in general, and especially contextual interpretation that permits the use of private languages, can truncate the set of efficient contracts that parties will write. We have just argued that contextualist interpretation can create moral hazard. It is plausible to believe that the more complex the contract, the easier it will be to create disputes regarding what the contract says and what language it was written in. In addition, litigation is more costly in a contextualist interpretive regime because the parties more frequently will have full trials. The fact that contextualist regimes increase the likelihood and cost of

86. *See Pac. Gas & Elec.*, 442 P.2d 641.

87. *See Columbia Nitrogen Corp.*, 451 F.2d 3.

88. *See* Hunt Foods & Indus., Inc. v. Doliner, 270 N.Y.S.2d 937, *aff'd*, 272 N.Y.S.2d 686 (App. Div. 1966).

disputes creates an incentive for parties to use simpler, but possibly less efficient, contracts.[89]

An example may make this point clear. Assume that parties can create a complex deal that will generate a net expected surplus of $15 at a cost of $5 to write the contract. The parties also can create a simpler deal that will generate a net expected surplus of $8 at a contract-writing cost of $1. The parties should write the complex contract because the net expected social surplus of $10 exceeds the expected social surplus of $7 that the simple contract would create. Suppose, however, that there is a .3 likelihood of an interpretive dispute when parties use a complex contract in a contextualist interpretive regime and no likelihood of a dispute when parties use a simple contract (formally, one needs only a higher likelihood of a dispute when the contract is complex). If there is an interpretive dispute, parties, as in the continuous-payoff case described above, expect the court to interpret the contract correctly. The correct interpretation would permit the parties to realize (and share) the surplus that the contract was written to create. In this illustration, the parties will use the simple contract if the cost of a trial would exceed $10.[90] Since trial costs are said often to approach the amount at stake in disputes, it is not fanciful to suggest that a contextualist interpretive regime will drive parties toward simple contracts. To complete the example, suppose that the cost of a trial would be $12, so that parties actually would use the simple contract. Now suppose that the state switches to a textualist interpretive regime. For convenience, let the probability of a dispute be the same in both regimes, but now assume that the typical interpretive dispute will be resolved by summary judgment at one-half the cost of a full trial. Then parties would use the complex contract and maximize social surplus.[91]

When parties consider what type of contract to write—what type of deal to create—they will not only consider the costs and gains from creating the deal initially; they will also consider the likelihood and costs of later disputes. In the example here, when the cost of disputes is added to the calculation, parties in a contextualist interpretive regime will use a simple but less efficient contract rather than a complex but more efficient contract. To be sure, an example cannot show that contextualist interpretive regimes

89. A formal treatment of the relation between contracting and enforcement costs and the parties' choice of contractual form will appear in Alan Schwartz & Joel Watson, *The Law and Economics of Costly Contracting*, 20 J.L. ECON. & ORG. (forthcoming Apr. 2004).

90. The expected return from a complex contract given the likelihood of dispute is $7(\$15) + .3(\$15 - t) - \$5$, where the first term is the expected gain when there is no dispute, the second term is the expected gain from a trial (t is the trial cost), and the third term is the cost of creating the deal. The net expected return from the simple contract is $7. This will exceed the return from the complex contract when t exceeds $10.

91. In this example, when t = $12, ½ t equals $6, and the expected return from using the complex contract rises to $8.20, which exceeds the gain of $7 from the simple contract.

always induce parties to use less efficient contracts. The lesson rather is that an overlooked cost of these regimes is that they sometimes buy greater accuracy at a cost of less efficient contracting. This disadvantage reinforces our view that business parties generally prefer a plain-meaning linguistic default, which implies the use of a Willistonian interpretive style.

The analysis here permits us further to clarify the debate between contextualists and textualists and also suggests a new understanding of the function of merger clauses. Contextualists claim that interpretations made on broad evidentiary bases are more likely to be correct than interpretations made on narrow bases. Textualists dispute this claim. The dispute between these camps, we have shown above, is irrelevant when the issue is what the contract says. Business firms are content with interpretations of their language that are correct on average, not always correct, and so prefer narrow evidentiary bases to broad ones. The interpretation dispute matters a great deal, however, when the interpretive issue concerns the language in which the parties wrote. In such a case, we have argued, a broad evidentiary base affords a disappointed party the opportunity to raise the language issue strategically. Broad evidentiary bases also increase the risk of judicial error and truncate the set of efficient contracts parties can write. Our analysis thus supports the conclusion that courts should interpret business contracts on minimal evidentiary bases whether the issue is what the contract language means or what language the contract was written in, unless parties explicitly instruct the court otherwise.[92]

Merger clauses are understood to restrict the evidentiary base available to courts when making interpretations.[93] Because courts today often search broadly for the correct answer, merger clauses are difficult for parties to enforce.[94] These clauses now can be seen to have an additional function. A

92. This conclusion may be thought to raise a regress problem. If parties can give interpretive instructions, then those instructions will themselves have to be interpreted, as would instructions regarding how to interpret the instructions and so forth. This problem does not appear to be serious. The default we advocate would require parties to tell courts to widen the evidentiary base. For example, the contract would recite: "Use custom." There seems little reason for parties to say "Use custom sympathetically." Courts seldom would need instructions as to how to interpret simple directives that refer the courts to evidentiary categories now in use.

93. A merger clause recites that the written agreement is the parties' final expression of their intentions. A common example states: "This contract contains the final understanding between the parties and represents the final agreement on all terms. There are no verbal agreements or representations in connection therewith. The writing is a merger of all proposals, negotiations, and representations with reference to the subject matter and provisions." For examples of similar clauses, see Luther Williams, Jr., Inc. v. Johnson, 229 A.2d 163, 165 (D.C. 1967); and UAW-GM Human Res. Ctr. v. KSL Recreation Corp., 579 N.W.2d 411, 412 (Mich. Ct. App. 1998).

94. *See* Franklin v. White, 493 N.E.2d 161, 166 (Ind. 1986) ("An integration clause is only some evidence of the parties' intentions. The trial court should consider an integration clause along with all other relevant evidence on the question of integration."); Sutton v. Stacey's Food Mart, Inc., 431 A.2d 1319, 1322 n.3 (Me. 1981) ("A merger clause does not control the question of whether a writing was intended to be a completely integrated agreement." (citing RESTATEMENT (SECOND) OF CONTRACTS § 242 cmt. e (Tentative Draft Nos. 1-7, 1973))); *see also* RESTATEMENT (SECOND) OF CONTRACTS § 209 cmt. b (1981) ("Written contracts . . . may

merger clause, if honored, would limit the court to the evidentiary base B_{min}. Parties aware that the base was so limited would have a strong incentive to write the contract in majority talk. By adopting the merger clause, therefore, the parties signal to the court that this incentive has motivated them to speak in majority talk. Thus, a court, even if not persuaded by our more general argument that majority talk should be the default language, should still interpret the inclusion of a merger clause to mean that the parties wrote in majority talk.[95]

2. *The Parol Evidence Rule*

A typical statement of the parol evidence rule provides that when terms are "set forth in a writing intended by the parties as a final expression of their agreement," the terms "may not be contradicted by evidence of any prior agreement or of a contemporaneous oral agreement but may be explained or supplemented by course of performance, course of dealing, or usage of trade."[96] This rule comes in a "hard" and a "soft" version. Courts that adopt the hard version of the parol evidence rule decide whether a written contract is ambiguous from the document itself; that is, the court

include an explicit declaration that there are no other agreements between the parties, but such a declaration may not be conclusive."). Professor Corbin is regarded as the most persuasive advocate of this position: "[I]t can never be determined by mere interpretation of the words of a writing whether it is an integration of anything, whether it is the final and complete expression of the agreement or is a mere partial expression of the agreement." 3 ARTHUR LINTON CORBIN, CONTRACTS § 581, at 442 (1960) (internal quotation marks omitted).

95. We argue that courts should adopt a Willistonian linguistic default because parties prefer it and because there are "external" efficiency reasons to justify the default. First, the state subsidizes the judicial system and thus has an independent reason to reduce the likelihood of disputes. For the reasons given above, our default would reduce this likelihood further than current law. Second, if contracts are written in majority talk, courts can create standard vocabularies in which commercial transactions can be conducted. When a phrase has a set, easily discoverable meaning, parties who use it will know what the phrase requires of them and what courts will say the phrase requires. Courts that insulate the meaning of terms in the majority language from deviant interpretations by interpreting the same words in the same ways across cases thus create a collective good: a set of terms with meanings that are already understood by a large majority of potential contracting parties. It follows that courts should encourage parties to use majority talk. *See* Alan Schwartz, *Contract Theory and Theories of Contract Regulation*, 92 REVUE D'ÉCONOMIE INDUSTRIELLE 101, 102-03 (2000); Robert E. Scott, *The Case for Formalism in Relational Contract*, 94 NW. U. L. REV. 847, 853-56 (2000); Scott, *supra* note 78, at 157. Another way to put this second point is that similar contracting problems recur in varying contexts. Efficiency is enhanced if parties can adopt prior solutions, and adoption is facilitated when parties know that if they use the language in which a prior solution was cast, they will be taken by courts and contract partners to have adopted that solution. Henry Smith has shown that courts are more hospitable to acontextual interpretation when they recognize that contract language is intended for third parties, as with deeds or third-party beneficiary contracts. *See* Henry E. Smith, *The Language of Property: Form, Context, and Audience*, 55 STAN. L. REV. 1105, 1177-90 (2003). Courts should recognize that the set of third parties who would benefit from acontextual interpretation is wider than is conventionally believed. This set includes parties who could profitably use the contracting solutions that prior parties had developed.

96. U.C.C. § 2-202 (2003).

makes an interpretation on the evidentiary base $B_{\min}$. Courts that adopt the soft version hear all extrinsic evidence before deciding whether there is an ambiguity.[97] The UCC is taken to have adopted the soft version, and a number of common law courts have adopted this version as well.[98] As should be apparent, this version of the parol evidence rule is justifiable if courts should consider all evidence that may bear on what the parties meant; it is not justifiable if courts should consider only the evidence that parties, ex ante, want courts to see.

Contrary to the conventional understanding, however, the text of the UCC's version of the parol evidence rule actually adopts the linguistic default for which we argue. Section 2-202 provides that extrinsic evidence can "explain or supplement" a writing but cannot "contradict" the writing. This language, when given its obvious meaning, creates a strong incentive for parties to write in majority talk. To see why, suppose that a contractual phrase has the plain meaning m_M in majority talk and the equally plain but different meaning m_P in party talk. If a party can introduce extrinsic evidence explaining that the contract actually had the meaning m_P, then extrinsic evidence never could contradict the writing. The meaning m_P would not contradict the meaning m_M, because the writing, properly understood, never meant m_M. Put more vividly, if a party could introduce extrinsic evidence to show that the parties meant "green" when they wrote "red," extrinsic evidence could "explain" but never "contradict" their contract. A contradiction could arise only if the contract *were* assumed to be written in majority talk and the extrinsic evidence was offered to show a meaning in that language other than m_M. Therefore, unless the word "contradict" is to be read out of the statute, section 2-202 must be taken to presuppose that contracts are written in majority talk, but to permit the introduction of extrinsic evidence to clarify ambiguities in this language.[99]

97. *See* Eric Posner, *The Parol Evidence Rule, the Plain Meaning Rule, and the Principles of Contractual Interpretation*, 146 U. PA. L. REV. 533, 534 (1998); Whitford, *supra* note 70, at 939.

