LAW & ECONOMICS CENTER | 50th 1974–2024

Nashville Regional Tournament Rounds 1 & 2 Readings
Table of Contents:

Round 1

Round 2

BLIND SPOT:
THE ATTENTION ECONOMY AND THE LAW

Human attention, valuable and limited in supply, is a resource. It has become commonplace, especially in the media and technology industries, to speak of an "attention economy" and of competition in "attention markets."[1] There is even an attentional currency, the "basic attention token," which purports to serve as a medium of exchange for user attention.[2] Firms like Facebook and Google, which have emerged as two of the most important firms in the global economy, depend nearly exclusively on attention markets as a business model.[3]

Yet despite the well-recognized commercial importance of attention markets, antitrust and consumer protection authorities have struggled when they encounter the attention economy.[4] Antitrust agencies, tasked with assessing the effects of mergers and controlling anticompetitive behavior, seem to lack a way to understand the market dynamics when the firms offer "free products" that are actually competing for attention.[5] Meanwhile, those tasked with consumer protection have no good paradigm for dealing with attentional intrusions stemming from non-consensual, intrusive advertising or debates over the use of telephones on airlines.

This essay aims to provide a legal and economic analysis to help face the challenges here described. In other work, I have described the rise and spread

---

[3] *See infra* Part II.A.
[4] *See infra* Parts I.A, III.A, III.B.
[5] *See infra* Part III.A.

of the "attention industry," the businesses that depend on the resale of attention, a global industry with an annual revenue of approximately $500 billion.[6] This essay builds on that work by focusing on the economic decisions implicit in "Attention Brokerage." As described here, brokerage is the resale of human attention. It is to attract attention by offering something to the public (entertainment, news, free services, and so on), and then reselling that attention to advertisers for cash. Examples of pure Attention Brokers include social media companies like Instagram and Facebook, search engines like Google or Bing, ad-supported publishers like Buzzfeed or AM News, and some television channels like CBS or NBC. The Brokers' activities are critical to the operation of attention markets, for the business model creates much of the competition for attention that this essay describes.[7]

This approach offers new promise for the antitrust law and some of the challenges it confronts in the attention economy. Markets and market definition are central to contemporary antitrust law, and this essay offers a new approach to the definition of attention markets in cases where enforcers and courts may otherwise become confused by the presence of a "free" product or by two-sided market analysis.[8] It suggests defining the relevant consumer markets based on "time spent" (or just "time") as the currency, and then making use of the familiar economic concept of substitution to find an appropriate market. And so, for example, in a case centered on online mapping products, an enforcement agency may ask whether products like Google Maps, Waze, and Apple Maps are, in fact, substitutes for each other in attention markets. The law can then address appropriate market definition by asking whether other products, like streaming video, compete for the same attention as online maps. That makes possible the use of an "Attentional Small but Significant and Non-Transitory Increase in Price" test, or "A-SSNIPS" test as an aid to finding the appropriate market definition for consumer markets.

The implications of this essay are not merely theoretical. Armed with a better analysis, it would not be too late for an American antitrust agency to challenge some of the relevant acquisitions consummated over the 2010s, like Facebook-Instagram, or Google-Waze, under either Section 7 of the Clayton Act, or Section 2 of the Sherman Act.[9] The passage of years might, in fact, provide clearer evidence of whether such mergers have, in fact, generated

---

[6] *See* TIM WU, THE ATTENTION MERCHANTS: THE EPIC SCRAMBLE TO GET INSIDE OUR HEADS (2016).

[7] *See infra* Part II.

[8] *Cf.* Ohio v. Am. Express Co., 138 S. Ct. 2274 (2018).

[9] *See* United States v. E.I. du Pont de Nemours & Co., 353 U.S. 586 (1957); Standard Oil Co. of N.J. v. United States, 221 U.S. 1 (1911); United States v. Microsoft Corp., 253 F.3d 34 (D.C. Cir. 2001); *see also* C. Scott Hemphill & Tim Wu, *Nascent Competitors*, U. PA. L. REV. (forthcoming 2020).

efficiencies, or instead yielded either higher advertising prices, or increased the ad-load experienced by consumers, or resulted in quality effects, like diminished privacy protections.

In any event, taking attention markets seriously will be essential for agencies confronting a new generation of challenges raised by the importance of the businesses that resell human attention. The goal of this essay is to encourage economists and agencies to develop workable models that help the law face these challenges.

I. BACKGROUND

A. THE BLIND SPOT

We may begin with two legal problems, one from antitrust, one from consumer protection, which demonstrate the challenges that attentional markets create for current laws.

1. *Mergers When Products Are "Free"*

Contemporary antitrust doctrine assumes cash markets where customers spend fiat currency to buy goods or services.[10] Consumer harm, under this approach, is primarily measured in terms of higher prices, reduced output, or other money-related harms. This approach has a number of known limitations.[11] But a particularly problematic blind spot—one that has become prominent over the last decade—emerges when the companies in question seem to give away their products for "free" and are, in fact, competing in attention markets.[12]

---

[10] *See infra* Part III.A.1. *Cf.* SCM Corp. v. Xerox Corp., 645 F.2d 1195, 1206 (2d Cir. 1981) (rejecting the argument that the threat of antitrust liability could "attach upon the acquisition of a patent at a time prior to the existence of the relevant market and, even more disconcerting, at a time prior to the commercialization of the patented art").

[11] Neglect of innovation-related harms and effects on product quality are two prominent limitations. *See* OECD, *The Role and Measurement of Quality in Competition Analysis*, DAF/COMP (2013); Mark McMillan, *Quality Collusion: News, If It Ain't Broke, Why Fix It?*, 39 FORDHAM URB. L.J. 1895 (2012); *see generally* Tim Wu, *Taking Innovation Seriously: Antitrust Enforcement If Innovation Mattered Most*, 78 ANTITRUST L.J. 313 (2012).

[12] This essay is not the only work to notice the problem with merger review in this area. Some scholars have argued that there has been a failure to understand the importance of big data. *See, e.g.*, MAURICE E. STUCKE & ALLEN P. GRUNES, BIG DATA AND COMPETITION POLICY 107–26 (2016). Others, like Newman, have focused on the problem of "zero price" markets. *See* John M. Newman, *Antitrust in Zero-Price Markets: Foundations*, 164 U. PA. L. REV. 149 (2015) [hereinafter Newman, *Foundations*]; John M. Newman, *Antitrust in Zero-Price Markets: Applications*, 94 WASH. U. L. REV. 49, (2016) [hereinafter Newman, *Applications*]; *see also* Michal S. Gal & Daniel L. Rubinfeld, *The Hidden Costs of Free Goods: Implications for Antitrust Enforcement*, 80 ANTITRUST L.J. 521 (2016).

B. THE PROTECTION OF CAPTIVE AUDIENCES

> I tremble for the sanity of a society that talks, on the level of abstract principle, of the precious integrity of the individual mind, and all the while, on the level of concrete fact, forces the individual mind to spend a good part of every day under bombardment with whatever some crowd of promoters want to throw at it.[132]
>
> – Charles Black, 1953

The ubiquity of screens and devices might be the first thing that a visitor from another decade would notice about our times. Whether carried in our hands, or found in a taxi, elevator, schoolhouse, waiting room, and many other places, these screens can be hard to escape. And while most of our usage of screens or phones is consensual—indeed a mainstay of socializing and work—some is not, yielding new regulatory challenges.

This Part focuses on problems posed by the seizure of attention that is *non-consensual*. Such attentional seizure can happen in several ways. First, over the last decade, non-consensual advertising has increased in volume and revenue thanks to the decreased prices of screens. Consider, for example, the spread of "Gas Station TV"—that is, the televisions embedded in gasoline pumps that bring advertising to the captive pumper as he or she gets gas.[133] As

---

[129] PHILLIP AREEDA ET AL., ANTITRUST ANALYSIS 5–10 (7th ed. 2013) (describing harms intended to be prevented by antitrust).

[130] *See, e.g.*, Bhattacharya, *supra* note 127; Adam Levy, *Hate Ads on Facebook? They're About to Get Worse*, NEWSWEEK (Feb. 19, 2017), www.newsweek.com/hate-ads-facebook-getting-worse-557630.

[131] *See* Wu, *Taking Innovation Seriously*, *supra* note 109, at 313.

[132] Charles L. Black, Jr., *He Cannot Choose but Hear: The Plight of the Captive Auditor*, 53 COLUM. L. REV. 960, 962 (1953).

[133] *See* GSTV, gstv.com.

the CEO of Gas Station TV puts it, "We like to say you're tied to that screen with an 8-foot rubber hose for about five minutes."[134] Second, as described earlier, agencies face situations where people are in enclosed, regulated spaces for long periods of times, like in airplanes, and must grapple with the rules for such situations.

Let us discuss non-consensual advertising first. From the neurological research discussed above, it is clear that advertisements, when they use motion and sound, are extremely difficult if not impossible to ignore due to the involuntary responses of the brain.[135] It is true that one might argue that we are responsible for placing ourselves in a position to be exposed. That is why the most clearly non-consensual taking of attention happens when we are in situations where we are compelled to be there and are unable to escape—in the position that First Amendment jurisprudence refers to as being a "captive audience." The captive audience doctrine originated in the late 1940s in cases like *Kovacs v. Cooper*, which concerned a city ban on sound trucks that drove around broadcasting various messages at loud volume so as to reach both pedestrians and people within their homes.[136] The Court wrote that "[t]he unwilling listener is not like the passer-by who may be offered a pamphlet in the street but cannot be made to take it. . . . [H]e is practically helpless to escape this interference with his privacy by loud speakers except through the protection of the municipality."[137] There are situations—on airplanes, at the gas pump, and so on—where one is simply incapable of escaping, as Black put it, "whatever some crowd of promoters want to throw at it."[138]

It seems undeniable that at least some of the time we end up captive audiences, involuntarily subjected to sounds and images that we cannot ignore. But does this really amount to more than a mere annoyance? It might be protested that seconds, or at most minutes of seized attention are matters too trivial to be taken seriously by the law—*de minimis non curat lex.*

Charles Black, addressing this very point—that non-consensual advertising was nothing more than "a bit of a fuss about nothing"—answered:

> Subjecting a man, willy-nilly and day after day, to intellectual forced-feeding on trivial fare, is not itself a trivial matter; to insist, by the effective gesture of coercion, that a man's right to dispose of his own faculties stops short of the interest of another in forcing him to endure paid-up banality, is

---

[134] *See* Tom Walsh, *Gas Station TV Moves Pumps More Growth into Downtown*, DETROIT FREE PRESS (Apr. 12, 2015), www.freep.com/story/money/business/columnists/tom-walsh/2015/04/11/gasstation-tv-gilbert/25527645/.

[135] *See supra* text accompanying notes 57 & 58.

[136] *See* Kovacs v. Cooper, 336 U.S. 77 (1949).

[137] *Id.* at 86–87.

[138] Black, *supra* note 132.

> not itself banal, but rather a sinister symbol of relative weighting of the independence of the mind of man and the lust to make a buck.[139]

In other words, it really is an unfortunate thing to accept the idea that a seizure of time and mental facilities is a *de minimis* triviality. Sure, we may waste plenty of time as it is, but at least it is ours to waste. It is, in this sense, and as Black later argued,[140] the deprivation of a liberty, more precisely liberty of thought, which is itself a constitutional value. As Justice Cardozo put it, "Of that freedom one may say that it is the matrix, the indispensable condition, of nearly every other form of freedom."[141]

But we might still insist on some more concrete demonstration of harm—some measure of value lost. In this respect, as in the antitrust discussion, it helps to look at what the industry itself thinks. For when looking not at individual effect, but the aggregate effect, we find that the attention taken, without anything in exchange, is sometimes quite valuable. We may observe, for example, that the non-consensual screen advertising ("digital out of home" in the jargon) claims some $12 billion per year in revenue.[142] We also might also look at individual deals—if Spirit Airlines earns a sum per flight by playing unavoidable in-seat advertising without the consent of its passengers, this provides a clear measure of the value of the attention which has been appropriated.

Measured in more concrete terms it becomes easier to accept the concept known as "attentional theft."[143] A typical definition of theft or larceny is the taking of control of property or a resource "under such circumstances as to acquire the major portion of its economic value or benefit."[144] Here, the time taken has its economic value lost to the consumers and transferred to the airline for resale. Conceptual support comes also from the common-law tort of conversion, which is currently defined by the *Restatement* as an intentional act of "dominion or control over a chattel which so seriously interferes with the right of another to control it that the actor may justly be required to pay the other the full value of the chattel."[145] In recent years, a number of courts have expanded the concept of conversion to reach purely intangible property or resources, like domain names, electronic computer data, and personal

---

[139] *Id.*

[140] *See id.* at 963.

[141] Palko v. Connecticut, 302 U.S. 319, 327 (1937).

[142] *See* Press Release, Cision PRweb (PQ Media), Global out-of-Home Media Revenues Up 6.2% in 2016, (Jan. 10, 2017) www.prweb.com/releases/2016/12/prweb13943072.htm.

[143] *See* Tim Wu, *The Crisis of Attention Theft—Ads that Steal Your Time for Nothing in Return*, WIRED (Apr. 14, 2017), www.wired.com/2017/04/forcing-ads-captive-audience-attention-theft-crime/.

[144] N.Y. PENAL LAW § 155.00 (McKinney 2000).