98. *See, e.g.*, Ben-Shahar, *supra* note 52, at 790; Snyder, *supra* note 52, at 624-25.

99. This proposed interpretation of section 2-202 actually was made in *Southern Concrete Services, Inc. v. Mableton Contractors, Inc.*, 407 F. Supp. 581 (N.D. Ga. 1975). The contract there required the defendant to take "'approximately 70,000 cubic yards' of concrete." *Id.* at 582. The defendant purchased 12,542 cubic yards. It defended the subsequent lawsuit with the claim that trade custom and supplementary agreements between the parties would show "that the quantity stipulated in the contract was not mandatory . . . and that both quantity and price were understood to be subject to renegotiation." *Id.* The court excluded the evidence because it believed that an explicit quantity requirement would be contradicted by an understanding that the requirement was not mandatory. It explained: "To admit evidence of an agreement which would *contradict* the express terms of the contract would clearly eviscerate the purpose of § 2-202." *Id.* at 585.

3. *Course-of-Performance Evidence*

The UCC explicitly invites courts to consider the parties' course of performance under a contract.[100] Such evidence is said to be "always relevant" to the contract's meaning.[101] If the state is to provide the interpretive theory that the parties want, however, then supplementing contracts with course-of-performance evidence would frequently be a mistake. The parties' course of performance under a contract differs from evidence in the other evidentiary categories because it can be offered not only to show what the parties originally meant, but also to prove that the parties' meaning has changed. Thus, evidence that the buyer accepted shipments at quality levels below those specified in the contract may show that the parties modified the contract's quality requirement. Admitting course-of-performance evidence to prove a change in meaning is consistent with the traditional contract law rule that the parties' agreement may be inferred from acts or silence.[102] Nevertheless, courts should be reluctant to admit act-or-acquiescence evidence to show a change in the meaning of a written contract. The existence of a writing indicates that the parties once believed that the gains from writing things down exceeded the costs. In the absence of evidence that this calculus has changed, the best inference to be drawn is that parties want modifications to be written as well. If this inference is not compelling, it should become so when the contract contains a term requiring modifications to be in writing. Courts, however, accepting the UCC's invitation, often hold that conduct can effect a waiver of the "no oral modification" clause.[103]

The Code and the courts' use of course-of-performance evidence to establish a change in meaning reflects a misunderstanding of the parties' likely intentions. To see why, consider the quality-level illustration in the preceding paragraph. Suppose that the contract price was $2000 per lot

100. § 2-202(a).

101. U.C.C. § 2-208 cmt. 2 (1989) (providing that "a course of performance is always relevant to determine the meaning of the agreement"). Revised Article 1, approved in 2001, substantially preserves the section 2-208 definition of "course of performance" in new section 1-303(a) but deletes the comments to former section 2-208. *See* U.C.C. §§ 1-303, 2-208 (2003). To date, revised Article 1 has been adopted only in a few states. Section 2-208 and its comments thus remain as part of the enacted UCC in the large majority of jurisdictions.

102. *See, e.g.*, U.C.C. § 2-208 cmt. 1 (1989) ("The parties themselves know best what they have meant by their words of agreement and their action under that agreement is the best indication of what that meaning was."). On the current legal status of section 2-208, see *supra* note 101.

103. *See* U.C.C. § 2-209(4) (2003) ("Although an attempt at modification or rescission does not satisfy the requirements of subsection (2) [excluding modifications except by a signed writing] . . . it can operate as a waiver."). Comment 4 explains that "[s]ubsection (4) is intended, despite the provisions of subsection[] (2) . . . , to prevent contractual provisions excluding modification except by a signed writing from limiting in other respects the legal effect of the parties' actual later conduct." *Id.* § 2-209 cmt. 4.

delivered and the contract contained a no-oral-modification clause in addition to the quality specification. The parties expect that deviations from the specified contract quality will sometimes occur. For the buyer, the expected loss from the average deviation (the probability of a deviation times the cost) is $100. Thus, sometimes a deviation will create a cost that approaches zero (perhaps the buyer has forgiving customers or the quality shortfall can be quickly corrected). At other times, a deviation can impose a loss whose expected value equals or exceeds $500 (say the buyer has a new, potentially large customer for whom quality is important, or the particular deviation would be slow to correct in a high-demand period). The parties also know that it would not be cost-justified to litigate against the average quality shortfall (the litigation cost would be, say, $150 per deviation).[104]

The contract in this example has three salient features: (1) The price does not fall with declines in the level of quality supplied, (2) there is a written specification of the quality the seller is to deliver, and (3) there is a no-oral-modification clause. These contractual features together with the cost of deviations support two conclusions with respect to the interpretive relevance of course-of-performance evidence. First, features (1) and (2) together with our cost assumptions imply that the parties do not expect to litigate the average quality deviation. Rather, the buyer will accept nonconforming deliveries that cause average losses, with the cost of these deviations reflected in a lower fixed price. Feature (3)—the no-oral-modification clause—implies that the seller should not infer from a series of acceptances of goods whose defects cause the buyer to incur losses in the neighborhood of the average (or less) that the buyer also will accept a nonconforming delivery that would cause it to incur a large loss. The second conclusion that the example supports, therefore, is that high-cost deviations in product quality are prohibited.

This example captures an important feature of contracting behavior. When business parties incur costs to cast obligations in written form, they do so partly to permit a party to stand on its rights under the written contract when standing on its rights matters. Course-of-performance evidence therefore commonly will be irrelevant to show what the contract originally meant or what it currently means. The parties' amicable behavior after the contract likely evidences only their view regarding how the average case should be treated. Courts, however, see the unusual case that

104. This example was stimulated by Ben-Shahar's model, but he apparently assumes that deviations impose the same loss in each period rather than the same loss on average. Ben-Shahar, *supra* note 52, at 796-800. On his assumption, parties are indifferent to whether courts admit course-of-performance evidence to show a change in meaning. Ben-Shahar later relaxes the same-loss assumption, however, and then concludes, as we do, that the soft version of the parol evidence rule will disadvantage typical parties.

the contract was written to govern.[105] Courts thus should recognize that parties sometimes have multiple intentions. Their actions under the contract will evidence their intentions for typical cases but seldom will evidence their intentions for the atypical case. Thus, a court is likely to make a category mistake when it relies on parties' behavior in nonlitigated cases to infer how parties want a litigated case to be treated.[106]

V. THE LEGAL DEFAULT PROJECT

In Part IV, we argued that sophisticated parties commonly prefer a default theory of interpretation that instructs courts to use narrow evidentiary bases and to presume that business contracts are written in the majority language. Courts can only interpret what is said, so our analysis assumed that the parties' writing was complete for the subjects at issue. Contracts, however, are often incomplete in relevant respects. Therefore, we now consider whether firms would prefer the state to complete these contracts with default legal terms. The somewhat surprising answer we derive from contract theory is that most state-created defaults will be useless or inefficient. Firms would prefer the state not to create inefficient defaults because firms will contract out of them; thus, the only effect these defaults will have is to increase transaction costs.

Commercial parties commonly want to condition their obligations on the nature of their contract partners or on the state of the world that will materialize after the contract is written. For example, a seller would like to condition its warranty obligation on the buyer's intensity of use: The seller would charge more or warrant less for more intense users because these users are more likely to make claims. A buyer also would like to condition price or quantity on the nature of demand ex post. If demand turns out to be high, the buyer would be willing to pay more or to take more product, but the buyer would want a lower price or less product if demand turned out to be low. These considerations imply that contracts will inevitably be

105. Our argument here is similar to Karl Llewellyn's view of custom. He believed that customs govern ordinary cases but seldom are relevant to the unusual cases that cause parties to litigate. *See* Alan Schwartz, *Karl Llewellyn and the Origins of Contract Theory*, *in* JURISPRUDENTIAL FOUNDATIONS, *supra* note 71, at 12, 16.

106. Arbitrators obey the parties' interpretive instructions, but courts do not. Our analysis thus identifies a reason for parties to use arbitration, but it cannot support the inference that parties use courts because they prefer the interpretive styles in current use. There are other reasons for using courts, such as the ability to get discovery, to have appeals, to have the substantive law applied by experts in it, to create effective precedents, and the like. *See* Kenneth S. Abraham & J.W. Montgomery, III, *The Lawlessness of Arbitration*, 9 CONN. INS. L.J. 355 (2003); Charles L. Knapp, *Taking Contracts Private: The Quiet Revolution in Contract Law*, 71 FORDHAM L. REV. 761 (2002). The widespread use of merger and no-oral-modification terms in contracts intended for courts also suggests, in line with our views, that parties who prefer judicial enforcement also prefer enforcement under a different interpretive style.

incomplete. There is an infinite number of possible future states and a very large set of possible partner types. When the sum of possible states and partner types is infinite and contracting is costly, contracts must contain gaps. Parties cannot write contracts about everything.

Incomplete contracts sometimes produce lawsuits because parties will not always agree ex post regarding the treatment of omitted contingencies.[107] Courts in such cases no longer can simply engage in interpretation because, by definition, the contracts lack words to interpret. The courts' task thus shifts to the development of rules to resolve gap cases. Hence, some default rules are judicially created.[108] *Restatement* and statutory drafters also create defaults when, in their view, certain gaps are likely to recur. Though our analysis has relevance for courts, we focus here principally on the *Restatement* and the UCC, asking just when the state can create good defaults for business parties.[109] In particular, we derive criteria for efficient defaults, and then argue that these criteria are difficult for drafters to satisfy.

A. *The Case for Defaults*

The conventional view is that, but for the cost, parties would write complete contracts. The state may increase efficiency in cases where contracting costs prevent parties from solving contracting problems. To see how, consider the problem of developing a damages rule—a contract term specifying the transfer, or the method of computing the transfer, that a party in breach must pay to its contract partner. Any particular set of parties would bear the full cost of solving this problem but likely could not capture the full gain because it can be difficult to prevent later parties from copying successful solutions that have evolved into contract terms.[110] In this

107. Parties tend to omit low-probability states. For example, if very low demand is thought quite unlikely, parties will not incur the contracting cost to specify their obligations in the very low-demand state. A lawsuit could arise if this state materializes and the parties cannot agree on a new price.

108. Since these rules are facilitative, courts should permit future parties to vary them. Judicial creation of such gap-filling defaults is not inevitable. A court instead can refuse enforcement when gaps cause a contract to be obligationally incomplete—that is, when the terms the contract does contain provide an insufficient basis on which to ground a remedy. The common law rule is that contracts leaving material terms incomplete or indefinite are not legally binding. It may be paradoxical that the indefiniteness rule is itself a default. Parties can opt out of it by more completely specifying their obligations in the contract.

109. The *Restatement* defaults differ from statutory defaults because a restatement is not self-executing. A restatement provision cannot become a legal default until it is both promulgated by the American Law Institute (ALI) and then adopted by a court when resolving a dispute. Thus, our analysis here applies both to the private legislatures that promulgate defaults and to the courts that adopt them.

110. For analyses of how contractual innovations spread, see Gerald F. Davis, *Agents Without Principles? The Spread of the Poison Pill Through the Intercorporate Network*, 36 ADMIN. SCI.

circumstance, the total social gain from having a rule—a solution to a contracting problem—may exceed the social cost, but parties themselves will not create the rule. There is a case, therefore, for a publicly supplied contract law that contains efficient solutions to common contracting problems.[111]

This conclusion implies that contract law rules should be defaults. This is because, in a large economy, parties are heterogeneous so that not everyone will want the same thing. The justification for a default rule is that it does for parties what they would have done for themselves had their contracting costs been lower. Hence, parties who can do for themselves—that is, who can create their own solutions—should be free to do so. While this conclusion may seem obvious, courts sometimes create barriers to contracting out of *Restatement* or UCC rules, on the ground that these rules reflect either better or fairer solutions than those that parties develop.[112] To the contrary, we argue that commercial parties themselves are the best judges of what constitutes a good contracting solution, and that default rules should not be chosen on fairness grounds.

Perhaps a more precise way to state the ground on which drafters should choose defaults is that a good default minimizes contracting costs. Parties, if left free, will supplant or modify state-supplied terms that they dislike. In a world of free contracting, unpopular defaults thus will raise contracting costs but not otherwise affect parties' behavior. As a consequence, the state's power is limited to reducing contracting costs, which it does best by enacting popular defaults. This leaves open the question of what makes a default popular. This is a difficult question to answer in general but, as we have seen, firms prefer contract terms that maximize joint surplus. Defaults thus will be popular with firms if they maximize joint surplus and unpopular otherwise.[113]

Q. 583 (1991); and Gerald F. Davis & Henrich R. Greve, *Corporate Elite Networks and Governance Changes in the 1980s*, 103 AM. J. SOC. 1 (1997).

111. *See* Charles J. Goetz & Robert E. Scott, *The Limits of Expanded Choice: An Analysis of the Interactions Between Express and Implied Contract Terms*, 73 CAL. L. REV. 261, 291-93 (1985).

112. *See id*. at 284-85. The courts' behavior explains the intensive lobbying by firms regarding the proposed revisions to Article 2, especially the warranty terms. *See* Robert E. Scott, *The Rise and Fall of Article 2*, 62 LA. L. REV. 1009, 1049-53 (2002). Firms know that judicial action can make a UCC default sticky in practice.

113. Russel Korobkin has argued that the "endowment effect" makes defaults hard to change. Individual subjects in experiments manifest the endowment effect when they demand much more to sell an item—for example, a coffee mug—than they would bid to purchase the item. Put another way, persons' preferences are in part a function of the legal assignment of property rights; they are not independent of that assignment, as the Coase Theorem supposes. In the realm of contract law, a party subject to the endowment effect would ask much more to give up a default that is favorable to it than the party would bid to obtain from its contract partner a clause that is identical to the default. As a consequence, for such parties defaults are more like mandatory rules; the endowment effect would prevent parties from conveniently altering the allocations that the defaults create. Korobkin conducted experiments casting first-year law students in the role of

This reasoning shows the error of choosing defaults just because they are fair. Parties have the incentive (and often the ability) to contract out of even fair defaults that do not maximize surplus.[114] As a good example of this response, commentators and courts once thought (and may still think) that it is fair for merchant sellers to make implied warranties of quality. Merchant sellers, however, routinely attempt to disclaim the implied warranty of merchantability in section 2-314 of the UCC, and their merchant buyers routinely consent.[115] Thus, section 2-314 in business contexts often does nothing more than increase transaction costs. Drafters and courts therefore should ask what parties would want, not what parties

attorneys, and found that they advised clients in a manner that manifested endowment effects. *See* Russell Korobkin, *The Status Quo Bias and Contract Default Rules*, 83 CORNELL L. REV. 608 (1998). A more recent treatment is Russell Korobkin, *The Endowment Effect and Legal Analysis*, 97 NW. U. L. REV. 1227 (2003).

Professor Korobkin's experiments are not relevant to the types of transactions we consider. Recent theory attributes the endowment effect to the existence of a real option on the subject of sale. A party uncertain about the value of a good to her can delay a purchase or sale while gathering more information. To buy or to sell thus exercises the option because information is no longer relevant after the transaction. As a consequence, the seller's price is the sum of the good's cost and the (positive) value of the option to delay sale; the buyer's bid is the sum of the value of the good to him and the (negative) value of the option to delay purchase. Adding option values to an ask price and subtracting them from a bid price produces differences between them that are of the same order of magnitude as are manifested in the psychological experiments. *See* Massimo Paradiso & Antonella Trisorio, *The Effect of Knowledge on the Disparity Between Hypothetical and Real Willingness To Pay*, 33 APPLIED ECON. 1359 (2001); Eric van Dijk & Daan van Knippenberg, *Buying and Selling Exchange Goods: Loss Aversion and the Endowment Effect*, 17 J. ECON. PSYCHOL. 517 (1996); Jinhua Zhao & Catherine L. King, *A New Explanation for the WTP/WTA Disparity*, 73 ECON. LETTERS 293 (2001). An implication of these articles is that the endowment effect should be small or absent when sophisticated parties who know what they are doing trade goods for money. Daniel Kahneman thus remarked: "Loss aversion plays little role in routine economic transactions, in which a seller and a buyer exchange a good and money, both of which were held for that purpose." Daniel Kahneman, *Reference Points, Anchors, Norms, and Mixed Feelings*, 51 ORGANIZATIONAL BEHAV. & HUM. DECISION PROCESSES 296, 301 (1992). Consistent with this view, sophisticated actors trading in real markets have been shown not to manifest endowment effects; their preferences are independent of the initial location of property rights to traded objects. *See* John A. List, *Does Market Experience Eliminate Market Anomalies?*, 118 Q.J. ECON. 41, 42-43 (2003) ("[M]arket experience matters: across all consumer types, marketlike experience and the endowment effect are inversely related. In addition, within the group of subjects who have intense trading experience (dealers and experienced nondealers), I find that the endowment effect becomes negligible."). Our analysis assumes away endowment effects because we study transactions between firms with "intense trading experience" who "trade goods for money."

114. If the state wants a particular rule actually to control behavior, it should make the rule mandatory. Currently, parties are permitted to create their own contracts, but sometimes only at the cost of contracting out of state-created defaults. The costs of contracting out will exceed the gains for some firms but not others. As a consequence, a fair but inefficient default disadvantages firms with relatively high contracting costs, but cannot fully realize the state's fairness concern because other firms will create their own deals.

115. Section V.C below argues that the UCC's implied warranty term is inefficient because it provides a seller with too little guidance regarding the performance obligation and also creates moral hazard.

should want.[116] We argue in the next Section that this is a difficult question to answer.

B. *The Cost Concern and Default Rules*

There are three criteria for a good default rule: It must be conditioned on only a few possible states of the world, be relatively simple in form, and be efficient for a wide variety of contract parties. The first criterion is necessary because there can be an infinite number of possible future states of the world to which a rule could apply, while the state's drafting costs are finite. Thus, a publicly supplied rule that had to address many possible future states seldom would be cost-justified to create. The second criterion, of simplicity, is a function of institutional competence. Contract law rules are created by courts and drafters. Courts cannot conduct investigations into the efficiency properties of possible rules and rule combinations. Drafters also have limited resources. The American Law Institute (ALI), which creates restatements, and the National Conference of Commissioners on Uniform State Laws (NCCUSL), which together with the ALI created the UCC, do not hold legislative-style hearings on proposed rules, cannot hire neutral economic or industry experts to help in rule creation, and generally rely on the part-time labor of law professors and private attorneys.[117] The drafters thus cannot, and do not, write rules for business contexts that match the complexity of, say, the tax code; rather, contract law necessarily is restricted to the consideration of problems whose solutions can be embodied in simple rules.

The third criterion, of efficiency, is perhaps the most difficult to satisfy, because parties in large economies are heterogeneous. Default rules would be too expensive to create if efficient solutions were party-specific. Then there would need to be as many legal rules as there are sets of contracting

116. Scholars have identified a set of default rules that are termed "information-forcing" because the rules create an incentive for a party to disclose relevant information to its contract partner. *See* Ayres & Gertner, *supra* note 1. Information-forcing rules include those providing that damages cannot be recovered unless they are foreseeable and reasonably certain; the rules create an incentive for a promisee who fears breach to disclose in advance to the promisor the loss the promisee would suffer. The promisor then will be able to price more accurately and to take optimal precautions against breach. Information-forcing rules fall within the analysis here because parties can create them on their own. For example, sellers routinely propose contracts that disclaim liability for consequential damages. This creates an incentive for buyers who may suffer these damages to disclose them to the seller and then to purchase insurance by bargaining away the disclaimer. An information-forcing default thus will be popular if parties would agree to disclose in the circumstances in which the rule applies. It has been shown that the foreseeability default may not always be popular in this sense. *See* Johnston, *supra* note 10, at 626-38 (stating that parties facing a monopolist may not want to disclose the profits they would earn from successful transactions because the monopolist then may price discriminate against them).

117. *See* Alan Schwartz & Robert E. Scott, *The Political Economy of Private Legislatures*, 143 U. PA. L. REV. 595, 607-37 (1995).

parties. The task, then, is to find rules that would be efficient—surplus-maximizing—in a wide variety of contexts.[118]

The difficulty of satisfying the three criteria explains why there are very few default rules[119] in contract and commercial law. We illustrate the stringency of the criteria by considering two variants of the investment example analyzed in Section III.B above. Assume first that, after the parties contract for the seller to produce the specialized product, but before the goods are produced, the seller's factory is destroyed in a fire. The seller no longer can produce the goods, but it can pay money damages.[120] The common law default rule nevertheless excuses the seller when an exogenous event makes the contemplated, specialized performance impossible to render.[121] This excuse rule seems to satisfy the criteria for an efficient default. First, the costs of creating the rule are low. There are only two possible future states of the world—either an exogenous event prevents performance or it does not—and there is need for a rule in only one of them (when the seller cannot perform). The rule also is simple in form; the seller is excused or not, depending on whether there was a fire or not. Finally, the rule likely is efficient for a wide variety of parties. This is because buyers in general are better insurers against lost valuations of specialized investments than are sellers; buyers usually are better informed than sellers about the consequences of sellers' breach. Excusing the seller requires the buyer either to insure on the market or to reveal its valuation to the seller.