[145] RESTATEMENT (SECOND) OF TORTS § 222A(1) (AM. LAW INST. 1965).

records.[146] The point is not that anyone is actually criminally liable, or that a tort action would actually lie. It is, rather, that the underlying concept of harm may support the adoption of policy designed to protect the resource.

Nor should this essay be taken as suggesting that all display advertising—posters, television, and so on—be barred as a seizure of attention. Instead, it only suggests that the consent concept should be key for regulators who want to protect consumers from non-consensual attentional theft. Consent models are, of course, commonplace among legal regimes that regulate social interactions. For example, when it comes to physical or sexual contact, consent is usually a defense to a battery or rape charge.[147] Implicit consent forms an obvious model for thinking about intrusions upon attention, or more technically, upon the brain.

When it comes to attentional theft, we should construe implicit consent to apply broadly to most social contexts and exchanges. By walking down the street, one should be understood to be necessarily consenting to seeing advertising posters and all manner of other involuntary attentional intrusions, from the sirens on fire trucks, to other people, to fast-moving cars. And, as discussed earlier, we regularly agree to sell our attention by, for example, agreeing to watch advertising in exchange for "free" television programming. Thus, our analysis should pay special attention to context and focus on situations where one becomes a captive audience.

The approach explored here also gives lawmakers and regulators a means of thinking about rules for regulated environments, like airplanes, where passengers may subject each other to loud annoyances, like phone conversations. For most of us, it is not hard, intuitively, to see why it might be annoying to spend an entire flight next to a person who has chosen to spend the flight speaking loudly on a telephone. But as we've seen above, consumer protection regulators, focused on the legal standard of policing "unfair and deceptive practices,"[148] have struggled to find a framework to give consumers the protections that they obviously want.

I'd suggest the following principles that local, state, or federal lawmakers or regulators should keep in mind:

---

[146] *See, e.g.*, Thyroff v. Nationwide Mut. Ins. Co., 8 N.Y.3d 283 (2007); Kremen v. Cohen, 337 F.3d 1024 (9th Cir. 2003) (applying California law); Superior Edge, Inc. v. Monsanto Co., 44 F. Supp. 3d 890 (D. Minn. 2014) (applying Missouri law); *In re* Yazoo Pipeline Co., L.P., 459 B.R. 636 (Bankr. S.D. Tex. 2011) (applying Texas law).

[147] *See, e.g.*, N.Y. PENAL LAW §§ 120.00, 130.05, 130.10 (McKinney 2000).

[148] *See, e.g.*, 49 U.S.C. § 41712 (empowering the Department of Transportation to investigate unfair or deceptive practices or unfair methods of competition); 15 U.S.C. § 45 (empowering the Federal Trade Commission to prevent the use of unfair or deceptive practices or unfair methods of competition).

- Attention is a valuable resource belonging to consumers to be spent or allocated as they wish.
- Regulators should protect consumers from non-consensual and entirely uncompensated transfers of attentional resources.
- While in many environments, attention is freely given, regulators should protect consumers when they are captive audiences, unable to make a meaningful choice about how their attention is spent.
- Regulators should also protect citizens from intrusions that are unavoidable due to high volume, bright flashing lights, and so on, especially in their homes.

These principles might, for example, lead a locality or state legislature to ban the uncompensated transfer from a captive audience implicit in Gas-Station TV, unless, perhaps, a discount on gas is offered in exchange for being exposed to the advertising. The approach also helps justify laws already in existence that protect the public against attentional seizure. Most municipalities, for example, have regimes that govern excessive noise, which protect people from noise while they are captive in their homes. For example, the Los Angeles Municipal Code has a "Noise Regulation" section that prohibits "unnecessary, excessive and annoying noises from all sources."[149] The city has enacted complex regulations that, for example, require miniature golf operators to post signs "[r]equesting patrons to refrain from unnecessary noise."[150] New York City, surely one of the noisier cities in the world, has a complex set of noise regulations,[151] such as an (unenforced) ban on honking in situations other than in response to actual dangers, and a ban on some types of car alarms.[152] Federal and local governments regulate highway billboards, and many ban digital billboards—billboards that change shape and sometimes flash in order to gain attention.[153]

These various laws were not, of course, enacted to protect attentional resources as discussed in this essay, but for aesthetic reasons, or occasionally

---

[149] L.A., CAL., MUN. CODE ch. 11 (1973).

[150] *See id.* art. 4 § 41.44(c).

[151] Many of which are not fully enforced. *See* Matt Flegenheimer, *Stop the Honking? New York Suggests It's a Lost Cause*, N.Y. TIMES (Jan. 28, 2013), www.nytimes.com/2013/01/29/nyregion/new-york-removes-no-honking-signs.html.

[152] In 2004, New York City's City Council passed a more comprehensive ban on the sale of car alarms, but the measure was vetoed. *See* Winnie Hu, *City Council Ban Car Alarm Sales, but a Mayoral Veto Looms*, N.Y. TIMES (July 22, 2004), www.nytimes.com/2004/07/22/nyregion/city-council-bans-car-alarm-sales-but-a-mayoral-veto-looms.html?_r=0.

[153] *See, e.g.*, Scenic Arizona v. City of Phoenix Bd. of Adjustment, 268 P.3d 370 (Ariz. Ct. App. 2011).

SHAREHOLDER WEALTH MAXIMIZATION: A SCHELLING POINT

I. INTRODUCTION

Imagine a reality television game show where two contestants begin the game in two different places in New York City.[1] The object of the game is for the two contestants to find each other, but they do not know anything about each other and they have no way of communicating. If they succeed, both contestants win a prize. If they fail, they get nothing. With no ability to explicitly bargain over the meeting, the parties have to make an educated guess about what the other person is most likely to do. Most people, confronted with this sort of tacit coordination game,[2] will attempt the meeting at a major New York City landmark such as the Empire State Building.[3] Absent any other clues as to the optimal equilibrium meeting point, both parties choose a place that is imaginatively unique and intuitive, expecting that the place will also be unique in the other's imagination.[4] The Empire State Building stands out not because it is a particularly optimal meeting place, but rather because it is iconic, nearly syn-

---

[4] SCHELLING, *supra* note 1, at 54.

onymous with New York City itself. This is called a "focal point," or "Schelling point," after Professor Thomas Schelling.

There are two important observations that arise from the New York City game: first, that people can coordinate without communication and, second, that value-creating outcomes[5] can be achieved despite multiple equilibria[6] and high transaction costs.[7] As to the former, the fact that many more people than would be expected by chance would likely collect the prize illustrates that coordination without communication is possible.[8]

The latter is more interesting because most real-life coordination occurs with some communication. Thus, logic might suggest that if communication were permitted, the parties would simply bargain to an optimal solution.[9] It could be that one contestant started at Columbia University and the other at Arthur Ashe Stadium, the optimal meeting point was actually equidistant between the two, and the optimal meeting time was around 5:18 a.m., when traffic on the streets was at the lowest.[10] But, even in explicit bargaining contexts, the bargainers still, far more often than logic would dictate, fall back on logically irrelevant but imaginatively simple solutions.[11]

Part of the reason for this is that real-life coordination, perhaps unlike the New York game, involves some degree of embed-

---

[5] "Value-creating" is the term used here because, importantly, focal points do not always lead to optimal outcomes. Law-and-economics, like economics, has long concerned itself with finding the optimal. Some definitions are formal statements of economic efficiency (e.g., Pareto efficient, Kaldor-Hicks efficient), while others are intuitive (e.g., appropriate level of corporate manager attention to voluntary efforts at protecting the environment). Similarly, some argue that optimality is derived from measures of utility or wealth, or that free exchange defines and begets utility and wealth. *See, e.g.*, FRANK H. EASTERBROOK & DANIEL R. FISCHEL, THE ECONOMIC STRUCTURE OF CORPORATE LAW 6–7 (1996).

[6] In the New York game, *any* meeting place could be a "winning" equilibrium.

[7] *See* N. Gregory Mankiw, *New Keynesian Economics*, *in* THE CONCISE ENCYCLOPEDIA OF ECONOMICS 379, 381 (David R. Henderson ed., 2008). Here, transaction costs are the cost of negotiating a meeting place if communication were available. *See* Robert Cooter, *The Cost of Coase*, 11 J. LEGAL STUD. 1, 20–21, 23 (1982).

[8] *E.g.*, SCHELLING, *supra* note 1, at 114 (describing the situation where two spouses lose track of each other in a store and must coordinate to find each other and go home).

[9] *See, e.g.*, *id.* at 50.

[10] *See* Ben Wellington, *Quantifying the Best and Worst Times of Day To Hit the Road in NYC*, I QUANT NY (Aug. 5, 2014, 12:34 AM), https://iquantny.tumblr.com/post/93845043909/quantifying-the-best-and-worst-times-of-day-to-hit [https://perma.cc/6Q5D-AN4K] (finding that the lowest-traffic time of day in New York City is 5:18 a.m., which is when taxis were able to maintain the highest average speed).

[11] SCHELLING, *supra* note 1, at 67.

ded conflict.[12] Both sides want to get the best deal possible for themselves which, while the transaction as a whole will make both sides better off if a deal is reached, does not mean there is not a conflict over who gets the greater share of the surplus created by the transaction. That conflict, in turn, makes the commitments proposed in the explicit bargain less than credible, as they might be made out of strategic motive to capture more of the surplus rather than one of facilitating the exchange. In a sense, this diminishment in credibility devolves the explicit bargaining back into tacit bargaining, because communication that lacks credibility is likely to lead to the same place as not communicating at all: settling for a Schelling point or not making a deal.[13]

The other reason is that discovering and implementing these optimizations is costly. Once the bargain reaches the stage where continuing to optimize gets too costly—either strategically or otherwise—the parties can, and often do, choose the logically irrelevant and perhaps suboptimal, yet contextually intuitive, solution to break the strategic deadlock or forgo the additional costs.[14]

Finally, a word on what makes a point contextually intuitive. One way is physical or geographic uniqueness—the Empire State Building is very tall and associated in the popular mind with New York City. Another way is that once bargainers agree on a Schelling point as the way to conclude their bargain, they will return to it in future bargains—again, to avoid costs.[15]

These, then, are the primary features of Schelling points: they permit coordination in high-transaction-cost, multiple-equilibrium environments because they are contextually intuitive. They are value-creating, though not always optimal—that is, another equilibrium might be more optimal by a given measure, but the one that is contextually intuitive makes coordination easier and more likely. They may be contextually intuitive due to some characteristic—for example, the Empire State Building's association with New York City—or due to having been productively used to coordinate people in prior dealings, or, often, both.[16]

---

[12] *See generally* Robert B. Ahdieh, *The Strategy of Boilerplate*, 104 MICH. L. REV. 1033 (2006). Of course, one party could always push the other to come to her location or a location more convenient for her.

[13] *See* SCHELLING, *supra* note 1, at 67; *see also* David Friedman, *A Positive Account of Property Rights*, 11 SOC. PHIL. & POL'Y 1, 13 (1994).

[14] *See* SCHELLING, *supra* note 1, at 67.

[15] Friedman, *supra* note 13.

[16] *See infra* Sections IV.B.1–2.

A Schelling point is an intuitive, if imperfect, place for bargainers to conclude a bargain where a suboptimal bargain is superior to no bargain at all.

Shareholder wealth maximization is the norm within corporate law and governance holding that directors' and officers' roles are, fundamentally, to maximize the long-term value of the corporation, and thus the value of its shares of stock.[17] The debate over whether it is or should be the norm or the law for corporations is usually contested on different theories of economic or social value, and, of course, legal precedent.[18] The extant law-and-economics theories of corporate law view shareholder wealth maximization as an efficient term in a hypothetical bargain between corporate managers and directors on one side and shareholders on the other side.[19] The argument goes that it has, by reason of this efficiency, emerged as the law.[20] Competing theories rest on a rejection of this "contractarian" theory of the firm due to empirical results that run counter to the theory, persistent counterfactuals that undermine the norm's existence, and scant hard precedent for the norm.[21] At times in the wide-ranging academic and public debate, commentators have called for dividing managerial focus among wealth maximization and other matters such as worker wealth, environmental sustainability, and other nonfinancial social concerns.[22]

---

[17] STEPHEN M. BAINBRIDGE, CORPORATE LAW 245 (3d ed. 2015); PRINCIPLES OF CORPORATE GOVERNANCE: ANALYSIS AND RECOMMENDATIONS § 2.01 (AM. L. INST. 2008). *But see* Eric Franklin Amarante, *What We Talk About When We Talk About Shareholder Wealth Maximization*, 19 TRANSACTIONS 455, 459–62 (2017) (illustrating the complexity buried just within the words "shareholder," "wealth," and "maximization"). One of the myriad debates that involves shareholder wealth maximization is about what it actually means. *See id.* at 459.

[18] Lynn A. Stout*, Bad and Not-So-Bad Arguments for Shareholder Primacy*, 75 S. CAL. L. REV. 1189, 1192, 1199–1202 (2002).

[19] Jonathan R. Macey, *A Close Read of an Excellent Commentary on* Dodge v. Ford, 3 VA. L. & BUS. REV. 177, 185 (2008).

[20] EASTERBROOK & FISCHEL, *supra* note 5, at 120–21.