118. This note puts the text's point formally; the text goes on to give examples. A market type is a two-element set composed of a seller and a buyer. Denote a particular type as t_i. There are $t_1, \ldots, t_n$ types in the market, and there are m sets of parties of each type. Thus, $m(t_i)$ denotes the total number of parties of a particular type. Let c_{pi} denote the cost for each t_i set of parties to create a contract term, and let c_{si} be the cost to the state of creating a default rule that is identical to that term. We suppose $c_{si} > c_{pi}$ because state-created rules require legislative or administrative action. Denote the benefit to a set of t_i parties from having a particular contract term (or default rule) as $b(t_i)$. Parties to a particular contract will not create the term if $c_{pi} > b(t_i)$. The state should create a default if the total benefits to all t_i parties would exceed the state's cost of rule creation. This efficiency condition can be written as

$$csi < \sum_{1}^{m(t_i)} b(t_i).$$

This condition cannot be satisfied when $m(t_i) = 1$ because $c_{si} > c_{pi}$, and we have assumed that private rule creation would be inefficient ($c_{pi} > b(t_i)$). State rule creation becomes more efficient as $m(t_i)$ increases (the number of parties who would benefit from a rule goes up), but the text next argues that for contract law rules, $m(t_i)$ is more likely to be close to one than close to a large number. A more extensive analysis of the heterogeneity point is in Alan Schwartz, *The Normative Implications of Transaction Cost Economics*, 152 J. INSTITUTIONAL & THEORETICAL ECON. 287 (1996).

119. Recall that a rule specifies the relevant conduct in advance—e.g., no driving above fifty-five miles per hour.

120. The buyer's expectation interest would be its valuation less the price, which was $15 in the example.

121. *See, e.g.*, Taylor v. Caldwell, 122 Eng. Rep. 309 (Q.B. 1863); RESTATEMENT (SECOND) OF CONTRACTS § 263 (1981).

This example is deceptively simple, however. The set of possible causes of a seller's incapacity to perform can be large, ranging from fire or flood, to a temporary or permanent supply shortage, to a government regulation, and so on. A seller's ability to anticipate and to take precautions against the causes of incapacity thus will differ across sellers and among causes. It would be efficient to excuse the seller only when its inability to perform resulted from causes that were difficult to anticipate and prevent; otherwise, sellers would use the excuse doctrine just to escape deals that turned out badly. Thus, an excuse rule must resolve difficult causation issues. The seller also may be able to perform in part or to perform in full but only for some contract buyers. There is a question how a seller should make the resulting allocation decisions. Moreover, some buyers may be more able than others to take precautions or to insure efficiently. Drafting rules to cover all of these possible causes, effects, and parties would be quite costly. As a consequence, the excuse case today is regulated by a standard, not by a set of rules. Section 2-615(a) of the UCC provides that a seller's failure to perform "is not a breach . . . if performance as agreed has been made impracticable by the occurrence of a contingency the non-occurrence of which was a basic assumption on which the contract was made."[122] Courts decide after the fact whether a performance would have been "impracticable" and whether the parties had "basically assumed" that the allegedly excusing cause would not occur or would not materially affect performance if it did occur.

Now consider a second illustration. Recall that the investment example implicitly assumed that the parties' contract specified what a specialized product would do, so that if the seller delivered something else, breach would be clear. A default rule might also specify the seller's quality obligation, but recall that the product was *specialized to the buyer's use*. A legal rule that set out this particular seller's quality obligation thus would apply only to one case. It is inefficient to draft for one case. A particular seller's quality obligation, however, sometimes can be generalized to an industry. It may be, for example, that book publishers would agree on what requirements a good printing press must meet. Even so, if an Article 2 drafter were to describe, for every industry, the quality that sellers must deliver to buyers, the drafting cost would likely exceed the social gain. Unsurprisingly, the law regulates the seller's quality obligation with standards. Under section 2-314(2), goods "to be merchantable must be at least such as (a) pass without objection in the trade . . . and (c) are fit for the ordinary purposes for which such goods are used."[123] Courts after the fact

122. U.C.C. § 2-615(a) (2003). Section 261 of the *Restatement* contains the same rule in almost the same language. RESTATEMENT (SECOND) OF CONTRACTS § 261 (1981).

123. U.C.C. § 2-314(2).

decide whether it was enough for a seller to satisfy subsection (a) or (c) ("at least such as"); whether the goods would pass without objection; what "ordinary" seller purposes are; and whether the goods "are fit" for such purposes.

Contract and commercial law thus contain very few default rules, because the criteria for good default rules are so difficult to satisfy. Rather, the *Restatement (Second) of Contracts* and Article 2 of the UCC primarily contain standards; these texts are replete with provisions requiring parties to behave "reasonably," "conscionably," "fairly," "in good faith," and the like.[124] These codifications also fail to address important contracting problems. For example, parties to long-term contracts face the problem of keeping each party's gain under the contract above the opportunity cost of performance in every period. This problem is best addressed with index clauses, which tie the current period price to current cost and demand conditions. Efficient indices are party- or industry-specific, however, and so are too costly for contract law to create.[125] Instead, the law ignores this long-term contracting problem in favor of letting parties solve it for themselves. In sum, the cost concern forces contract law commonly to regulate with standards rather than rules, when it regulates at all.

C. *The Moral Hazard Concern and Default Standards*

The project of creating publicly supplied default standards has been unsuccessful for two reasons. First, firms often need specific guidance regarding the performance obligation. Second, standards create unacceptable moral hazard. Regarding the need for guidance, a seller commonly must know what quality level to produce. Telling the seller that its product should "at least" satisfy a buyer's "ordinary purposes" is generally unhelpful. In practice, sellers of complex products thus disclaim the implied warranty of merchantability in favor of an express warranty that describes important aspects of the seller's performance obligation. As another illustration, when an exogenous event induces the seller to consider breaching, both parties need to know whether breach would be legally permissible. An erroneous decision could expose one or the other party to a

124. *See id*. §§ 2-205 to -209, 2-302, 2-305 to -307, 2-309, 2-311 to -312, 2-314 to -315, 2-317, 2-503 to -504, 2-508, 2-511, 2-513, 2-602 to -610, 2-612, 2-614 to -615, 2-704 to -706, 2-709 to -710, 2-712, 2-714 to -716, 2-718 to -719, 2-723. The UCC also adopts standards in order to avoid resolving controversial issues. *See* Schwartz & Scott, *supra* note 117. Our analysis here is meant to show that even if the ALI and NCCUSL were bolder, the cost concern would generate many standards.

125. An index clause links the transaction price between the parties to verifiable external indices, such as the Consumer Price Index (CPI), that correlate with the economic conditions facing the parties. The degree of correlation is a function of the type of party (some firms' costs will move more closely with the CPI than other firms' costs will, for example), so that index clauses vary with particular parties or party types.

damages judgment. Telling the seller at this point, with reference to section 2-615, that breach is permissible if performance is "impractical" is also unhelpful. And telling the buyer that its return performance (i.e., a payment of an installment of the price) can be suspended under section 2-609 if suspension would be "commercially reasonable" is similarly of little use. In response to this difficulty, parties commonly ignore the standards in sections 2-615 and 2-609 in favor of writing force majeure and insecurity clauses that specify precisely both the exogenous causes that will excuse the seller's performance obligation and the permissible responses that the buyer can take when insecure about the seller's performance.[126] Parties that need specific guidance write detailed rules in their contracts. The state thus wastes drafting resources when it creates a standard that parties routinely reject.

We introduce the moral hazard difficulty with an example. Suppose that a set of contracts is obligationally complete, but some contracts in the set lack a solution to a contracting problem. For example, the contracts state prices and quantities but do not index prices to inflation. The contracts thus are insufficiently state-contingent. The prices will be suboptimal in some possible future states of the world. A court could enforce such a contract by awarding a disappointed promisee the difference between the contract and market prices. The issue is whether the state should fill the gap with a standard that could recite: "A contract price cannot be unreasonable in light of the conditions obtaining when performance is due." The state should be reluctant to adopt this standard because the case for a public default is absent. Contracting costs could not have prevented parties from drafting the standard for themselves. It would have cost the parties little to have required the seller to deliver a "reasonable" quality within a "reasonable" time but to have excused the seller from delivery if its performance became "impractical." Drafters should be reluctant to enact default standards without first asking why the standards were missing from private contracts.

Standards are unpopular because they increase the likelihood of moral hazard. Parties sometimes have incentives to take actions that are privately

126. Force majeure clauses apparently have a long pedigree. Consider, as an example, the following clause in a contract between Michelangelo and the heirs of Pope Pius III requiring Michelangelo to produce fifteen marble statues for the cathedral in Siena within a specified time:

> Item, whereas the said Michelangelo, by virtue of the said agreement, has bound himself to cause marble to be brought to Florence from the mountains of Carrara for the making of the said statues; and whereas, by reason of the besieging of the Pisans within the territory of Pisa, war has once more broken out; and whereas the Florentine Republic may endeavour to divert the course of the River Arno in such wise that the transport of the said marble from the mountains of Carrara to the city of Florence may therefore be hindered; and whereas also the said Michelangelo may fall ill, which God avert . . . then . . . the said [new contractual period] shall be suspended during the time of the said hindrance.

ANTON GILL, IL GIGANTE: FLORENCE, MICHELANGELO, AND THE *DAVID*, 1492-1504, at 215-16 (2002).

optimal but publicly undesirable. For example, a buyer has an incentive to claim that a product is unsuitable for one of its "ordinary" purposes when the market price drops.[127] A seller has an incentive to claim that its performance has become "impractical" when its costs rise;[128] or to attempt to get a price increase by using our illustrative standard to claim that inflation has made the contract price "unreasonable." When a standard governs, the party who wants to behave strategically must ask what a court will later do if the party is sued. The vaguer the legal standard and the more that is at stake, the more likely the party is to resolve doubts in its own favor. A party that resolves doubts in this way will attempt to maximize private gains at the expense of joint welfare maximization. Therefore, a standard is efficient only when the party on whom it confers discretion has the incentive to maximize joint returns in the course of maximizing its private gain.[129] Few UCC standards pass this test.[130] Consequently, parties contract away from them. For example, the customary product warranty for machines and other equipment contracts out of the quality standard (section 2-314), the cure standard (section 2-508), and the revocation-of-acceptance standard (section 2-608).[131] The common force majeure clause contracts out of the excuse standard (section 2-615), and many contracts contract out of the consequential-damages standard (section 2-715), the warranty-damages standard (section 2-714), and the standards governing insecurity and anticipatory repudiation (sections 2-609 to 2-610).[132] This evidence strongly suggests that typical parties dislike contract law's standards.[133]

127. *See, e.g.*, T.W. Oil, Inc. v. Consol. Edison Co., 443 N.E.2d 932 (N.Y. 1982).

128. *See, e.g.*, Aluminum Co. of Am. v. Essex Group, Inc., 499 F. Supp. 53 (W.D. Pa. 1980).

129. This test sometimes can be met. For example, the standard merger agreement contains a term permitting the buyer to exit without penalty if a "material adverse change" occurs in the interim between signing the contract and concluding the merger. This term does not specify the events that could constitute a material adverse change because there can be many such events whose effects will vary with the ex post state of the world. A standard nevertheless is efficient here because the buyer will exit only if the seller's value has fallen materially. The threat of exit creates an incentive for the seller to take synergy-increasing actions that lower the probability that its value to the buyer will fall materially. *See* Ronald Gilson & Alan Schwartz, Understanding MACs and MAEs (2003) (unpublished manuscript, on file with authors), *available at* http://islandia.law.yale.edu/ccl/papers/merger15.pdf. This example is offered to suggest that some statutory standards may be efficient, but it may also support the inference that parties will write a standard in their contracts when a standard would maximize surplus.

130. *See, e.g.*, Victor P. Goldberg, *Discretion in Long-Term Open Quantity Contracts: Reining in Good Faith*, 35 U.C. DAVIS L. REV. 319 (2002) (analyzing in detail the ways in which courts have used the good faith standard that regulates discretion in output and requirements contracts under section 2-306(1) to undermine the controls on discretion that the parties had set for themselves).

131. *See* ALAN SCHWARTZ & ROBERT E. SCOTT, COMMERCIAL TRANSACTIONS: PRINCIPLES AND POLICIES 204-24 (2d ed. 1991).

132. Before 1978, parties routinely contracted out of section 2-609 with ipso facto clauses, which stated precisely when a party could cancel the agreement in consequence of its partner's financial difficulty. The current Bankruptcy Code outlaws these clauses on the (erroneous) ground that they permit solvent parties to deplete the bankruptcy estate. *See* 11 U.S.C. § 365(e)(1) (2000). For an analysis, see Yeon-Koo Che & Alan Schwartz, *Section 365, Mandatory Bankruptcy Rules*

Therefore, when contracts have a gap that a standard could fill, the best inference for a decisionmaker to draw is not that the standard is "missing" from these contracts, but rather that the standard has been rejected. This is the best inference because of the ease with which parties could have drafted their own standard and because many standards create moral hazard. Parties probably rejected an ex ante solution—drafting a standard—in favor of renegotiating when circumstances changed, with the written terms constituting the disagreement points that would determine whether a new deal was efficient. And to summarize, filling gaps with standards ordinarily will be inefficient for three reasons: The state wastes resources in drafting them; the parties waste resources in contracting out; and when courts are expected to use standards actively to police bargains, parties may create sets of rules they would otherwise have preferred to omit.[134]

In sum, the project of creating default rules founders on the costs of rule creation for heterogeneous parties that function in complex commercial environments. These costs essentially preclude the creation of all but a few default rules.[135] The project of creating default standards founders on the need of parties for specific guidance as to what they are supposed to do, and on the tendency of standards to create moral hazard. These gloomy conclusions hold even when the information that a rule or standard would require is assumed to be known to the parties and accessible to the courts.

and Inefficient Continuance, 15 J.L. ECON. & ORG. 441 (1999). A concrete example of contracting out occurs in *Northwest Lumber Sales, Inc. v. Continental Forest Products, Inc.*, 495 P.2d 744, 748-50 (Or. 1972).

133. Ian Ayres has suggested that standards may be good defaults because precedent can crystallize around them, thereby providing parties with guidance. *See* Ian Ayres, *Making a Difference: The Contractual Contributions of Easterbrook and Fischel*, 59 U. CHI. L. REV. 1391, 1415-16 (1992). This argument assumes that legal standards will become operative implied terms in parties' contracts, thereby giving courts the opportunity to apply them. The premise generally fails because parties contract out of the standards.

134. The last danger is said to have occurred in England when the courts adopted an excuse standard. As English lawyer Andrew Rogers, Q.C., wrote in his introduction to a book on the excuse doctrine:

> It is only in relatively recent times that the English courts relented in their demand that the strict words of the contract be adhered to, notwithstanding changes in circumstance. The resulting doctrine of frustration has not worked satisfactorily. The courts found it difficult to determine . . . the limits for its application. . . .
>
> . . . .
>
> . . . [F]rustration under the general law brings in its train automatic discharge. To avoid such results, the parties to the contract are required to draft their own particular code. That in turn means a substantial increase in transaction costs.
>
> Notwithstanding the costs involved, in an effort to meet the difficulty many contractors have undertaken fairly detailed contractual risk allocation.

Andrew Rogers, *Foreword* to FORCE MAJEURE AND FRUSTRATION OF CONTRACT, at v, v-vi (Ewan McKendrick ed., 2d ed. 1995).

135. A rare example of a successful default rule is the requirement that a breacher pay the other party the difference between the contract and market prices. The rule applies in only one state of the world: when there is breach. It is simple to apply because the court only compares the contract and market prices, and it is efficient for many parties because the rule protects the expectation interest. Other good default rules are hard to find.

For example, the parties and a court will know whether a fire destroyed the seller's factory in whole or in part. Information often is asymmetric, however. Parties may not know relevant things about each other, and courts may not know relevant things about parties. The existence of asymmetric information exacerbates the moral hazard concern and thus makes the creation of efficient defaults even more difficult.

D. *The Asymmetric Information Concern*

The existence of asymmetric information truncates the set of contracts that parties can write. As a consequence, it also truncates the set of defaults that drafters can write. To see why, it is helpful to begin with an information taxonomy from the contract theory literature. A datum of information is "unobservable" if a party cannot observe it. Buyers ordinarily cannot observe a seller's production cost. A datum of information is "observable but not verifiable" if a party can observe it, but cannot verify the information's existence to a third party such as a court at an acceptable cost. For example, an employer usually can know which employees sometimes shirk, but it would be expensive relative to the gains to prove to a court that a particular employee shirked twenty percent of the time. A datum of information is "verifiable" if a party can both observe it and establish its existence to a third party. Information is asymmetric when it is either unobservable or unverifiable.

To understand the constraints on contracting behavior caused by asymmetric information, we change the relation-specific investment example of Section III.B in two ways. First, we let the seller's costs be stochastic. In some states of the world, the seller's cost to produce the specialized product will be lower than the buyer's valuation, as in the original example; in other states of the world, the seller's costs will exceed the buyer's valuation, so that trade would be inefficient. Second, we let the buyer invest to increase the value to the buyer of the seller's performance. These changes generate the well-known overinvestment problem.[136] The buyer's damages if the seller breaches will be the difference between its realized valuation and the price. The buyer thus will invest until the marginal cost of further investment equals the marginal increase in value. This is too much investment, however. In those states of the world in which it is inefficient for the parties to trade, the buyer's investment has no social value; it serves only to increase the damages the seller will pay.

136. The problem was identified in Steven Shavell, *Damage Measures for Breach of Contract*, 11 BELL J. ECON. 466 (1980), and elaborated on in William P. Rogerson, *Efficient Reliance and Damage Measures for Breach of Contract*, 15 RAND J. ECON. 39 (1984).

This problem would vanish if the seller could observe the buyer's production function—the functional relation between the possible investment levels the buyer could choose and the values associated with those levels. The parties then could write a liquidated damages clause that would award the buyer the difference between (1) the value the goods would have had were the buyer to invest optimally and (2) the price. This value would be lower than the unconstrained value that would be produced if the buyer never took into account that in some states of the world its investment would generate no value at all (because the parties would not trade in those states). Such a contract can seldom be written, however, because sellers can seldom observe buyers' production functions. As a consequence, a buyer would specify as liquidated damages the unconstrained value to the buyer of the seller's performance, and then invest to realize that value.[137]

To be sure, this example does not show that parties could never induce efficient investment when information is asymmetric. When parties can observe, though not verify, relevant variables such as the buyer's valuation, theorists have developed a number of contracts that could induce sellers to invest efficiently to reduce costs and buyers to invest efficiently to increase value.[138] These contracts, however, are "parameter specific"—that is, the prices the contracts set and the actions they require depend on the particular values, costs, and probability distributions that parties face.[139]

137. Courts cannot prevent this practice because a court can observe only what the parties can observe. If the court does not know the value that optimal investment would have generated, it cannot award that value to the buyer. It must instead award the value the buyer actually lost. A nice, though technical, discussion of the impossibility of writing first-best efficient contracts when both the seller's costs and the buyer's valuation are unobservable is Patrick W. Schmitz, *On the Interplay of Hidden Action and Hidden Information in Simple Bilateral Trading Problems*, 103 J. ECON. THEORY 444 (2002).

138. Concise but moderately technical discussions of these contracts are Patrick W. Schmitz, *The Hold-Up Problem and Incomplete Contracts: A Survey of Recent Topics in Contract Theory*, 53 BULL. ECON. RES. 1 (2001); and Alan Schwartz, *Incomplete Contracts*, *in* 2 THE NEW PALGRAVE DICTIONARY OF ECONOMICS AND THE LAW, *supra* note 30, at 277. A nice—though also technical—example is Benjamin E. Hermalin & Michael L. Katz, *Judicial Modification of Contracts Between Sophisticated Parties: A More Complete View of Incomplete Contracts and Their Breach*, 9 J.L. ECON. & ORG. 230 (1993).