[21] *See* Stephen M. Bainbridge, *In Defense of the Shareholder Wealth Maximization Norm: A Reply to Professor Green*, 50 WASH. & LEE L. REV. 1423, 1428 (1993). *See infra* Section II.A; Stout, *supra* note 18, at 1192–93, 1208 (stating that ownership and residual claimant theories are "empirically incorrect" and "false"); Ann M. Lipton, *What We Talk About When We Talk About Shareholder Primacy*, 69 CASE W. RSRV. L. REV. 863, 868–69 (2019) (proposing the different hypothetical desires of shareholders that may be in conflict with increasing the value of the residual claim).

[22] Ronald M. Green, *Shareholders as Stakeholders: Changing Metaphors of Corporate Governance*, 50 WASH. & LEE L. REV. 1409, 1411 (1993).

This Article, purposely, does not stake a position on whether shareholder wealth maximization is socially optimal, or whether expanding corporate governance to include greater focus on other matters would be better. Instead, it reflects on the role the norm plays in coordinating people and money. It strives to illustrate that shareholder wealth maximization "works" because it is a contextually intuitive way to coordinate within the high-cost, multiple-equilibrium bargaining environment associated with large public corporations.

Capital is necessary to run a corporation and generally, corporations cannot obtain enough capital from debt-financing sources. For that reason, they must sell stock or equity. That arrangement, like all the other contracts that make up the corporation, must consist of some kind of bargain. This Article proposes that selling stock or equity is an arrangement of the kind Professor Schelling describes—a coordination game with embedded conflict.[23] Given the practical impediments to actual bargaining, the value of coordinating investment and resolving of conflicts between the shareholders and the managers and directors by settling on shareholder wealth maximization is reflected in the fact that this bargain is so often struck. One way to describe this process is as Professor Schelling did—that there was something contextually intuitive about the idea of maximizing shareholder wealth.

Everyone who might want to invest in any corporation or run one understands, in general, what it means for a corporation to orient itself toward maximizing shareholder wealth.[24] The key observation is that the shareholders recognize, generally, that managers and directors are going to be trying to maximize wealth. This recognition is independent of any individual shareholder's actual desires or purposes, regardless of whether those desires or purposes are advanced or frustrated by the norm—that is, even if shareholders wanted managers and directors to do something else, they generally know that managers and directors

---

[23] *See also* Manuel A. Utset, *Towards a Bargaining Theory of the Firm*, 80 CORNELL L. REV. 540, 582 (1995) (describing the bargaining between individual shareholders and the managers as a tacit bargain of the kind Schelling analyzed).

[24] This is not to say that every person's actual decisions in any given transaction or case will be in alignment. It just means that the concept of focusing on maximizing corporate wealth is relatively simple. *Cf.* Lipton, *supra* note 21, at 891–92.

will be trying to maximize wealth.[25] This sheds some light on why it might have emerged as the norm for animating director and officer conduct in the beginning, why it has been enshrined in the law in Delaware and elsewhere, and why it continues as a norm today. In sum, I hope this Article answers the question: Why is shareholder wealth maximization so persistent, even granting its inefficiencies, shortcomings, and even granting its occasional failure to result in anyone's wealth being maximized? The posited answer is that shareholder wealth maximization's contextual intuitiveness has resulted in its becoming a load-bearing part of corporate governance. It is persistent for its usefulness, if not always preferred for its optimality. In the hypothetical bargain between the managers and directors on one side, and the shareholders on the other, shareholder wealth maximization represents a value-creating, if not always optimal, equilibrium.

It is important to add that the public is taking a more keen interest in the role of corporations in society.[26] Many view shareholder wealth maximization as socially harmful, or, at minimum, a powerful impediment to social progress or change.[27] While describing shareholder wealth maximization as a Schelling point is not the sort of idea that would necessarily respond to calls for reform premised upon the need for social change, understanding its persistence should be a part of the debates to come. There is no doubt that the economics, social value, and legal doctrine will be up for debate over the next few years, but the debate will be incomplete without a full account of the norm's coordinating function. This Article, hopefully, contributes part of such an account.

This Article proceeds in five parts. Following this Introduction, Part II provides a brief overview of the extant theories of the firm, the way these theories interact with existing doctrines,

---

[25] *Cf.* SCHELLING, *supra* note 1, at 58–59 (discussing how one person would "grimly acknowledge" that he or she must go to the other if the only location that both recognize is where the other person is located); Lipton, *supra* note 21, at 867.

[26] David Gelles & David Yaffe-Bellany, *Shareholder Value Is No Longer Everything, Top C.E.O.s Say*, N.Y. TIMES (Aug. 19, 2019), https://www.nytimes.com/2019/08/19/business/business-roundtable-ceos-corporations.html [https://perma.cc/3SDT-JJKU].

[27] *See generally* Addisu Lashitew, *Building a Stakeholder Economy*, BROOKINGS (Oct. 28, 2020) https://www.brookings.edu/blog/future-development/2020/10/28/building-a-stakeholder-economy/ [https://perma.cc/PG75-FAC7].

quality, maximizes firm profitability, and decreases price. Thus, it seems clear that consumers are the prime beneficiaries of advertising.

Production and sale of goods and services is a complicated and competitive process. Specialization creates economic efficiencies derived from dividing the manufacturing, distribution, and retail sales functions among separate firms. Vertical relationships are arrangements down the chain of distribution—from input suppliers to manufacturers, to wholesale distributors, to retailers. A typical vertical relationship is organized in the following way: A bicycle manufacturer recognizes that certain efficiencies can be gained by selling its product to wholesale distributors (who, in turn, sell the bikes to local retailers) rather than performing all of these functions itself.

For the manufacturer, however, there is a fundamental danger in this type of vertical relationship. Specifically, the bicycle manufacturer wants to sell the profit-maximizing quantity of bikes; however, it does not control the retail sales function. The manufacturer wants its retailers to engage in aggressive competitive behavior geared (no pun intended) toward selling its bikes. Retailers can take several actions that affect the sales of the manufacturer's bikes. Obviously, the price set by the retailer will affect sales. Further, the retailer can promote the manufacturer's bikes on its web site, it can place the bikes prominently in the store, and train employees to provide information about the bikes or to promote the bikes over other brands.

Because retailers and producers are different entities, however, they have different objectives; retailers do not necessarily have the same incentive—a desire to maximize the sale of the manufacturer's product. As a result, manufacturers have resorted to numerous contractual devices to encourage their retailers to promote their products, for example, limiting retail pricing discretion or requiring certain types of point-of-sale service. The case below involves an antitrust challenge to once such contractual limitation—resale price maintenance.

Leegin Creative Leather Products, Inc. v. PSKS, Inc.

Supreme Court of the United States
551 U.S. 877 (2007)

JUSTICE KENNEDY delivered the opinion of the Court

In Dr. Miles Medical Co. v. John D. Park & Sons Co., 220 U.S. 373 (1911), the Court established the rule that it is *per se* illegal under § 1 of the Sherman Act for a manufacturer to agree with its distributor to set the minimum price the distributor can charge for the manufacturer's goods. The question presented by the instant case is whether the Court should overrule the *per se* rule and allow resale price maintenance agreements to be judged by the rule of reason, the usual standard applied to determine if there is a violation of § 1. The Court has abandoned the rule of *per se* illegality for other vertical restraints a manufacturer imposes on its distributors. Respected economic analysts, furthermore, conclude that vertical price restraints can have procompetitive effects. We now hold that *Dr. Miles* should be overruled and that vertical price restraints are to be judged by the rule of reason.

I

Petitioner, Leegin Creative Leather Products, Inc. (Leegin), designs, manufactures, and distributes leather goods and accessories. In 1991, Leegin began to sell belts under the brand name "Brighton." The Brighton brand has now expanded into a variety of women's fashion accessories. It is sold across the United States in over 5,000 retail es-

tablishments, for the most part independent, small boutiques and specialty stores. Leegin's president, Jerry Kohl, also has an interest in about 70 stores that sell Brighton products. Leegin asserts that, at least for its products, small retailers treat customers better, provide customers more services, and make their shopping experience more satisfactory than do larger, often impersonal retailers. Kohl explained: "[W]e want the consumers to get a different experience than they get in Sam's Club or in Wal-Mart. And you can't get that kind of experience or support or customer service from a store like Wal-Mart."

Respondent, PSKS, Inc. (PSKS), operates Kay's Kloset, a women's apparel store in Lewisville, Texas. Kay's Kloset buys from about 75 different manufacturers and at one time sold the Brighton brand. It first started purchasing Brighton goods from Leegin in 1995. Once it began selling the brand, the store promoted Brighton. For example, it ran Brighton advertisements and had Brighton days in the store. Kay's Kloset became the destination retailer in the area to buy Brighton products. Brighton was the store's most important brand and once accounted for 40 to 50 percent of its profits.

In 1997, Leegin instituted the "Brighton Retail Pricing and Promotion Policy." Following the policy, Leegin refused to sell to retailers that discounted Brighton goods below suggested prices.

* * *

Leegin adopted the policy to give its retailers sufficient margins to provide customers the service central to its distribution strategy. It also expressed concern that discounting harmed Brighton's brand image and reputation.

* * *

In December 2002, Leegin discovered Kay's Kloset had been marking down Brighton's entire line by 20 percent. Kay's Kloset contended it placed Brighton products on sale to compete with nearby retailers who also were undercutting Leegin's suggested prices. Leegin, nonetheless, requested that Kay's Kloset cease discounting. Its request refused, Leegin stopped selling to the store. The loss of the Brighton brand had a considerable negative impact on the store's revenue from sales.

PSKS sued Leegin in the United States District Court for the Eastern District of Texas. It alleged, among other claims, that Leegin had violated the antitrust laws by "enter[ing] into agreements with retailers to charge only those prices fixed by Leegin." Leegin planned to introduce expert testimony describing the procompetitive effects of its pricing policy. The District Court excluded the testimony, relying on the *per se* rule established by *Dr. Miles*.

* * *

The jury agreed with PSKS and awarded it $1.2 million. Pursuant to 15 U.S.C. § 15(a), the District Court trebled the damages and reimbursed PSKS for its attorney's fees and costs. It entered judgment against Leegin in the amount of $3,975,000.80.

The Court of Appeals for the Fifth Circuit affirmed. On appeal Leegin did not dispute that it had entered into vertical price-fixing agreements with its retailers. Rather, it contended that the rule of reason should have applied to those agreements.

* * *

We granted certiorari to determine whether vertical minimum resale price maintenance agreements should continue to be treated as *per se* unlawful.

* * *

A

Though each side of the debate can find sources to support its position, it suffices to say here that economics literature is replete with procompetitive justifications for a manufacturer's use of resale price maintenance. See, *e.g.*, Brief for Economists as *Amici Curiae* 16 ("In the theoretical literature, it is essentially undisputed that minimum [resale price maintenance] can have procompetitive effects and that under a variety of market conditions it is unlikely to have anticompetitive effects"); Brief for United States as *Amicus Curiae* 9 ("[T]here is a widespread consensus that permitting a manufacturer to control the price at which its goods are sold may promote *inter*brand competition and consumer welfare in a variety of ways"); ABA Section of Antitrust Law, Antitrust Law and Economics of Product Distribution 76 (2006) ("[T]he bulk of the economic literature on [resale price maintenance] suggests that [it] is more likely to be used to enhance efficiency than for anticompetitive purposes"); *see also* H. Hovenkamp, The Antitrust Enterprise: Principle and Execution 184–191 (2005) (hereinafter Hovenkamp); R. Bork, The Antitrust Paradox 288–291 (1978) (hereinafter Bork). Even those more skeptical of resale price maintenance acknowledge it can have procompetitive effects. See, *e.g.,* Brief for William S. Comanor et al. as *Amici Curiae* 3 ("[G]iven [the] diversity of effects [of resale price maintenance], one could reasonably take the position that a *rule of reason* rather than a *per se* approach is warranted"); F. Scherer & D. Ross, Industrial Market Structure and Economic Performance 558 (3d ed.1990) (hereinafter Scherer & Ross) ("The overall balance between benefits and costs [of resale price maintenance] is probably close").

* * *

The justifications for vertical price restraints are similar to those for other vertical restraints. Minimum resale price maintenance can stimulate interbrand competition — the competition among manufacturers selling different brands of the same type of product — by reducing intrabrand competition — the competition among retailers selling the same brand. The promotion of interbrand competition is important because "the primary purpose of the antitrust laws is to protect [this type of] competition." [*State Oil Co. v.*] *Khan,* 522 U.S. [3,] 15 [(1997)]. A single manufacturer's use of vertical price restraints tends to eliminate intrabrand price competition; this in turn encourages retailers to invest in tangible or intangible services or promotional efforts that aid the manufacturer's position as against rival manufacturers. Resale price maintenance also has the potential to give consumers more options so that they can choose among low-price, low-service brands; high-price, high-service brands; and brands that fall in between.

Absent vertical price restraints, the retail services that enhance interbrand competition might be underprovided. This is because discounting retailers can free ride on retailers who furnish services and then capture some of the increased demand those services generate. Consumers might learn, for example, about the benefits of a manufacturer's product from a retailer that invests in fine showrooms, offers product demonstrations, or hires and trains knowledgeable employees. R. Posner, Antitrust Law 172–173 (2d ed.2001) (hereinafter Posner). Or consumers might decide to buy the product because they see it in a retail establishment that has a reputation for selling high-quality merchandise. Marvel & McCafferty, *Resale Price Maintenance and Quality Certification*, 15 Rand J. Econ. 346, 347–349 (1984). If the consumer can then buy the product from a retailer that discounts because it has not spent capital providing services or developing a quality reputation, the high-service retailer will lose sales to the discounter, forcing it to cut back its services to a level lower than consumers would otherwise prefer. Minimum resale price maintenance alleviates the problem because it prevents the discounter from undercutting

the service provider. With price competition decreased, the manufacturer's retailers compete among themselves over services.