139. Asymmetric information also can create "ambiguity aversion," and when it exists an attempt by the state to fill contractual gaps would usually be futile. To see why, assume that $n > 1$ probability distributions may characterize the occurrence of a possible event or state of affairs. A party behaves in an ambiguity-averse fashion when it calculates its expected payoffs by using the lowest probability in any of the n possible distributions. Suppose, then, that party F would be better off if state ω materializes than party G would be, while party G may be better off if state ψ materializes than party F would be. When F is calculating its expected payoffs, it will use the probability distribution that puts a relatively higher weight on state ψ obtaining, in which it is worse off, than it would put on state ω obtaining, in which it is better off. For similar reasons, G will put a relatively greater weight on state ω obtaining. When the parties act in this fashion, they calculate their expected payoffs differently even though they agree on the monetary payoffs; the parties estimate differently because they are estimating with different probability distributions. This behavior can create an inefficiency. The sum of the payoffs that ambiguity-averse parties

The lesson is that the state cannot help when asymmetric information prevents parties from writing the efficient contract. Parties thus would reject default standards that permitted the seller to obtain "a commercially reasonable price in proportion to its cost of production," or that permitted the buyer to recover "damages in proportion to commercially reasonable reliance investments made in good faith." Under the former proposed standard, the seller always would claim to have high costs, while under the latter, the buyer always would claim that all its investments were commercially reasonable and were made in good faith. As another example, the UCC provides in section 2-715 that a seller must either perform or pay the buyer consequential damages measured by the difference between the contract price and the value of the goods to the buyer.[140] This requirement is superficially efficient because it induces the seller to perform when performance would increase value and to breach otherwise. Valuations, however, often are difficult to verify. As a consequence, buyers have an incentive to overstate their valuations, thereby inducing sellers to perform even when it would be less costly to breach and pay true damages.[141] Commercial parties respond to this problem by routinely contracting out of section 2-715. In place of the law, parties create complex repair-and-replacement provisions that strive for efficiency in other ways.[142] An appreciation of the information problems that parties confront compels this conclusion: A good default does for parties what parties would have done for themselves had their contracting costs been lower. When asymmetric information prevents parties from writing certain types of contracts (even

expect to occur will be less than the true expected surplus (since each party is calculating with the most unfavorable probability). As a consequence, the parties may perceive an efficient deal, which in fact would create gains in excess of costs, as creating costs in excess of gains. A response to this problem is to use an incomplete contract and to rely on renegotiation to determine the terms of trade in some possible future states. This can be more efficient than writing a more complete ex ante contract because when a state of the world materializes, the parties will know what their actual gains from trading in that state would be. They will then trade when trade is efficient and otherwise will not. Ambiguity aversion is more likely to occur when parties face high uncertainty (so that it can be hard to know just which probability distribution is the correct one to use). And when uncertainty is great, the parties' preference for contractual incompleteness suggests that filling gaps with state-created rules or standards would not advance the parties' objectives. A recent discussion thus stated: "The argument . . . suggests why incomplete contracts may not be a paradox in a world with ambiguity aversion since their inefficiency will not be 'readily fixable'—complete contracts do not help very much." Sujoy Mukerji, *Ambiguity Aversion and Incompleteness of Contractual Form*, 88 AM. ECON. REV. 1207, 1218 (1998).

140. Under section 2-715(2)(a), the buyer can recover (in addition to direct damages under sections 2-712, 2-713, or 2-714) "any loss resulting from general or particular requirements and needs of which the seller at the time of contracting had reason to know." U.C.C. § 2-715(2)(a) (2003).

141. The prospect of inefficient performances also would induce excessive precautions against breach.

142. These clauses disclaim the seller's liability for consequential damages and restrict the buyer's right to reject, but obligate the seller to repair or replace defective tenders. *See* SCHWARTZ & SCOTT, *supra* note 131, at 204-09.

when their contracting costs are zero), a state-supplied default serves no purpose.[143]

E. *Summary*

Much of what is commonly called contract law consists of the sets of default rules and, much more frequently, default standards contained in the *Restatement* and UCC Article 2. We argue here that the efforts of the *Restatement* writers and UCC drafters to create this law have been largely wasted. Default rules have proved too expensive to write for large heterogeneous economies. Default standards founder over the parties' need for ex ante guidance and the property of standards to create moral hazard. The moral hazard difficulty is exacerbated by asymmetric information. Defaults that are conditioned on behavior that parties or courts cannot observe will be exploited for private ends. Thus, parties contract away from them. As a consequence, inefficient defaults only raise transaction costs unnecessarily.[144]

What, then, is the proper role of courts in resolving disputes over incomplete contracts? It is appropriate for courts to apply a default standard

143. Karen Eggleston, Eric Posner, and Richard Zeckhauser recognize that parties may write simple contracts as a consequence of asymmetric information, but they urge courts to interpret these contracts liberally. That is, courts should "freely" insert terms in a way that "(roughly speaking) maximizes the ex ante joint surplus of the contract." Karen Eggleston et al., *The Design and Interpretation of Contracts: Why Complexity Matters*, 95 NW. U. L. REV. 91, 127 (2000). These authors do not show how courts could create efficient terms when sophisticated parties with money at stake could not, nor do they consider the possibility that sophisticated parties would contract out of unhelpful judicial efforts.

144. This claim may appear vulnerable to the objection that inefficient defaults are "harmless bromides" because parties are already writing the contract; their marginal cost of avoiding a bad legal rule, thus, is slight. This objection is mistaken on two levels. First, the total social cost of avoiding an inefficient legal rule is not slight. The small cost of changing the rule for a particular contract must be summed over all of the transactions that the rule affects in the decades that the rule will exist. Second, it is actually costly to change a bad rule when writing a particular contract because courts tend to regard state-created defaults as presumptively fair or efficient. This institutional bias raises the cost of contracting out. *See, e.g.*, Hayward v. Postma, 188 N.W.2d 31, 33 (Mich. Ct. App. 1971) (holding that parties must use clear and unequivocal language to shift liability for the risk of loss from seller to buyer); Davis v. Small Bus. Inv. Co. of Houston, 535 S.W.2d 740, 744 (Tex. Civ. App. 1976) (holding that a contractual provision purporting to allocate to the debtor the burden of "all" expenses incurred in preserving collateral was insufficient to trigger the "unless otherwise agreed" provision of Texas's equivalent of former section 9-207(2)(a) of the UCC, which required that "reasonable expenses" be chargeable to the debtor); Caudle v. Sherrard Motor Co., 525 S.W.2d 238, 240 (Tex. Civ. App. 1975) (taking a similar approach with respect to Texas's equivalent of section 2-509 of the UCC). Moreover, judicial interpreters may be reluctant to give the express language of the contract a meaning that conflicts with the relevant default. *See, e.g.*, Nanakuli Paving & Rock Co. v. Shell Oil Co., 664 F.2d 772, 794-805 (9th Cir. 1981) (holding that a merger clause that excludes evidence of prior dealings does not bar evidence of usage of trade to alter the price term in the contract); Legnos v. United States, 535 F.2d 857, 858 (5th Cir. 1976) (holding that despite the express term "F.O.B. vessel" in a contract, international context requires that "the intention of the contracting nations, rather than definitional niceties, must be given controlling weight").

as long as that standard does not create moral hazard. Otherwise, courts have a choice: to dismiss a case on the ground that a contract is too indefinite to enforce, or to read the contract to reach a reasonable result. Courts exercise both choices today, sometimes refusing to enforce,[145] and sometimes, particularly for partly performed contracts whose dismissal would create hardship, interpreting the contract to reach a reasonable solution. We argue here that drafters should not infer default standards from some courts' understandable efforts to do justice in particular cases. Rather, our view is that the UCC or *Restatement* drafters commonly should do nothing, and that courts should be hospitable to attempts by later parties to alter or avoid earlier rules of the case.

We have argued in Part V that firms would reject the "default project" if they could because the project generates many inefficient terms that contracting parties must incur costs to avoid. There is, however, a welfare-maximizing role for the state in creating certain default *structures*. While parties often will incur the costs of creating contracts, they almost never will find it cost-justified to create a structure, such as a close corporation law or a bankruptcy act. Structures that create default procedures for parties to follow when conducting (or terminating) different forms of enterprise, and that specify the legal consequence of following these procedures, have created very large welfare gains. Substantive contract law, therefore, may simply function on the wrong level of generality: Business parties are best suited to create their own contracts, whereas the state is best suited to create the broad structures within which the parties' contracts fit. These structures have been largely unexamined from the perspective of contract theory.

VI. MANDATORY RULES

Mandatory contract law rules ban terms that parties choose; hence, these rules are inconsistent with the commitment to party sovereignty that we have defended. The rules nevertheless are justifiable on two grounds. The first is to prevent externalities, the classic example of which is price

145. It is widely believed that contemporary courts ignore the indefiniteness doctrine. Conventional wisdom holds that courts should (and do) fill contractual gaps with general standards of reasonableness and good faith. *See, e.g.*, 1 CORBIN, *supra* note 94, § 95, at 400. This conventional wisdom is misleading. In a recent article, one of us shows that the indefiniteness doctrine has survived the influence of the UCC and the widespread academic support for filling gaps with standards. *See* Scott, *supra* note 33. In dozens of cases, American courts dismiss claims for breach of contract on the grounds of indefiniteness, often without granting any relief to the disappointed promisee. Scott argues that parties write incomplete contracts in the face of a rigorous indefiniteness doctrine when they can create efficient extralegal mechanisms for coping with problems of hidden information and hidden action. Using the language of this Article, his claim is that parties sometimes can solve moral hazard problems when courts cannot.

fixing.[146] The second ground is to ameliorate a market failure that disclosure cannot cure. As an illustration, consumers are thought to be poorly informed about the odds of product defects, but it is very difficult to communicate probability information in the format of a product label. The law's response is to ban disclaimers under the doctrine of strict liability in tort. But neither of these traditional grounds can justify the mandatory interpretation rules that we discussed in Part IV. Terms such as merger clauses and no-oral-modification provisions affect only the parties to the contract. There is no market failure to which the rules barring their enforcement respond.

This suggests that there is a third, largely unremarked, ground for mandatory contract law rules such as the rules that govern contract interpretation. This third ground for mandatory rules is a form of paternalism whose logic, in the interpretation context, runs like this:

(A) There are good reasons for courts to enforce the parties' intentions.

(B) Parties sometimes use contract terms, such as merger clauses or no-oral-modification clauses, that restrict the ability of courts to find the parties' intentions.

(C) There is no good reason for the existence of such "intention-blocking" terms.

(D) The terms thus are inconsistent with the parties' true or deep intention, which is to have the court enforce their actual deal; hence, courts should enforce the actual deal.

The interpretation rules are paternalistic because courts override the parties' expressed preferences out of concern for the parties' welfare. Courts, that is, do not conceive of themselves as imposing an agreement that parties would reject had they considered the matter under ideal deliberative conditions. Rather, courts believe that they are supplying the agreement that those conditions would have produced.[147]

The difficulty lies with the premise specified in step (C), that parties have no good reason to write contracts with the suspect terms. Our discussion in Part IV shows that rational, well-informed and uncoerced

146. The rules prohibiting fraud and duress also function to prevent a party from externalizing costs. Fraud and duress do not create social wealth but rather redistribute it (unfairly) between the parties. Were these practices permitted, the costs of enforcing the redistributions would be externalized to society. Thus, fraud and duress are grounds for declining to enforce a contract, because the resulting deal does not maximize joint welfare.