* * *

Resale price maintenance can also increase interbrand competition by encouraging retailer services that would not be provided even absent free riding. It may be difficult and inefficient for a manufacturer to make and enforce a contract with a retailer specifying the different services the retailer must perform. Offering the retailer a guaranteed margin and threatening termination if it does not live up to expectations may be the most efficient way to expand the manufacturer's market share by inducing the retailer's performance and allowing it to use its own initiative and experience in providing valuable services. See Mathewson & Winter, *The Law and Economics of Resale Price Maintenance*, 13 Rev. Indus. Org. 57, 74–75 (1998); Klein & Murphy, *Vertical Restraints as Contract Enforcement Mechanisms*, 31 J. Law & Econ. 265, 295 (1988); see also Deneckere, Marvel, & Peck, *Demand Uncertainty, Inventories, and Resale Price Maintenance*, 111 Q.J. Econ. 885, 911 (1996) (noting that resale price maintenance may be beneficial to motivate retailers to stock adequate inventories of a manufacturer's goods in the face of uncertain consumer demand).

* * *

While vertical agreements setting minimum resale prices can have procompetitive justifications, they may have anticompetitive effects in other cases; and unlawful price fixing, designed solely to obtain monopoly profits, is an ever–present temptation. Resale price maintenance may, for example, facilitate a manufacturer cartel. An unlawful cartel will seek to discover if some manufacturers are undercutting the cartel's fixed prices. Resale price maintenance could assist the cartel in identifying price-cutting manufacturers who benefit from the lower prices they offer. Resale price maintenance, furthermore, could discourage a manufacturer from cutting prices to retailers with the concomitant benefit of cheaper prices to consumers.

Vertical price restraints also "might be used to organize cartels at the retailer level." *Business Electronics* [*Corp. v. Sharp Electronics Corp*., 485 U.S. 717,], 725–726 [(1988)]..A group of retailers might collude to fix prices to consumers and then compel a manufacturer to aid the unlawful arrangement with resale price maintenance. In that instance the manufacturer does not establish the practice to stimulate services or to promote its brand but to give inefficient retailers higher profits. Retailers with better distribution systems and lower cost structures would be prevented from charging lower prices by the agreement. Historical examples suggest this possibility is a legitimate concern. See, *e.g.,* Marvel & McCafferty, *The Welfare Effects of Resale Price Maintenance*, 28 J. Law & Econ. 363, 373 (1985) (providing an example of the power of the National Association of Retail Druggists to compel manufacturers to use resale price maintenance); Hovenkamp 186 (suggesting that the retail druggists in *Dr. Miles* formed a cartel and used manufacturers to enforce it).

* * *

Resale price maintenance, it is true, does have economic dangers. If the rule of reason were to apply to vertical price restraints, courts would have to be diligent in eliminating their anticompetitive uses from the market. This is a realistic objective, and certain factors are relevant to the inquiry. For example, the number of manufacturers that make use of the practice in a given industry can provide important instruction. When only a few manufacturers lacking market power adopt the practice, there is little likelihood it is facilitating a manufacturer cartel, for a cartel then can be undercut by rival manufacturers. Likewise, a retailer cartel is unlikely when only a single manufacturer in a competitive market uses

resale price maintenance. Interbrand competition would divert consumers to lower priced substitutes and eliminate any gains to retailers from their price-fixing agreement over a single brand. See Posner 172; Bork 292. Resale price maintenance should be subject to more careful scrutiny, by contrast, if many competing manufacturers adopt the practice. Cf. Scherer & Ross 558 (noting that "except when [resale price maintenance] spreads to cover the bulk of an industry's output, depriving consumers of a meaningful choice between high-service and low-price outlets, most [resale price maintenance arrangements] are probably innocuous"); Easterbrook 162 (suggesting that "every one of the potentially-anticompetitive outcomes of vertical arrangements depends on the uniformity of the practice").

The source of the restraint may also be an important consideration. If there is evidence retailers were the impetus for a vertical price restraint, there is a greater likelihood that the restraint facilitates a retailer cartel or supports a dominant, inefficient retailer. See Brief for William S. Comanor et al. as *Amici Curiae* 7–8. If, by contrast, a manufacturer adopted the policy independent of retailer pressure, the restraint is less likely to promote anticompetitive conduct. Cf. Posner 177 ("It makes all the difference whether minimum retail prices are imposed by the manufacturer in order to evoke point-of-sale services or by the dealers in order to obtain monopoly profits"). A manufacturer also has an incentive to protest inefficient retailer-induced price restraints because they can harm its competitive position.

As a final matter, that a dominant manufacturer or retailer can abuse resale price maintenance for anticompetitive purposes may not be a serious concern unless the relevant entity has market power. If a retailer lacks market power, manufacturers likely can sell their goods through rival retailers. See also *Business Electronics, supra*, at 727, n. 2 (noting "[r]etail market power is rare, because of the usual presence of interbrand competition and other dealers"). And if a manufacturer lacks market power, there is less likelihood it can use the practice to keep competitors away from distribution outlets.

* * *

The manufacturer has a number of legitimate options to achieve benefits similar to those provided by vertical price restraints. A manufacturer can exercise its *Colgate* right to refuse to deal with retailers that do not follow its suggested prices. The economic effects of unilateral and concerted price setting are in general the same. The problem for the manufacturer is that a jury might conclude its unilateral policy was really a vertical agreement, subjecting it to treble damages and potential criminal liability. Even with the stringent standards in *Monsanto* and *Business Electronics,* this danger can lead, and has led, rational manufacturers to take wasteful measures. A manufacturer might refuse to discuss its pricing policy with its distributors except through counsel knowledgeable of the subtle intricacies of the law. Or it might terminate longstanding distributors for minor violations without seeking an explanation. The increased costs these burdensome measures generate flow to consumers in the form of higher prices.

Furthermore, depending on the type of product it sells, a manufacturer might be able to achieve the procompetitive benefits of resale price maintenance by integrating downstream and selling its products directly to consumers. *Dr. Miles* tilts the relative costs of vertical integration and vertical agreement by making the former more attractive based on the *per se* rule, not on real market conditions. [S]ee generally Coase, The Nature of the Firm, 4 Economica, New Series 386 (1937). This distortion might lead to inefficient integration that would not otherwise take place, so that consumers must again suffer the consequences of the suboptimal distribution strategy. And integration, unlike vertical price restraints, eliminates all intrabrand competition.

There is yet another consideration. A manufacturer can impose territorial restrictions on distributors and allow only one distributor to sell its goods in a given region. Our cases have recognized, and the economics literature confirms, that these vertical nonprice restraints have impacts similar to those of vertical price restraints; both reduce intrabrand competition and can stimulate retailer services.... The same legal standard (*per se* unlawfulness) applies to horizontal market division and horizontal price fixing because both have similar economic effect. There is likewise little economic justification for the current differential treatment of vertical price and nonprice restraints. Furthermore, vertical nonprice restraints may prove less efficient for inducing desired services, and they reduce intrabrand competition more than vertical price restraints by eliminating both price and service competition. See Brief for Economists as *Amici Curiae* 17–18.

* * *

For these reasons the Court's decision in Dr. Miles Medical Co. v. John D. Park & Sons Co., 220 U.S. 373 (1911), is now overruled. Vertical price restraints are to be judged according to the rule of reason.

V

* * *

The judgment of the Court of Appeals is reversed, and the case is remanded for proceedings consistent with this opinion.

Notes and Questions

1. Resale Price Maintenance and the Leegin Factors: The business conduct at issue in this case was minimum resale price maintenance (RPM). RPM occurs where a manufacturer demands that retailers only sell the manufacturer's product above an established price floor. As evidenced in the opinion, the Court gave substantial weight to academic literature claiming that this practice carries with it both a possibility for anticompetitive harm to markets and procompetitive benefits. The Court rejected the previously standing 100 year rule of *per se* condemnation of RPM and replaced it with the rule of reason. The Court placed the onus of further development of the rule of reason on the district courts and advised them to be "diligent in eliminating their anticompetitive uses from the market." Justice Kennedy points to three factors as guidance for when to be concerned about the anticompetitive harms of minimum RPM. Those three factors include: (1) when RPM covers the bulk of an industry's goods, (2) when a single dominant retailer, rather than manufacturers, implements or encourages RPM, and (3) when RPM is instituted by firms with discernible market power. Do these three factors, purportedly increasing the dangers of RPM, make economic sense? How should a court go about weighing the magnitude of economic harm and procompetitive benefit? Is evidence of what the court calls free-riding by low-cost retailers on the customer service of higher cost retailers, such as showrooms, product demonstrations, and knowledgeable employees, dispositive evidence of economic harm?

2. Intrabrand v. Interbrand Rivalry: The Court recognizes a distinction between intrabrand and interbrand competition. Should these two different types of competition be analyzed in the same manner for purposes of antitrust regulation? The Court observes that some vertical nonprice restraints on intrabrand competition may actually improve interbrand competition. How does this work? Does this also mean that the losses in efficiency resulting from restraints in intrabrand competition are made up for by improved

interbrand competition? At least one prominent commentator is quite skeptical of the balancing the court seems to support. Judge Frank Easterbrook has written that:

> No one can sensibly weigh inter- and intrabrand competition against one another; they are not commensurable. The reduction in "intrabrand competition" is the *source* of the competitive benefit that helps one product compete against another. Intrabrand competition as such is worthless; one might as well complain when a corporation does not have internal competition to make the product most cheaply. Vertical integration eliminates this form of "competition," but in so doing it may enable the manufacturer to reduce its delivered price. No manufacturer wants to have less competition among its dealers for the sake of less competition. The reduction in dealers' rivalry in the price dimension is just the tool the manufacturer uses to induce greater competition in the service dimension. As I spelled out above, restricted dealing alters the product's attributes. There is no "less" in one column to "balance against a gain" in the other, any more than the manufacturer's sole prerogative to decide what physical product to make creates a "reduction in intrabrand competition."

Frank H. Easterbrook, *Vertical Arrangements and the Rule of Reason*, 53 Antitrust L.J. 135 (1984).

***3. Development of the Rule of Reason in Vertical Restraint*:** Justice Kennedy's opinion recognized that the Court had slowly but surely been moving in the direction of handling most business conduct under the rule of reason rather than condemning the same conduct under a rule of *per se* illegality. *Leegin* was foretold by a major shift in the court's approach to nearly all vertical restraints. The Court cited *GTE Sylvania* for the proposition that non-price vertical restraints should be evaluated under a rule of reason. In that case, a manufacturer of television sets used vertical restraints in the form of location-specific exclusive contracts. The agreements only allowed the retailer to sell the product at specific stores specified in the agreement. One retailer, despite agreeing to this term and having notable success with the sale of Sylvania TVs, sued under the antitrust laws because it wanted to sell this product in other geographical regions disallowed under the contract. The Supreme Court, in an argument similar to the one adopted in *Leegin*, highlighted the possibility for vertical restrictions on goods to increase interbrand competition and allow smaller manufacturers to compete. It reasoned that:

> Vertical restrictions reduce intrabrand competition by limiting the number of sellers of a particular product competing for the business of a given group of buyers. Location restrictions have this effect because of practical constraints on the effective marketing area of retail outlets. Although intrabrand competition may be reduced, the ability of retailers to exploit the resulting market may be limited both by the ability of consumers to travel to other franchised locations and, perhaps more importantly, to purchase the competing products of other manufacturers. None of these key variables, however, is affected by the form of the transaction by which a manufacturer conveys his products to the retailers.
>
> Vertical restrictions promote interbrand competition by allowing the manufacturer to achieve certain efficiencies in the distribution of his products. These "redeeming virtues" are implicit in every decision sustaining vertical restrictions under the rule of reason. Economists have identified a number of ways in which manufacturers can use such restrictions to compete more effectively against other manufacturers. See, e. g., Preston, *Restrictive Distribution Arrangements: Economic Analysis and Public Policy Standards*, 30 Law & Contemp. Prob. 506, 511 (1965).

For example, new manufacturers and manufacturers entering new markets can use the restrictions in order to induce competent and aggressive retailers to make the kind of investment of capital and labor that is often required in the distribution of products unknown to the consumer. Established manufacturers can use them to induce retailers to engage in promotional activities or to provide service and repair facilities necessary to the efficient marketing of their products. Service and repair are vital for many products, such as automobiles and major household appliances. The availability and quality of such services affect a manufacturer's goodwill and the competitiveness of his product. Because of market imperfections such as the so-called "free rider" effect, these services might not be provided by retailers in a purely competitive situation, despite the fact that each retailer's benefit would be greater if all provided the services than if none did.

Economists also have argued that manufacturers have an economic interest in maintaining as much intrabrand competition as is consistent with the efficient distribution of their products. Bork, *The Rule of Reason and the Per Se Concept: Price Fixing and the Market Division (II)*, 75 Yale L.J. 373, 403 (1966). Although the view that the manufacturer's interest necessarily corresponds with that of the public is not universally shared, even the leading critic of vertical restrictions concedes that *Schwinn*'s distinction between sale and nonsale transactions is essentially unrelated to any relevant economic impact. Comanor, *Vertical Territorial and Customer Restrictions:* White Motor *and Its Aftermath*, 81 Harv. L. Rev. 1419, 1422 (1968).25 Indeed, to the extent that the form of the transaction is related to interbrand benefits, the Court's distinction is inconsistent with its articulated concern for the ability of smaller firms to compete effectively with larger ones. Capital requirements and administrative expenses may prevent smaller firms from using the exception for nonsale transactions. See, e. g., Phillips, Schwinn *Rules and the "New Economics" of Vertical Relation*, 44 Antitrust L.J. 573, 576 (1975).