147. Traditional accounts of paternalism require the paternalist—here a judge—to aim at improving the agent's welfare. For a discussion, see Richard J. Arneson, *Mill Versus Paternalism*, 90 ETHICS 470 (1981); and Seana Valentine Shiffrin, *Paternalism, Unconscionability Doctrine, and Accommodation*, 29 PHIL. & PUB. AFF. 205 (2000).

parties have good reasons to use the intention-blocking contract terms that the Code disfavors and that many courts refuse to enforce. If the premise in (C) is rejected, however, then the conclusion in (D)—that courts are implementing the deep intentions of the parties—must fall as well. Paternalism with respect to the interpretation rules thus is misplaced because it frustrates rather than advances the parties' welfare: Parties choosing under ideal conditions would want what the state prohibits. In this Part, we discuss three other contexts in which the rules are mandatory. In each case, the rules can only be justified on the premise that parties have no good reason to use the terms that the legal rules prohibit, but in each case that premise fails. These rules, too, reflect a misplaced paternalism.

A. *Parties Cannot Ban Modifications*

Parties are formally free to require modifications to be written, but they are not formally free to prevent themselves from modifying their contract in the future.[148] In the courts' view, the best inference from the existence of a modification is that the parties' original intention to prevent modifications has become outmoded. Expressed more fully, parties would not change their minds regarding the deal's substance without good reasons, so parties can have no good reason to prevent themselves from changing their minds regarding the deal's substance. Contracting parties, however, do have good reasons to freeze their original deal. We next give an example to show that modification bans can enable parties to induce efficient relation-specific investment.[149] Since a major reason for enforcing contracts is to encourage relation-specific investment,[150] the refusal to enforce a modification ban violates a basic justification for the existence of contract law itself.

The example involves the production and sale of a product. The finished product can have two values for the buyer—high or low—depending on the ex post state of the world and how the product is made and delivered. The seller can make a relation-specific investment that will increase the likelihood that the product will have the high value. We

148. *See* RESTATEMENT (SECOND) OF CONTRACTS § 311 cmt. a (1979); *see also* Zumwinkel v. Leggett, 345 S.W.2d 89 (Mo. 1961) (holding unenforceable a no-oral-modification clause); Beatty v. Guggenheim Exploration Co., 122 N.E. 378, 381 (N.Y. 1919) (Cardozo, J.) ("Those who make a contract may unmake it. The clause which forbids a change may be changed like any other. . . . Whenever two men contract, no limitation self-imposed can destroy their power to contract again."); Jolls, *supra* note 1 (discussing reasons why parties might want to ban modifications). Civil law codes, on the other hand, discourage renegotiation; these codes thus are more consistent with the argument in the text than with the common law. *See* Eric Brousseau, *Did the Common Law Biased* [sic] *the Economics of Contract . . . and May It Change?*, *in* 6 LAW AND ECONOMICS IN CIVIL LAW COUNTRIES 79, 83-85 (Bruno Deffains & Thierry Kirat eds., 2001).

149. A more extensive formal discussion of the parties' preferences regarding modifications is in Schwartz & Watson, *supra* note 89.

150. *See supra* Section III.B.

assume that the parties can observe the value the product turns out to have, but that values are unverifiable. Also, the buyer cannot observe whether the seller made the investment or not. These information assumptions prevent the parties from inducing efficient investment with a simple fixed-price contract. Such a contract would either pay the seller to invest efficiently or condition the price on the product's realized value. But since neither the seller's behavior nor the product's value could be proven in court (as both values are unverifiable), neither party could enforce this simple contract. There is, however, a more complex contract that would induce the seller to invest efficiently. This contract would require that: (1) the seller make an up-front payment to the buyer that divides the expected surplus from the deal; and (2) after the product's value is realized, the seller make a take-it-or-leave-it offer to the buyer that requires the buyer to pay a price equal to the value (i.e., a high or a low price, depending on whether the value is high or low).

To see how this contract would work, we assume that each party's payoff should the parties not deal—its disagreement payoff—is zero. Further, if the seller invests, at a cost of $c > 0$, the product will have a high value for the buyer with .8 probability. If the seller does not invest, its cost is zero but the product will be certain to have a low value. After the seller invests (or not), it produces the product. We normalize the production cost to zero, so that it is always efficient for the parties to trade. In the example, a high value is $40 and a low value is $20. It would be efficient for the seller to invest if the expected gain less the cost would exceed the low value: i.e., if .8($40) + .2($20) – $c \geq$ $20, or if c is $16 or less. We let c = $12 so that investment would be efficient. The expected surplus under the contract is .8($40) + .2($20) – $12 = $24. Suppose for convenience that the parties have equal bargaining power; then they will divide the expected surplus equally. On these assumptions, the buyer will accept the contract if the seller makes an up-front payment of half the expected surplus, or $12. The buyer's ex post gain would be zero because the trading price would exactly equal the product's value, but the buyer's ex ante gain is the $12 payment. The seller would offer the contract and invest because its net expected gain also would be $12: the $24 expected surplus less the up-front payment of $12.

The difficulty is that following the seller's investment the buyer will pursue a strategy that will reduce the seller's return to below the return it could earn without a contract. To see how, suppose the product turns out to be of low value. The buyer will pay the $20 price and earn the product's value less the price plus the up-front payment, or $20 – $20 + $12 = $12. Now let the product turn out to have a high value. The buyer will reject the seller's take-it-or-leave-it offer, return the up-front payment and offer to renegotiate the deal. At this stage, the parties know that they will earn zero

if they fail to trade and will share $40 if they do trade. Consequently, the parties will renegotiate the contract. Just as in the example in Section III.B, the seller's investment cost is sunk. Our assumptions that the parties have equal bargaining power and engage in Nash bargaining (under which sunk costs are ignored) thus imply that the parties will agree to share the $40 value, so that the new price will be $20.[151] The buyer will reject the contract and renegotiate in the high value state because it will then earn $40 (the product value) – $20 (the new price) = $20 under the renegotiated contract, which exceeds the $12 it would have earned under the original contract.

The buyer's strategy, if anticipated, would cause the seller not to contract, however. To see why, assume that the seller contracts and invests. If product value turns out to be high, the seller will earn the renegotiated $20 price less the $12 investment cost. If product value turns out to be low, the buyer will enforce the contract, so that the seller will not have the right to the return of the up-front payment. In the low-value case, the seller thus will earn the $20 price less the $12 investment cost and less the $12 down payment, for a net loss of $4. On the probabilities assumed above for high and low values, the expected return to the seller from contracting and investing will be $5.60. If the seller instead chooses not to incur the $12 investment cost, the product will have a low value for sure and the seller will earn $20 (the contract price) – $12 (the up-front payment) = $8. If the parties do contract, then, the seller will do better by not investing. The seller, however, will do better still by not contracting at all. Then it also will not invest. Rather, the seller will produce a product with low value at a cost of zero and offer it to the buyer in a spot purchase. The transaction price will split the $20 surplus. The seller thus will earn a net $10 if it does not contract and a net $8 if it does. Therefore, the parties will not contract, the seller will not invest, and product quality will be inefficiently low.

The lesson is that parties sometimes are unable to write contracts that induce efficient relation-specific investment because renegotiation—a contract modification—cannot be banned.[152] An enforceable ban on modifications would permit the seller to reinstate the contract's original price term. A court could observe the original contract, the renegotiated contract, and the monetary transfers the parties made under each. Thus,

151. The seller in this example could not sue for the $40 price because we assume that product values are not verifiable. Thus, the seller could not prove that the value was high.

152. The investment in this example is termed "cooperative" because the seller invests to increase the buyer's value. When renegotiation is permitted, cooperative investment can be impossible to motivate by contract. *See* Yeon-Koo Che & Donald B. Hausch, *Cooperative Investments and the Value of Contracting*, 89 AM. ECON. REV. 125 (1999); Ilya Segal, *Complexity and Renegotiation: A Foundation for Incomplete Contracts*, 66 REV. ECON. STUD. 57 (1999). Our earlier examples involved self-investment (the seller invests to lower her costs; the buyer invests to increase his value). Efficient self-investment is easier to motivate with appropriate contracts. Both kinds of investment are common.

when the buyer returned the $12 up-front payment and paid $20 under a renegotiated contract, the court would know that product quality was high. The court then would order the buyer to pay another $20. Expecting that a court will undo any renegotiation, the buyer would realize that its options were limited to purchasing the low-quality product on the spot market, earning a net $10, or making the contract, accepting whatever ex post take-it-or-leave-it offer the contract permitted the seller to make, and earning a net $12. The buyer thus will make and comply with the contract. Anticipating the buyer's behavior, the seller will agree to contract, and it will invest efficiently. To summarize, parties can have good reasons for banning modifications.[153]

B. *Parties Must Accept Substantial Performance*

Courts will generally require parties to accept substantial rather than full performance unless, in the court's view, the deviation is material.[154] Parties sometimes try to opt out of this substantial-performance default by making full performance an express condition of the promisee's duty to pay. Courts, however, frequently refuse to require exact compliance with the express condition because "the law abhors a forfeiture."[155] The basic premise is that the performing party would not have agreed to a contract that would penalize it severely for minor deviations. Since the paying party would have known this, neither party would have thought that the contract required the equivalent of a forfeiture for a slight nonconformity, in spite of

153. Modification bans also can be efficient when one of the parties is risk-averse. To see why, consider a possible contract between a risk-neutral principal and a risk-averse agent who is supposed to do a task. The principal, it is commonly assumed, cannot verify the agent's behavior to a court, so the contract must motivate the agent. This will require the agent to bear risk; that is, her pay must be contingent on the outcome, so she will try to produce a good result. The agent bears risk because the outcome is a function both of her efforts, which increase the likelihood of a good result, and chance. After the agent acts, but before the result is known, a Pareto-superior deal between the parties becomes possible: Since the agent has acted, she no longer needs to be motivated, but she still bears risk because nature has yet to act. The deal will transfer risk from her to the risk-neutral principal by paying the agent a fixed fee that will lie somewhere between the contract's good- and bad-state payoffs to her. If the parties anticipate this renegotiation, however, they will know that the agent's payoff will not be a function of the outcome; rather, it will be the fixed fee the agent expects to get in the renegotiation. Thus, the agent cannot be motivated to try hard because her payoff actually is noncontingent. When trying hard is efficient, permitting renegotiation thus is inefficient. The parties, ex ante, again have good reasons to prevent themselves from changing their minds. *See* Jolls, *supra* note 1, at 209-24.

154. *See* RESTATEMENT (SECOND) OF CONTRACTS §§ 237, 241 (1981). If the performance is substantial, the default rule is that the promisee must transfer the price less a sum that would compensate it for the deviation from full performance. As we will see below, when the promisee's valuation is not verifiable the right to deduct damages is hollow.