* * *

In sum, we conclude that the appropriate decision is to return to the rule of reason that governed vertical restrictions prior to Schwinn. When anticompetitive effects are shown to result from particular vertical restrictions they can be adequately policed under the rule of reason, the standard traditionally applied for the majority of anticompetitive practices challenged under § 1 of the Act. Accordingly, the decision of the Court of Appeals is Affirmed.

Cont'l T. V., Inc. v. GTE Sylvania Inc., 433 U.S. 36, 54–59 (1977).

Another shift in the court's approach to vertical restraints came in *State Oil v. Khan*. There, a gasoline supplier terminated an agreement with an owner of a gas station who promptly sued alleging that State Oil's policies mandated a maximum resale price that should be condemned as a per se violation of the antitrust laws. The agreement provided the gas station owner with a 3.25 cent margin on the purchase price of every gallon of gasoline. If the gas station charged any price higher than the supplier's suggested retail price the extra money had to be rebated back to State Oil. In dismissing the application of the *per se* rule, a unanimous Court affirmed the reasoning, although not the decision, of Judge Posner in the Seventh Circuit. There, Judge Posner applied the *per se* rule in accordance with *stare decisis* but not without analyzing the answer to the question:

Why might competitors fix a maximum resale price? The difference between what a supplier charges his dealer and what the dealer charges the ultimate

customer is, functionally, compensation to the dealer for performing the resale service; so by agreeing on the resale prices of their goods competing sellers can reduce their dealers' margin below the competitive price for the dealers' service. This is a form of monopsony pricing, which is analytically the same as monopoly or cartel pricing and so treated by the law. E.g., Mandeville Island Farms, Inc. v. American Crystal Sugar Co., 334 U.S. 219 (1948).

The questionable next step (logically, not chronologically, next) in the evolution of antitrust law was to affix the per se label to contracts in which a single supplier, not acting in concert with any of its competitors, fixed its dealers' retail prices. Dr. Miles Medical Co. v. John D. Park & Sons Co., 220 U.S. 373 (1911). Here the economic difference between fixing a minimum resale price and fixing a maximum resale price becomes more pronounced, although most economists believe that neither form of price fixing is pernicious when the supplier is neither the cat's paw of colluding distributors nor acting in concert with his competitors. A supplier acting unilaterally might fix a minimum resale price in order to induce his dealers to furnish valuable point-of-sale services (trained salesmen, clean restrooms — whatever) to customers, which they could not afford to do without a guaranteed margin to cover the costs of the services, because the customers would use the services provided by the full-service dealers but then purchase the product from a competing dealer who could sell the product at a discount because he had not borne the expense of providing the services. Lester G. Telser, *Why Should Manufacturers Want Fair Trade?*, 3 J. Law & Econ. 86 (1962).

As for maximum resale price fixing, unless the supplier is a monopsonist he cannot squeeze his dealers' margins below a competitive level; the attempt to do so would just drive the dealers into the arms of a competing supplier. A supplier might, however, fix a maximum resale price in order to prevent his dealers from exploiting a monopoly position. We do not know anything about the competitive environment in which Khan and State Oil operate — which is why the district judge was right to conclude that if the rule of reason is applicable, Khan loses. But suppose that State Oil, perhaps to encourage the dealer services that we mentioned, has spaced its dealers sufficiently far apart to limit competition among them (or even given each of them an exclusive territory); and suppose further that Union 76 is a sufficiently distinctive and popular brand to give the dealers in it at least a modicum of monopoly power. Then State Oil might want to place a ceiling on the dealers' resale prices in order to prevent them from exploiting that monopoly power fully. It would do this not out of disinterested malice, but in its commercial self-interest. The higher the price at which gasoline is resold, the smaller the volume sold, and so the lower the profit to the supplier if the higher profit per gallon at the higher price is being snared by the dealer.

Khan v. State Oil Co., 93 F.3d 1358, 1361–62 (7th Cir. 1996) *vacated*, 522 U.S. 3 (1997).

4. A Spirited Dissent: There was a four-Justice dissent in the *Leegin* opinion written by Justice Breyer that stressed the value of *stare decisis* and the limited authority of the court to overrule well-established decisions. Arguing that antitrust policy required stable and predictable rules he contended that "the fact that a rule of law has become 'embedded' in our 'national culture' argues strongly against overruling. The *per se* rule forbidding minimum resale price maintenance agreements has long been 'embedded' in the law of antitrust. It involves price, the economy's 'central nervous system.' It reflects a basic

antitrust assumption (that consumers often prefer lower prices to more service). It embodies a basic antitrust objective (providing consumers with a free choice about such matters). And it creates an easily administered and enforceable bright line, 'Do not agree about price,' that businesses as well as lawyers have long understood." Leegin Creative Leather Products, Inc. v. PSKS, Inc., 551 U.S. 877, 926–27 (2007) (citations omitted).

Is predictability a good reason to keep old precedent? What do you think happens in the business world when the Supreme Court changes a long-established rule that substantially affects the way that corporations compete? The Supreme Court frequently overrules itself. Why should this opinion be any different?

5. Manufacturers' Interest in Retailer Service and Prices: For minimum resale price maintenance, a manufacturer conditions a retailer's right to sell the manufacturer's products on the retailer's agreement not to sell the product below a certain price. This seems counterintuitive for the manufacturer because a retailer's margin over cost does not impact the manufacturer's profit. Put another way, a manufacturer sets its price regardless of what the retailer does so why does he care about the retailer's price? Some insight may be provided by a statement made by Leegin's president: "[W]e want the consumers to get a different experience than they get in Sam's Club or in Wal-Mart. And you can't get that kind of experience or support or customer service from a store like Wal-Mart." 551 U.S. at 882. Why does the president care how much service retailers provide? Are you convinced that high quality service benefits Leegin, as the manufacturer? If so, what is the logic?

6. Rule of Reason's Structure: The majority recognizes that there is a valid concern that minimum resale price maintenance is nothing but a ruse among retailers to cartelize the market. This valid concern, as elaborated by the dissent, is why most scholars promulgate a rule or reason for minimum resale price maintenance instead of a standard of per se legality. *But see* Richard A. Posner, *The Next Step in the Antitrust Treatment of Restricted Distribution: Per Se Legality*, 48 U. Chi. L. Rev. 6 (1981). *Leegin*'s actual impact will not be fully established until courts determine how the rule of reason should apply. If the factfinder finds that the minimum resale price maintenance was imposed by the retailers, what does this say about the likely anticompetitive effect of the agreement? Is the analysis different if the factfinder finds that the manufacturer initiated the minimum resale price agreement? How should the answers to this question impact how courts should apply the rule of reason to these agreements? For more information on the potential structure for the rule of reason, see Thomas A. Lambert, Dr. Miles *Is Dead. Now What?: Structuring A Rule of Reason for Evaluating Minimum Resale Price Maintenance*, 50 Wm. & Mary L. Rev. 1937 (2009).

7. The Intersection of State and Federal Antitrust Law: *Leegin*'s impact on firms' actual pricing practices is very difficult to predict given the complicated overlap of state and federal law in this area. Some state antitrust laws are identical to the Sherman Act, and precedent or statutes require the state court to follow federal precedent interpreting the Sherman Act. Those states presumably will follow *Leegin*. Other states, however, have codified the per se illegal standard for minimum resale price maintenance. In these states, a firm implementing resale price maintenance may not be liable under federal law but still will be liable under state law. See Michael A. Lindsay, *An Update on State RPM Laws Since* Leegin, Antitrust Source, Dec. 2010 (providing a table of all the states and how each treats minimum resale price maintenance). With this structure of laws, is it likely that national firms will even bother with resale price maintenance because they will essentially have a complicated patchwork of pricing schemes depending on state boundaries and statutes? If a national firm proceeds and implements a patchwork of pricing strategy even

though it is very complicated and costly to implement, does this lend support to the majority's view that minimum resale price maintenance is procompetitive?

8. Who Is the Loser: The majority argues that minimum resale price maintenance can alleviate the market of the free-rider problem:

> If the consumer can … buy the product from a retailer that discounts because it has not spent capital providing services or developing a quality reputation, the high-service retailer will lose sales to the discounter, forcing it to cut back its services to a level lower than consumers would otherwise prefer. Minimum resale price maintenance alleviates the problem because it prevents the discounter from undercutting the service provider. With price competition decreased, the manufacturer's retailers compete among themselves over services.

551 U.S. at 891. This argument is criticized because preventing the free-rider problem converts consumer welfare into producer welfare — certainly not the goal of antitrust. Critics argue it is better to know that consumers are getting lower prices than to hope that pricing schemes that increase prices somehow result in higher demand and increased total welfare. Does the knowledge that rational consumers will choose more consumer welfare over less resolve this concern since rational profit maximizing firms will choose a strategy that benefits consumers?

The most likely loser under the rule of reason is an individual that does not want point of service sales and is satisfied with little or no service and lower prices. Is the Court right to ignore this consumer in its analysis? Is this resolved by the difference in a per se illegal standard, the rule of reason standard, and a per se legal standard?

9. RPM and Increased Prices: Justice Breyer stated that "[m]ost economists today agree that, in the words of a prominent antitrust treatise, 'resale price maintenance tends to produce higher consumer prices than would otherwise be the case.'" 551 U.S. at 912. At least one empirical study attempted to measure *Leegin*'s impact on prices and found some evidence that prices did increase after *Leegin* was decided. Nathaniel J. Harris, Leegin*'s Effect on Prices: An Empirical Analysis*, 9 J.L. Econ. & Pol'y 251 (2013). Is this price increase evidence dispositive and mean *Leegin* was incorrectly decided? If price increases, does consumer welfare — the ultimate standard for antitrust — necessarily have to decrease? If minimum resale price maintenance shifted the demand curve outward, would price increase? Would this shift necessarily increase consumer welfare?

10. Vertical Restraints and Evidence: Economists have continued to seek ways to use data, empirical methods, and quantitative approaches to test and confirm economic theories. Reviews of empirical literature seem to suggest that vertical restraints are associated with higher levels of output. For example, economists Francine Lafontaine and Margaret Slade concluded from their review that:

> Specifically, it appears that when manufacturers choose to impose such restraints, not only do they make themselves better off, but they also typically allow consumers to benefit from higher quality products and better service provision.… The evidence thus supports the conclusion that in these markets, manufacturer and consumer interests are apt to be aligned, while interference in the market is accomplished at the expense of consumers (and of course manufacturers).

For the details of their quantitative results, read Francine Lafontaine & Margaret Slade, *Exclusive Contracts and Vertical Restraints: Empirical Evidence and Public Policy*, 45 J. Econ

administration that will manage the public lands in the public interest.[151] Former Senator Wirth asserts that a decade of conflict and indifference has now ended.[152] In other words, under the Clinton Administration, politics will no longer govern. But is there any objective basis for concluding that the Clinton Administration's public lands policies are more clearly in the public interest than were the policies of the Bush and Reagan Administrations? No. All three administrations have claimed to be acting in the public interest. That these policies have varied only evidences the vicissitudes of American politics over the last dozen years.

IV. THE INEVITABILITY OF PRIVATE RIGHTS IN THE PUBLIC LANDS

From the foregoing analysis we can conclude that the federal public lands are the property of the United States Government which, like all property within the national borders, is under the sovereign jurisdiction of the United States.[153] Although occasionally still treated as a commons, the public lands are not a commons under the law. It would not be inconsistent with democratic theory for the public lands to be managed by the United States Government as a trustee for the body politic or for the citizenry as tenants in common, but neither practice nor law supports this interpretation.

It is important to understand that the United States Government controls the federal public lands in its proprietary capacity. Although theoretically plausible, the alternative explanations of the status of the public lands are legally inaccurate and politically misleading. It is also important to understand that the government acting in its proprietary capacity is different from a private party acting in its proprietary capacity because the government possesses unique coercive powers. The coercive powers of government exist whether government is acting in its proprietary or sovereign capacities. The United States Government has the legal authority to buy, sell or trade the public lands. The government

151. Tape of Conference on Public Lands, held by the Natural Resources Law Center and the University of Colorado Law Review in Boulder, Colorado (Sept. 23-25, 1993) (Senator George Miller speaking) (on file with Natural Resources Law Center).

152. *Id.* (Former Senator Tim Wirth speaking).

153. However, ownership of the public lands by the United States Government does not completely exempt them from the sovereign jurisdiction of state governments. Rather, certain activities conducted within a state's borders are properly regulated by the state no matter who owns the land upon which the activity occurs. For example, "[u]nder their general police powers, the states have traditionally regulated the taking and transport of fish and wildlife within their borders. Resting on well established law, such police powers have extended to the regulation of harvesting game on the public lands in the absence of conflicting Federal legislation." PLLRC, *supra* note 2, at 158. Similarly, state police power arguably extends to the regulation of mining activities on public lands. William R. Marsh & Don H. Sherwood, *Metamorphosis in Mining Law: Federal Legislative and Regulatory Amendment and Supplementation of the General Mining Law Since 1955*, 26 ROCKY MTN. MIN. L. INST. 209, 295 (1980). In fact, certain uses of the public lands may be subject to local regulation. "The police power over land use is ineluctable. It is reasonable to assume that local governments will shortly have extended zoning or other discretionary land use controls to cover virtually all land containing valuable mineral resources." Paul J. Schlauch, *Tripartite Federalism–The Emerging Role of Local Government as a Regulator of the Extractive Industries*, 20 ROCKY MTN. MIN. L. INST. 359, 397 (1974).

has the legal authority to grant or preclude access to and use of the public lands. Subject to the minimal limits of the Constitution,[154] the government may exercise these powers so as to grant or deny benefits to whomever it chooses. Even where the Constitution limits the government's discretion in allocating public land resources and distributing the benefits of those uses, the government cannot avoid the ultimate need to grant benefits to some and deny them to others. So long as the government controls resource allocation, whether in its sovereign or proprietary capacities, it cannot avoid this ultimate assignment of benefits. Only the creation of a true commons or privatization will relieve the government, and therefore the political system, of the task of deciding who gets what—of effectively assigning private rights in the public lands.[155]

A. *The Physical Realities of Resource Use*

When resources are in unlimited supply,[156] allocation to alternative uses and distribution of the benefits of those uses are not social problems.[157] Individuals may employ such resources as they choose, subject only to their own capacity for resource utilization.[158] But few, if any, resources are in unlimited supply. Resources not in unlimited supply, like those of the public lands, are scarce.

When resources are scarce, societies require mechanisms for allocating them among alternative uses and for distributing the benefits of their use among individual members of the society. Whatever the chosen mechanism, whether pri-

---

154. As with other functions of the federal government, actions on the public lands are limited by the various constitutional guarantees of individual liberty such as due process, equal protection, and free speech. U.S. CONST. amend. I; U.S. CONST. amend. XIV.

155. For the purposes of this discussion, I assume that in a commons, each member of the commons acquires a right in the commons by virtue of membership in the relevant community. Members of the community may do as they please with common resources, while nonmembers of the community are excluded from access except on terms dictated by the unanimous agreement of the members of the community. The public lands do not meet this definition because there is not open access. The privatization of public lands will require an initial assignment of right to particular individuals. In the interest of equality, the government may seek to assign identical rights to all individuals, but as the former Communist countries of eastern Europe have discovered, this often is not practical.

156. Scarcity is a function of many factors in addition to actual physical supply. It is affected by technology, by tastes, and by the availability of substitutes. Thus, things that were once scarce may become effectively of unlimited supply as a consequence of changed preferences, new technologies and the discovery of alternative resources.

157. As used in this article, the term allocation refers to the assignment of resources to particular uses. The term distribution refers to the assignment to individuals of the costs and benefits of those uses.

158. Although resource abundance relieves the society of the need for allocational and distributional institutions, individuals must still make choices in their own resource use due to the finite consumptive capacity of individuals and to the associated diminishing marginal utility which influences individual choices among alternative consumptive patterns. *See* WILLIAM OPHULS, ECOLOGY AND THE POLITICS OF SCARCITY 8 (1977) ("the institution of government, whether it takes the form of primitive tabu or parliamentary democracy, therefore has its origins in the necessity to distribute scarce resources in an orderly fashion"); HAROLD J. BARNETT & CHANDLER MORSE, SCARCITY AND GROWTH: THE ECONOMICS OF NATURAL RESOURCE AVAILABILITY 3-7 (1963).

vate, public or some combination, the allocations which are made will benefit some individuals more than others. Saying that the allocation is in the public interest, or that the government is acting on behalf of the body politic, does not alter this reality. Scarce resources, because they are scarce, will be allocated to uses which some individuals prefer and not to uses which other individuals prefer. Thus benefits of a particular resource use will be distributed to some individuals and not to other individuals.

The only way to avoid these distributional results of allocating scarce resources is to provide equal access for everyone. Unlimited equal access will lead to the destruction of the resource. It is the tragedy of the commons.[159] Limited equal access, depending upon how it is accomplished, may alleviate the distributional problem, but at great cost to average individual welfare as well as to the welfare of the least well off individuals.[160] Equal access, in a procedural sense,[161] might be accomplished through a lottery, but a lottery will not eliminate the uneven distribution of the benefits of resource use.[162] Some will win the lottery and others will lose.[163] Substantive equality might be provided by proportional distribution of benefits, but few resource uses can be divided indefinitely without losing most or all of the benefit of the use.[164] Many alternative uses of the same resource conflict one with the other,[165] making the determination of present day proportionality an immensely challenging task.

159. *See* Garrett Hardin, *The Tragedy of the Commons*, 162 SCI. 1243 (1968).

160. For a theoretical explanation of the relationship between differences in wealth and the welfare of the least well off, see JOHN RAWLS, A THEORY OF JUSTICE 75-83 (1971).

161. The strong commitment in American law to due process evidences that people can care about equal process even when it does not deliver equal outcomes. For example, people often prefer a process in which they have the opportunity to participate over a process in which they are not allowed to participate independent from the benefits which they derive from the process. In a sense, people are consumers of process as well as of tangible benefits.

162. In theory a lottery could evenly distribute the benefits of scarce resource use over time, but only if the resource is sufficiently abundant to assure every individual use during his or her lifetime and if present uses are properly discounted to be equal to future uses. A lottery would also require a prior determination of sustainability in order to determine appropriate individual uses. Absent complete information, which we can never have, sustainability will elude the lottery master, as it will elude the Babbittian ecosystem manager.

163. For example, hunting licenses for some game and permits for white water boating on some rivers are allocated by lotteries. The result is that some people get to hunt and boat and other people are excluded.

164. This is true for several reasons. Most resources have multiple alternative uses, many of which are incompatible. My use of wildlife for viewing is not compatible with your use of wildlife for hunting. Our equality masters might assign some wildlife to me for viewing and some to you for hunting, but if everyone is to have an equal share in the potential benefits of wildlife, the limit of divisibility of the wildlife resource will soon be exceeded with benefits resulting to no one. Just as individual consumers tend to have diminishing marginal utilities with increased consumption of a resource, so too they have diminishing marginal utilities below some level of consumption. When an individual's share of a resource is less than the amount necessary to experience the benefit the individual values, it is worth little or nothing to that individual. An exclusive right to grazing land insufficient to feed a cow, or to timber insufficient to operate a mill, or to wilderness insufficient in which to be alone is worthless to the individual who values those uses.

165. This fact often leads policy advocates and policy makers to avoid the comparison of conflicting use claims by asserting that the alternative uses are incommensurables. This implies that there is no way to make a reasoned choice. Yet the choice is ultimately made, one use or another among incompatible uses is permitted, and the incommensurables are thus commensurated.

Individual tastes differ. Tastes can be alternatively satisfied through resource substitutions.[166] Changes in technology, information and tastes will assure that today's proportional allocation, if it could be determined,[167] will be unequal tomorrow.[168]

The system of federally owned lands is one mechanism for the allocation of scarce resources. Resources get allocated politically, pursuant to the various public land laws and public land management agencies. Although these allocations are invariably made in the name of the public interest, individuals ultimately derive the benefits and costs of the allocations made by the public land managers. Congress may assert that the preservation of specified lands for wilderness, or the harvesting of some number of board feet of timber, or the continued exploration for hard rock minerals is in the public interest, but the reality is that some individuals will benefit from these actions and other individuals will not.

Those whose use of the public lands is based upon a grant, permit, lease, or contract, even if for a short period of time, may have an enforceable property or contract right. But those whose use is merely pursuant to generally applicable legislation and regulations have something analogous to a private right in the public lands.[169] So long as the legislation and regulations remain unchanged, those who engage in authorized uses have a right to do so while those who would engage in unauthorized uses have no such right. Those who value the authorized uses benefit, while those who value unauthorized uses do not benefit. As with all politically provided benefits, the users and nonusers share in the costs.[170]

---

166. For example, a taste for heat can be satisfied through coal, oil, electricity, etc.

167. To make a proportional distribution of costs and benefits we must be able to compare the costs and benefits on an individual basis. Kenneth Arrow's "impossibility theorem" suggests that such an interpersonal utility comparison cannot be done. KENNETH J. ARROW, SOCIAL CHOICE AND INDIVIDUAL VALUES (1951). *See also* James S. Coleman, *The Possibility of a Social Welfare Function*, 56 AM. ECON. REV. 1105 (1966); James M. Buchanan, *Positive Economics, Welfare Economics, and Political Economy*, 2 J.L. & ECON. 124 (1959).

168. On June 26, 1787, Alexander Hamilton told the Philadelphia convention: "It was certainly true that nothing like an equality of property existed: that an inequality would exist as long as liberty existed, and that it would unavoidably result from that very liberty itself." 1 THE RECORDS OF THE FEDERAL CONVENTION OF 1787 424 (Max Farrand rev. ed., 1966). James Madison explained further:

> The diversity in the faculties of men, from which the rights of property originate, is not less an insuperable obstacle to a uniformity of interests. The protection of these faculties is the first object of government. From the protection of different and unequal faculties of acquiring property, the possession of different degrees and kinds of property immediately results; and from the influence of these on the sentiments and views of the respective proprietors, ensues a division of the society into different interests and parties."

THE FEDERALIST NO. 19, at 53 (James Madison) (Mentor Books 1961).

169. Coggins et al., *supra* note 5, at 551 (permits issued pursuant to the Taylor Grazing Act do not grant the holder a fixed property right; nevertheless, holders, as well as the BLM "act as if other uses cannot be accommodated because the ranchers have 'rights' to the forage.").

170. Uses of the public lands for which user fees are paid are treated as permitted uses for the purposes of this discussion. The payer of a user fee is presumably entitled to a prescribed use once the fee is paid.

B. *Public Lands as Wealth Distribution*

Congressman Miller says that the public have changed their values. He suggests that the people of the Pacific Northwest are in favor of protecting spotted owls and old growth timber stands.[171] Who are the people of whom the Congressman speaks? The majority in the cities of Seattle and Portland? Most assuredly. The majority in the states of Washington and Oregon? Perhaps. The majority in the cities of Longview and Klamath Falls? Most assuredly not. That different majorities would not agree on this question, and that each of these majorities has a corresponding minority underscores the fact that majorities consist of individuals who value different things. When resources are allocated politically, people get what they want by convincing a majority to agree with them. Thus public lands management is fundamentally about politics. It is about gaining and losing individual wealth through the political process.

Ecosystem management proposed by the Clinton Administration,[172] like the scientific management of the progressive era about which Bob Nelson has commented,[173] will not avoid this essential political characteristic of public lands management. Ecosystem management will proceed on the assumption that ecosystem protection is to be valued over alternative uses of the public lands. Some people will benefit from this management policy, others will not. To the extent that current management practices sacrifice ecosystems for the provision of timber, grass, minerals, recreation, and game animals, wealth will be shifted from those presently benefiting from public lands policies to those who will benefit under a new regime of ecosystem management. Proponents of ecosystem management will claim that it is required by the public interest and that it is about science rather than politics. But while science can help the pursuit of chosen goals, it offers no assistance with the definition of the public interest.[174] Ecosystem management on the public lands will have wealth distribution consequences just as any other public lands policy goal. The advocates of ecosystem management are surely not ignorant of this fundamental truth about politics.

---

171. Tape of Conference on Public Lands, *supra* note 151.

172. "Interior Secretary Bruce Babbitt has put the [Clinton] [A]dministration squarely behind the concept of 'ecosystem management.'" Bill Dawson, *Jobs vs. Environment at Stake in Portland Forest Conference*, HOUS. CHRON., Mar. 28, 1993, at A13; *see also* William K. Stevens, *Babbitt to Map Ecosystems Under New Policy to Save Them*, N.Y. TIMES, Mar. 14, 1993, at A29 (Babbitt has championed a major policy shift towards "long-term care of entire ecosystems"); John Husar, *Higher Fees Bring Hope on the Range*, CHI. TRIB., Aug. 12, 1993, at N11 ("Babbitt unveiled a new ecosystem management philosophy.").

173. Nelson, *supra* note 12.

174. *See* James L. Huffman, *Truth, Purpose and Public Policy: Science and Democracy in the Search for Safety*, 21 ENVTL. L. 1091, 1106 (1991) (reviewing JOHN D. GRAHAM ET AL., IN SEARCH OF SAFETY: CHEMICALS AND CANCER RISK (1988)):

> If we do not recognize the conceptual distinction between science and values, we will almost certainly have public policies that are based upon unreliable data and controlled by nondemocratic means. Experts are very important to the policy-making process, particularly in our complex, modern world, but their expertise is limited. No individual is expert on the public interest in a democracy. By definition, the only expert is the democracy itself.