155. *See id*. § 227 cmt. b ("The policy favoring freedom of contract requires that, within broad limits . . . the agreement of the parties should be honored even though forfeiture results. When, however, it is doubtful whether or not the agreement makes an event a condition of an obligor's duty, an interpretation is preferred that will reduce the risk of forfeiture."); *see also* ROBERT E. SCOTT & JODY S. KRAUS, CONTRACT LAW AND THEORY 718-20 (3d ed. 2002).

what the written words appear to say. The parties, courts believe, have no good reason to require forfeitures; thus, courts strictly construe express conditions that require full performance.

We offer an example, drawn from the famous case of *Jacob & Youngs, Inc. v. Kent*,[156] to show that parties have good reasons to construct deals that make forfeitures possible. Consider a contract that requires an owner to make progress payments to a builder as construction progresses. The last payment, which is sizable, is due when construction is completed. We assume that values are unverifiable. Thus, a court cannot observe the value to the owner of a building completed in accordance with the contract. On this assumption, there can be moral hazard on both sides. If the builder has the burden of proof, the owner may claim that defects in the final version reduce her value substantially, even though they do not. The unverifiability of valuations would make it difficult for the builder to disprove this claim. Conversely, if the owner has the burden of proof, the builder may deliberately render a defective performance but claim that, in fact, it had substantially complied.

To understand the potential effect of such moral hazard, assume that the builder has the burden of proof and that the owner will cheat by withholding the entire final payment if the building is less than perfect. Since perfection is difficult to achieve, the builder will expect not to receive the final payment. Therefore, it will not render the final performance. But then the owner will know that the contract's penultimate performance will be the final one, and it will then cheat. The builder, anticipating this, will not render the penultimate performance, and so forth. In equilibrium, therefore, the builder will render no performance at all. The parties' contracting problem thus is to induce the builder to perform when the unverifiability of values makes strategic behavior likely.

The parties' solution follows from the contextual nature of verifiability. A datum of information may not be verifiable to a court because explaining matters to a generalist judge or a lay jury can be costly in relation to the gains. The same datum of information may be verifiable to an arbitrator, however. The arbitrator's expertise makes her cheaper to inform; she acts in an informal setting; and she has a reputational stake in not appearing to be biased in favor of builders or owners. To ensure that parties cooperate with the arbitrator, the parties will make her decision conclusive in the absence of fraud, bias, or mistake.

In *Jacob & Youngs*, the parties had adopted this common solution to the moral hazard problem, making an architect the arbitrator.[157] The architect refused to certify that the builder had fully complied, though the defect

156. 129 N.E. 889 (N.Y. 1921).
157. *See id*. at 890.

appeared trivial.[158] The seeming disjunction between the size of the withheld final payment and the nature of the noncompliance suggested possible fraud or mistake by the architect. The builder, however, did not attempt to impeach the architect's decision. Rather, the builder asked a court to hold that perfect compliance was not a condition to receiving the entire last payment; the court agreed.[159] It believed that forfeiture of the entire last payment would have been unfair and that the parties could not have intended this result.[160] The issue the case posed, however, was not whether the parties had or lacked good reasons to permit forfeitures. Instead, the issue was whether the parties had good reasons to make the architect's findings conclusive. When the parties adopt a sophisticated governance scheme, it is mistaken paternalism for a court to require a promisee to accept substantial performance on the ground that the parties lack good reasons to require a perfect tender.

C. *Parties Cannot Agree to Penalties*

Contracting parties are permitted to specify the damages the breaching promisor must pay, provided that the specified damages represent a reasonable estimate of the promisee's lost expectation.[161] Parties, however, are not permitted to specify damages that exceed a reasonable estimate of the promisee's expectation. The logic here is the same as in the above examples. The parties made the contract in order to give the promisee a particular performance for a price. They know that breaches sometimes occur, so the parties also have good reasons to ensure that the promisee will receive a monetary substitute for performance in the event of breach. But contracting parties supposedly do not have good reasons to award the promisee much more than its lost expectation when the promisor fails to perform. The parties' deep contractual goal is advanced, therefore, by a mandatory rule declaring penalty clauses unenforceable.

This premise fails for two reasons. First, courts will sometimes implement the penalty rule inaccurately. Courts regulate liquidated damages clauses by comparing the difference between anticipated damages under the expectation damages default rule to the stipulated damages in the contract. Expectation damages, in turn, are based on the verifiable losses that the promisee anticipates from breach. Any liquidated damages clause

158. *See id*.
159. *See id*. at 890-91.
160. *See id*.
161. *See* U.C.C. § 2-718(1) (2003) ("Damages for breach by either party may be liquidated in the agreement but only at an amount which is reasonable in the light of the anticipated or actual harm caused by the breach . . . . A term fixing unreasonably large liquidated damages is void as a penalty.").

that incorporates observable but nonverifiable values thus will be vulnerable to a penalty claim even when the clause accurately measures the promisee's lost expectation.[162] Moreover, courts sometimes find compensatory liquidated damages clauses to be penalties in complex cases because the courts have failed to understand just how the clause protected the promisee's expectation.[163] These difficulties with current law permit promisors to invoke the penalty doctrine strategically. As a consequence, sophisticated parties are discouraged from using liquidated damages clauses when these clauses would otherwise be optimal.

The second problem with current law is directly relevant to the argument advanced above. The premise that parties have no good reason to contract for penalties is itself mistaken. Rather, penalties can permit parties to induce efficient relation-specific investments in certain asymmetric information environments. Commercial parties thus have good reasons to agree to penalties in these circumstances. As an illustration, return to the example in Section VI.A that showed the error of preventing parties from writing antimodification clauses in contracts. An alternative way to produce efficient investment in that case would be for the parties to contract with a third party that if either party should request a modification, it would have to pay a large penalty to the third party. This party would have an incentive to enforce the prohibition on modifications in order to collect the penalty. Anticipating such a suit, the buyer in our example would be deterred from requesting renegotiation. These third-party schemes are not seen in practice because penalty terms are unenforceable under current law.[164] Banning a liquidated damages clause because, and only because, it requires the breaching party to make a transfer that exceeds its contract partner's expectation thus wrongly interferes with the parties' sovereignty and may generate inefficiency.[165]

To summarize, contract law contains a number of mandatory rules that apply in the absence of an externality or a market failure. These rules override contractual terms that appear to be inconsistent with the intentions

162. *See* Goetz & Scott, *Liquidated Damages*, *supra* note 1, at 568-76**.**

163. *See* Alan Schwartz, *The Myth That Promisees Prefer Supracompensatory Remedies: An Analysis of Contracting for Damage Measures*, 100 YALE L.J. 369, 383-87 (1990).

164. For a review of the various ways, including the one above, that enforcing penalties can enhance efficiency, see Aaron S. Edlin & Alan Schwartz, *Optimal Penalties in Contracts*, 78 CHI.-KENT L. REV. 33 (2003).

165. *See id*. When the seller has market power, parties may use penalties to deter the entry of competitors into the seller's market. *See id*. at 40-42. Under current law, however, courts strike what they perceive to be penalty terms whether those terms were used to increase investment or to impede entry. We argue here that courts should only ban inefficient penalties; to ban other penalties is only misplaced paternalism. Another way to put this claim is that a party should always be free to argue that a damages term would create a negative externality or perpetuate a market failure—grounds that we maintain would still be sufficient reason for courts to strike such a term. But it is a mistake to treat as a sufficient proxy for these inefficiencies a liquidated damages clause that would overcompensate a party in expectation.

that rational, informed, and uncoerced parties would have under the circumstances. It appears to courts and to the drafters of the UCC that actual parties have no good reasons for choosing these terms, and that a defensible paternalism thus would not enforce them. In the cases we have analyzed, however, commercial parties turn out to have good reasons for the things they do. Put another way, the contract terms that courts and the UCC refuse to enforce actually advance the parties' welfare. A paternalistic justification for contract law's mandatory rules therefore fails. This should not be surprising. In Part II, we showed that commercial parties pursue a goal—joint welfare maximization—that the state supports, and generally can choose the means that best implement this goal. The rules regulating contracts between business firms thus should be mandatory only when the parties' contract creates an externality or is the product of market failure.

VII. CONCLUSION

This Article does not ask the conventional normative question: What contract law should the state provide? Rather, it asks: What contract law do business firms want the state to provide? A contract law for firms, we answer, would be narrower and more deferential to contracting parties than the contract law we now have. Of first-order importance, firms want the state to enforce the contracts that they write, not the contracts that a decisionmaker with a concern for fairness would prefer them to have written. Enforcement of written agreements presupposes a theory of interpretation. The commitment to party sovereignty that we defend in this Article requires courts to delegate to parties both the choice of a contract's substantive terms *and* the choice of the interpretive theory that will be used to enforce those terms. Commercial parties, we show, commonly prefer adjudicators to be accurate on average in ascertaining the meaning of their agreements rather than accurate in every instance; therefore, these parties want courts to make interpretations on the smallest evidentiary bases that will support interpretations that are accurate on average. Courts that defer to party preferences regarding interpretation thus will use a textualist interpretive style, one that restricts the evidentiary base to the written agreement and not much more. In addition, most firms prefer courts to interpret their contracts with the presumption that the contract is written in what we call majority talk, the language that firms and courts usually speak. For this reason, too, a textualist interpretive theory is the best default.

Much of today's contract law is in the form of default rules and standards. These defaults cause more harm than good. An efficient default rule—one that firms will accept—is simple in form, conditioned on few states of the world, and maximizes joint gains in a wide variety of contexts. A default standard is efficient only when parties can live with vague

definitions of their contracting obligations. Because standards confer considerable discretion on parties, a standard will be unsatisfactory if, as a result of that discretion, parties are likely to behave strategically under it. As we show, parties are heterogeneous, drafting costs are finite even for public decisionmakers, rules must sometimes be complex, parties commonly will exploit standards to redistribute rather than to maximize joint surplus, and information often is asymmetric. When the state tries to write efficient default rules and standards, these circumstances create obstacles that the state can seldom overcome. Unsurprisingly, business parties often contract out of these failed defaults. The effective domain of state-supplied contract law thus is smaller than is widely believed.

The welfare-maximization goal that we advance justifies courts in refusing enforcement to unconscionable contracts, contracts affected by fraud or duress, and contracts that create externalities. This goal, however, cannot support many of the mandatory rules that today govern much contracting behavior between firms. These rules bar enforcement to contract terms that efficiently cope with problems of hidden information and hidden action.

A normative theory of contract law that takes party sovereignty seriously shows that much of the expansion of contract law over the last fifty years has been ill-advised. Contract law today is composed of a few default rules, many default standards, and a number of mandatory rules. Most of the mandatory rules should be repealed or reduced to defaults, and most of the defaults should vanish from the law. Advocating freedom of contract for firms is uncontroversial. Taking freedom of contract seriously, however, would radically truncate current contract law. A law merchant appropriate to our time would be a merchants' law; and for merchants, the less publicly supplied law the better.