The history of public lands politics evidences that potential users of public lands resources have always understood the wealth implications of public lands management and policies. Throughout most of the nineteenth century, public lands policy on the vast western frontier was driven by the prospect of private wealth, along with anticipated public revenues. Congress would have been hard pressed to justify a policy of federal land retention in the face of political pressures calling for cheap land for western settlement.[175] During the second half of the nineteenth century, mining interests persuaded Congress to enact laws favorable to the private acquisition of mineral wealth.[176] During the same era, the agricultural interests, which had benefited from the land disposal laws, lobbied for free water from the public lands.[177] By the end of the century, the agriculture industry even managed to get the federal government to pay most of the bill for water development.[178] Early conservationists managed to garner a small piece of the wealth with the creation of Yellowstone National Park in 1876,[179] and a larger portion of the public land resource with the withdrawal of the forest reserves at the end of the century.[180] Timber interests were quick to turn these reserves to their advantage, at first by lobbying for the withholding of public land timber from the market,[181] and later by lobbying for a subsidized supply of timber from the public lands.[182] After several decades of essentially free grass from the unfenced public lands,[183] livestock ranchers persuaded Congress to create a system of forage leasing upon which they were able to base much of the value in their private ranches.[184]

---

175. "The pioneer won, and was destined to win, every encounter, sooner or later, with the cold business calculations of the East. Financiers could make paper requirements providing for a handsome payment by settlers for land. The settlers did not have the money, were destined not to have it, but settle the land they would." HIBBARD, *supra* note 19, at 5.

176. GATES, *supra* note 17, at 711-23.

177. MARC REISNER, CADILLAC DESERT 25-53 (1986) (detailing the efforts of John Wesley Powell to convince Congress that equitable sharing of all western water was the only practicable way to settle the West). *But see id.* at 50-53, 114-15 (noting that it was in fact the western agricultural industry which was to blame for the initial failure of reclamation, because of its unyielding faith in private enterprise. The western yeoman farmer was generally unwilling to accept the fact that the arid west couldn't bloom without the help of government irrigation projects; rather, he believed that "private capital, individual initiative, and hard work" were the solutions. More simply, the "indomitable individualist," who was the western pioneer, didn't, in the first place, want to hear that the West was dry, but if it was then he "didn't want the federal government parceling out water and otherwise meddling in [his] affairs.").

178. Reclamation Act of 1902, ch. 1093, 32 Stat. 388.

179. Although reservation of Yellowstone was no doubt inspired by a conservationist spirit in part, it was relatively easily done given the low resource value of the high elevation lands in question. It has also been demonstrated by Anderson and Hill that the railroads played a significant role in promoting Yellowstone as a playground for prospective railroad passengers. *See* Terry L. Anderson & P.J. Hill, *The Evolution of Property Rights: A Study of the American West*, 18 J.L. & ECON. 163 (1975).

180. *See supra* text accompanying notes 62-63.

181. *See* STEEN, *supra* note 75, at 90-96.

182. *See* Harbison, *supra* note 145, at 889-97 (discussing the willingness of the Forest Service to placate timber interests by modifying contract terms which would unduly burden timber companies).

183. *See* Anderson & Hill, *supra* note 179.

184. *See* Coggins, et al., *supra* note 5.

By the 1960's the competition for public land wealth had become so intense among diverse interests that Congress enacted the Multiple-Use Sustained-Yield Act[185] which was designed to assure that the wealth was shared.[186] Only four years later, however, Congress enacted the Wilderness Act,[187] an early warning that the concept of multiple use was flawed. The history of the enactment of the Wilderness Act was a reflection of the entire history of public lands politics. Potential wilderness areas containing commercial timber were excluded at the insistence of the timber industry. Indeed virtually all lands with commodity values were excluded, except for some mineral lands. But the mineral industry was given another twenty years to locate mineral claims within wilderness areas.[188] The wealth of the public lands was not to be transferred easily from some interests to others.

Pursuant to the Multiple-Use Sustained-Yield Act and other subsequent legislation,[189] the public land management agencies, particularly the Forest Service, engaged in two decades of planning for multiple use. It was a doomed exercise in trying to satisfy every claimant to public land resources. Congress had simply passed the buck to the agencies, but the agencies had little to guide them in deciding who should get what. The primary effect of combining multiple use with public lands planning was to give noncommodity interests greater influence through participation in a land management process driven by few substantive goals. The 1964 Wilderness Act was an early indication of the growing influence of these noncommodity, environmental interests. The emergence of ecosystem management, without any congressional authorization,[190] evidences that the wealth of the public lands has shifted dramatically to the protectionist interests of the environmental movement.

V. CONCLUSION

Federal public lands history has progressed through four overlapping stages: acquisition, disposal, retention and management. Through the first two stages it was assumed that land and resources should and would be privately managed. The late nineteenth century shift to a policy of retention led inevitably to the

185. 16 U.S.C. §§ 528-531 (1988).
186. *Id.* § 528.
187. 16 U.S.C. §§ 1131-1136 (1988).
188. *Id.* § 1133(d)(2)-(d)(3).
189. *See, e.g.*, Forest and Rangeland Renewable Resources Planning Act of 1974, Pub. L. No. 93-378, 88 Stat. 476; National Forest Management Act of 1976, 16 U.S.C. §§ 1600-1614 (1988).
190. The Clinton Administration is committed to ecosystem management, *see supra* note 172, and there is indication that Congress has begun to take seriously this environmental ideology. First, Congress is expected to approve an Interior Department request to transfer $12 million from elsewhere in the Interior budget to finance the reorganization of the Department's biological research, which includes the proposed national biological survey to map United States ecosystems and biological diversity. Stevens, *supra* note 172, at A29. Moreover, at least with respect to protection of Northwest salmon and their habitat, many members of Congress, like the Clinton Administration, "are calling for greater attention to ecosystems." Marla Williams, *Salmon Crisis Pits Ecology, Economy–Efforts to Save Coho Could Affect Millions of People*, SEATTLE TIMES, Oct. 21, 1993, at B1.

current era of federal management of the public lands. Throughout this history, the pursuit of the public interest was invariably the articulated objective. Disagreements in the public debates were about how best to achieve the public interest. The nature of democratic politics does not permit advocacy of private interests.

Yet American politicians and citizens have always been realists, whether dealing with public lands policy or some other area of governmental concern. We may lament government by special interests, but at least since James Madison wrote *Federalist No. 10*,[191] we have known what it takes to get political results. Our rhetoric may appeal to the greater good, but the votes are counted on the basis of who gets what. And special interests are nothing but aggregated private interests—private interests aggregated so that they might compete in the politics of, in this case, public lands.

In public lands politics our rhetoricians have offered several versions of the public lands system which they seek to influence. As Rousseau would have wanted it, the most frequent rhetoric is of lands belonging to the people—to all the people, or to the people as the body politic. It is effective political rhetoric because it obscures the agendas of the special interest groups who will have no standing in the public debate unless they claim to speak for the public interest. Some come, somehow, to believe their rhetoric. They convince themselves that theirs is the true rendering of the public interest, and that others are just out to benefit private interests. Others are less naive, or less self-confident, but they know what it takes to win at public lands politics.

Public lands politics is finally about distributing the benefits of public land resource use to individual members of the community. On rare occasions there may be near unanimity about the desired uses. In most cases there will be great diversity of opinion. It is inevitable that there should be such diversity in a liberal society. Indeed that is the point of a liberal society. However one chooses to define the public interest, there is no escaping that some private interests will be served while others are denied.

Charles Wilkinson has written that the "lords of yesterday"—the General Mining Law, the grazing laws, the appropriation water doctrine, the timber harvesting laws, and the water development laws—continue to rule western resource management.[192] To a significant extent those laws do govern public lands and western resource management, but the laws are not the lords. The lords of the public lands are and always have been private interests. So long as half of the

191 . THE FEDERALIST NO. 10, *supra* note 168, at 53:

> Complaints are everywhere heard from our most considerate and virtuous citizens, equally the friends of public and private faith and of public and personal liberty, that our governments are too unstable, that the public good is disregarded in the conflicts of rival parties, and that measures are too often decided, not according to the rules of justice and the rights of the minor party, but by the superior force of an interested and overbearing majority. However anxiously we may wish that these complaints had no foundation, the evidence of known facts will not permit us to deny that they are in some degree true.

192 . CHARLES F. WILKINSON, CROSSING THE NEXT MERIDIAN (1992).

American West is owned by the United States Government, the pursuit of public land wealth by private interests will be a dominant factor in national politics. But it will always be done in the name of the public interest. That is the way the game is played. Public lands are unavoidably political lands, and politics is inescapably about competing private interests. In the public lands debate, the rhetoric is about public rights, the reality is about private rights.

Positive, Normative and Functional Schools in Law and Economics

Abstract

During its relatively short history, the law and economics movement has developed three distinct schools of thought. The first two schools of thought, often referred to as the Chicago or positive school and the Yale or normative school, developed almost concurrently. The functional school of law and economics, which developed subsequently, draws from public choice theory and the constitutional perspective of the Virginia school of economics to offer a third perspective which is neither fully positive nor fully normative. Various important methodological questions have accompanied the debate between these schools concerning the appropriate role of economic analysis in the institutional design of lawmaking and the limits of methods of evaluation of social preferences and aggregate welfare in policy analysis. These debates have contributed to the growing intellectual interest in the economic analysis of law.

Keywords: law and economics, intellectual history, methodology

JEL Classification: K00, B0, B5, B25, B41

1. The origins and the evolved domain of law and economics

Various important methodological questions have accompanied the growth and evolution of law and economics. Economists and jurists alike have debated the appropriate role of economic analysis in the institutional design of lawmaking and the limits of methods of evaluation of social preferences and aggregate welfare in policy analysis. In many respects, these methodological debates have contributed to the diversification of methodologies in the economic analysis of law.

1.1. The origins of modern law and economics

Law and economics is probably the most successful example of the recent surge of applied economics into areas that were once regarded as beyond the realm of economic analysis and its study of explicit market transactions. Methodologically, law and economics applies the conceptual apparatus and empirical methods of economics to the study of law.

Extensive research has been carried out to identify the historical and antecedents to modern law and economics. The encyclopedic work edited by Jürgen Backhaus (2003) contains several biographical entries devoted to precursors and early European exponents of the law and economics movement. It is interesting to see that, although the recognition

of law and economics as an independent field of research is the result of studies carried out in the United States after the 1970s, it is in Europe that most of the precursors can be found. Notable antecedents to law and economics include the work of Adam Smith on the economic effects of legislation (1776), and Jeremy Bentham's theory of legislation and utilitarianism (1782, 1789).

In the United States, it was not until the mid-twentieth century—through the work of Henry Simon, Aaron Director, Henry Manne, George Stigler, Armen Alchian, Gordon Tullock, and others—that the links between law and economics became an object of serious academic pursuit. The regulation of business and economic law fell within the natural interest of the first American scholars of law and economics. Early research concentrated on areas related to corporate law, tax law, and competition law. In so doing, the first generation of law and economics scholars paralleled the efforts of other economists, trying to explain the functioning of explicit economic markets and the impact of alternative legal constraints, such as taxes and regulation, on the market.

In the 1960s the pioneering work of Ronald Coase and Guido Calabresi brought to light the pervasive bearing of economics in all areas of the law. The methodological breakthrough occasioned by Coase and Calabresi allowed immediate extensions to the areas of tort, property and contract. The analytical power of their work was not confined to these fields, however, and subsequent law and economics contributions demonstrate the explanatory and analytical reach of its methodology in a number of other areas of the law.

A difference in approach is detectable between the law and economics contributions of the early 1960s and those that followed in the 1970s. While the earlier studies appraise the effects of legal rules on the normal functioning of the economic system (i.e., they consider the impact of legal rules on the market equilibrium), the subsequent generation of studies utilizes economic analysis to achieve a better understanding of the legal system. Indeed, in the 1970s a number of important applications of economics to law gradually exposed the economic structure of basically every aspect of a legal system: from its origin and evolution, to its substantive, procedural, and constitutional rules.

Despite some resistance to the application of economics to nonmarket behavior, the important bonds between legal and economic analysis, as well as the social significance of the object of study, were in themselves a guarantee of success and fruitfulness for law and economics.

An important ingredient in the success of law and economics research has come from the establishment of specialized journals. The first such journal, the *Journal of Law and Economics*, appeared in 1958 at the University of Chicago. Its first editor, Aaron Director, should be credited for this important initiative, successfully continued by Ronald Coase. Other journals emerged in the following years: in 1972, the *Journal of Legal Studies*, also housed at the University of Chicago, was founded under the editorship of Richard Posner; in 1979, *Research in Law and Economics*, under the editorship of Richard Zerbe, Jr.; in 1981, the *International Review of Law and Economics* was established in the United Kingdom under the editorship of Charles Rowley and Anthony Ogus (later joined by Robert Cooter and Daniel Rubinfeld); in 1982, the *Supreme Court Economic Review*, under the editorship of Peter Aranson (later joined by Harold Demsetz and Ernest Gellhorn); in 1985, the *Journal of Law, Economics and Organization*, under the editorship of Jerry Mashaw

and Oliver Williamson (later joined by Roberta Romano); in 1994, the *European Journal of Law and Economics* was launched under the editorial direction of Jürgen Backhaus and Frank Stephen; in 1999, the *American Law and Economics Review,* under the editorship of Orley Ashenfelter and Richard Posner; and, most recently, in 2004 the *Journal of Empirical Legal Studies* under the editorship of Theodore Eisenberg, Jeffrey J. Rachlinski, Stewart J. Schwab, and Martin T. Wells; and the *Review of Law and Economics* established in 2005 under the editorship of Robert Cooter, Ben Depoorter, Lewis Kornhauser, Gerrit De Geest, Nuno Garoupa, and Francesco Parisi. These specialized journals provided—and continue to provide—an extremely valuable forum for the study of the economic structure of law.

In many respects, the impact of law and economics has exceeded its planned ambitions. One effect of the incorporation of economics into the study of law was to irreversibly transform traditional legal methodology. Legal rules began to be studied as a working system—a clear change from the Langdellian tradition, which had relied almost exclusively on the self-contained framework of case analysis and classification, viewing law as little more than a filing system. Economics provided the analytical rigor necessary for the study of the vast body of legal rules present in a modern legal system. This intellectual revolution came at an appropriate time, when legal academia was actively searching for a tool that permitted critical appraisal of the law, rather than merely strengthening the dogmatic consistencies of the system.

The marriage of law and economics has also affected the economic profession, contributing to the expansion of the original domain of microeconomic analysis—the study of individual and organizational choices in the market—to the study and understanding of other institutions and non-market phenomena.

1.2. The evolved domain of law and economics

Despite the powerful analytical reach of economics, it was clear from the outset that the economist's competence in the evaluation of legal issues was limited. While the economist's perspective could prove crucial for the positive analysis of the efficiency of alternative legal rules and the study of the effects of alternative rules on the distribution of wealth and income, economists generally recognized the limits of their role in providing normative prescriptions for social change or legal reform.

Recognition of the positive nature of the economic analysis of law was not sufficient to dispel the many misunderstandings and controversies in legal academia engendered by the law and economics movement's methodological revolution. As Coase (1978) indicated, the cohesiveness of economic techniques makes it possible for economics to move successfully into another field, such as law, and dominate it intellectually. But methodological differences played an important part in the uneasy marriage between law and economics. The Popperian methodology of positive science was in many respects at odds with the existing paradigms of legal analysis. Rowley (1981) characterizes such differences, observing that positive economics follow the Popperian approach, whereby testable hypotheses (or models) are derived by means of logical deduction and are then tested empirically. Anglo-American legal analysis, on the other hand, is generally inductive: lawyers use individual judgments to construct a general premise of law. Much work has been done in law and economics

despite these methodological differences, with a reciprocal enrichment of the analytical tools of both disciplines.

Law and economics relies on the standard economic assumption that individuals are rational maximizers, and studies the role of law as a means for changing the relative prices attached to alternative individual actions. Under this approach, a change in the rule of law will affect human behavior by altering the relative price structure—and thus the constraint—of the optimization problem. Wealth maximization, serving as a paradigm for the analysis of law, can thus be promoted or constrained by legal rules.

The early years of law and economics were characterized by the uneasiness of some traditional legal scholars in the acceptance of the notion of wealth maximization as an ancillary paradigm of justice. Although most of the differences gradually proved to be largely verbal—and many others were dispelled by the gradual acceptance of a distinction between paradigms of utility maximization and wealth maximization—two objections continue to affect the lines of the debate. The first relates to the need for specifying an initial set of individual entitlements or rights, as a necessary prerequisite for operationalizing wealth maximization. The second springs from the theoretical difficulty of defining the proper role of efficiency as an ingredient of justice, vis-à-vis other social goals.

In his well-known defense of wealth maximization as a guide for judicial action, Posner (1985a) distinguishes wealth or expected utility from market prices. While market prices may not always fully reflect idiosyncratic valuations, they avoid an undertaking of interpersonal utility comparisons, with the opportunity for ex post rationalization of positions taken on emotional grounds. Posner's view is sympathetic to the premises of a property right approach to legal relationships, and he stresses the importance of an initial distribution of property rights prior to any calculation of wealth maximization. His paradigm of wealth maximization serves as a common denominator for both utilitarian and individualist perspectives. By combining elements of both, Posner provides a theory of wealth maximization that comes closer to a consensus political philosophy than does any other overarching political principle.

In contrast, Calabresi (1980) claims that an increase in wealth cannot constitute social improvement unless it furthers some other goal, such as utility or equality. Denying that one can trade off efficiency against justice, he argues instead that efficiency and distribution are ingredients of justice, which is a goal of a different order than either of these ingredients. Calabresi thus defends law and economics as a worthy examination of certain ingredients of justice, rather than a direct examination of justice itself.

The intellectual resistance that has characterized the birth of law and economics can only be temporary. Both legal practitioners and policymakers are becoming aware of the important role of economic analysis in their discipline, and we have already mentioned notable contributions to mainstream economic theory from lawyers in the law and economics movement. Likewise, as Coase (1978) noted, economists have come to realize that the other social sciences are so intertwined with the economic system as to be part of the system itself. For this reason, law and economics can no longer be appraised as a branch of applied microeconomics; rather, it must be seen as contributing to a better understanding of the economic system itself. The study of the effects of other social sciences on the economic system will, Coase predicts, become a permanent part of the field of economics.

Coase also examines the reasons for the movement of economists into the other social sciences, and attempts to predict the future of this phenomenon. Groups of scholars are bound together by common techniques of analysis, a common theory or approach to the subject, and/or a common subject matter. In the short run, Coase maintains, one group's techniques of analysis may give it such advantages that it is able to move successfully into another field and maybe even dominate it. In the long run, however, the subject matter tends to be the dominant cohesive force. While the analytical techniques employed by economists—such as linear programming, quantitative methods, and cost-benefit analysis—may recently have aided the entry of economists into the other social sciences, Coase predicts that such a movement can only be temporary. After all, the wisdom possessed by economists, once its value is recognized, will be acquired by some of the practitioners in these other fields (as is happening in the field of law).

As the domain of law and economics continues to expand, its perspective on methodological issues has not been stagnant. While this introductory essay emphasizes the wide range of substantive applications, some degree of controversy still surrounds several of the methodological, normative, and philosophical underpinnings of the economic approach to law. Most of the ideological differences tend to lose significance because their operational paradigms often lead to analogous results when applied to real cases. Some scholars, however, perceive that the current state of law and economics as comparable to the state of economics prior to the advent of public choice theory, insofar as an understanding of "political failures" was missing from the study of market failures (Buchanan, 1974; Rowley, 1989) Public choice may indeed inject a skeptical—and at times disruptive—perspective into the more elegant and simple framework of neoclassical economics, but this added element may well be necessary to better understand a complex reality. In a way, the systematic incorporation of public choice theory into the economic approach to law has contributed to bridging the conflicting normative perspectives in law and economics, at least by bringing the debate onto the more solid ground of collective choice theory.

Economics is a powerful tool for the analysis of law. If humans are rational maximizers of their utility, wealth or well-being then they respond rationally to changes in exogenous constraints, such as laws. This rationality assumption provides the basic foundation for much law and economics literature. Building upon the standard economic assumption that individuals are rational maximizers, the sophisticated tools of price theory become a useful aid in the study and choice of legal rules (Cooter, 1984). While there is much consensus on the value of economic theory in the study of legal rules, important methodological differences arise with respect to the choice of the appropriate instruments of legal analysis and the choice of method for evaluation of social preferences. I will briefly discuss these methodological issues in turn.

2. Schools and intellectual perspectives in law and economics

Most practitioners of law and economics believe that there is an important common ground that unifies all scholars in the discipline, regardless of their ideological creed: a search for new insights in the law by applying economic concepts and theories (MacKaay, 2000). Despite this common statement of purpose, various schools of law and economics can

be identified, each with an elaborate research program and a distinct methodological approach.

2.1. The Chicago and Yale schools: Positive versus normative approaches to law and economics

During the early period of the discipline, law and economics scholarship was labeled as Chicago-style or Yale-style. These labels made reference to the respective positive or normative approach utilized by each school. The origins of the Chicago and Yale schools of law and economics are attributable to the early work of a handful of scholars, including the pioneering work of Ronald Coase and Guido Calabresi in the early 1960s.

At this point, methodological differences came to surface with substantive practical differences. The Chicago school laid most of its foundations on the work carried out by Richard Posner in the 1970s. An important premise of the Chicago approach to law and economics is the idea that the common law is the result of an effort—conscious or not—to induce efficient outcomes. This premise is known as the efficiency of the common law hypothesis. According to this hypothesis, first intimated by Coase (1960), and later systematized and greatly extended by Ehrlich and Posner (1974), Rubin (1977) and Priest (1977), common law rules attempt to allocate resources in either a Pareto or Kaldor-Hicks efficient manner. Further, according to the positive school, common law rules are said to enjoy a comparative advantage over legislation in fulfilling this task because of the evolutionary selection of common law rules through adjudication. Several important contributions provide the foundations for this claim; the scholars who have advanced theories in support of the hypothesis are, however, often in disagreement as to its conceptual basis.

The primary hypothesis advanced by positive economic analysis of law is thus the notion that efficiency is the predominant factor shaping the rules, procedures, and institutions of the common law. Posner contends that efficiency is a defensible criterion in the context of judicial decision-making because "justice" considerations—on the content of which there is no academic or political consensus—introduce unacceptable ambiguity into the judicial process.

In arguing for positive use of economics, Posner (1974) is not denying the existence of valuable normative law and economics applications. In fact, law and economics often has many objective things to say that will affect one's normative analysis of a policy.[1]

Despite the powerful analytical reach of economic analysis, Chicago scholars acknowledged from the outset that the economist's competence in the evaluation of legal issues was limited. While the economist's perspective could prove crucial for the positive analysis of the efficiency of alternative legal rules and the study of the effects of alternative rules on the distribution of wealth and income, Chicago-style economists generally recognized the limits of their role in providing normative prescriptions for social change or legal reform.

To the contrary, the Yale school of law and economics, often described as the "normative" school believes that there is a larger need for legal intervention in order to correct for pervasive forms of market failure.[2] Distributional concerns are central to the Yale-style literature. The overall philosophy of this group is often presented as more value-tainted and more prone to policy intervention than the Chicago law and economics school.

Unlike its Chicago counterpart, this school has attracted liberal practitioners who employ the methodology of the Chicago school but push it to formulate normative propositions on what the law ought to be like (MacKaay, 2000). Given the overriding need to pursue justice and fairness in distribution through the legal system, most Yale-style scholars would suggest that efficiency, as defined by the Chicago school, could never be the ultimate end of a legal system.

2.2. The Virginia school: The functional approach and the return to normative individualism

In recent years, a new generation of literature–developed at the interface of law, economics and public choice theory–pushes the methodological boundaries of economic analysis of law. The resulting approach is in many respects functional in its ultimate mission, cutting across the positive and normative distinction and unveiling the promises and pitfalls of both the normative and the positive alternatives. This approach to legal analysis has the potential of shedding light on the traditional conception of lawmaking, suggesting that the comparative evaluation of alternative sources of law requires an appropriate analysis of the incentive structure in the originating environment. This line of research is attentive to the identification of political failures in the formation of law, stressing the importance of market-like mechanisms in the creation and selection of legal rules.

The functional approach to law and economics is still in its initial phase of development and far from a point of theoretical maturity, but this approach is unquestionably successful in raising some crucial questions regarding the difficult link between individual preferences and social outcomes, with an emphasis on institutional mechanism design and individual choice. The resulting approach is quite skeptical of both the normative and the positive alternatives.[3] Public choice theory provides strong methodological foundations for the functional school of law and economics: the systematic incorporation of the findings of public choice theory into the economic analysis of law may serve to bridge the conflicting normative perspectives in law and economics, at least by bringing the debate onto the more solid ground of collective choice theory.

The functional approach is wary of the generalized efficiency hypotheses espoused by the positive school. In this respect, the functionalists share some of this skepticism of the normative school. Nothing supports a generalized trust in the efficiency of the law in all areas of the law. Even more vocally, the functional school of law and economics is skeptical of a general efficiency hypothesis when applied to sources of the law other than common law (e.g., legislation or administrative regulations).

The functional approach is also critical of the normative extensions and ad hoc corrective policies, which are often advocated by the normative schools. Economic models are a simplified depiction of reality. Thus, functionalists think it is often dangerous to utilize such tools to design corrective or interventionist policies. In this respect, the functionalists are aligned with the positive school in their criticism of the normative approach. According to both the positivists and the functionalists, normative economic analysis often risks overlooking the many unintended consequences of legal intervention.

An important premise of the functional approach to law and economics is its reliance on methodological individualism. According to this paradigm of analysis, only individuals choose and act (see, e.g., Buchanan (1990) and the various contributions of the Virginia school of political economy). The functional approach to law and economics is informed by an explicit recognition that whatever social reality we seek to explain at the aggregate level, ought to be understood as the result of the choices and actions of individual human beings who pursue their goals with an independently formed understanding of the reality that surrounds them (Vanberg, 1994:1). Normative individualism further postulates that only the judgment of the single individuals can provide a relevant benchmark against which the merits of alternative rules can be evaluated.

The findings of public choice theory, while supporting much of the traditional wisdom, pose several challenges to neoclassical law and economics. In spite of the sophisticated mathematical techniques of economic analysis, judges and policymakers in many situations still lack the expertise and methods for evaluating the efficiency of alternative legal rules. Courts and policymakers should thus undertake a functional analysis. Such an analysis requires them to first inquire into the incentives underlying the legal or social structure that generated the legal rule, rather than directly attempting to weigh the costs and benefits of individual rules.[4] In this way, the functionalist approach to law and economics can extend the domain of traditional law and economics inquiry to include both the study of the influence of market and non-market institutions (other than politics) on legal regimes, and the study of the comparative advantages of alternative sources of centralized or decentralized lawmaking in supplying efficient rules.

3. Pareto, Bentham and Rawls: The Dilemma of preference aggregation

The need to make comparative evaluations between different rules motivates much of law and economics. Consequently, the second methodological problem in law and economics concerns the choice of criteria for carrying out such comparative analysis. In practical terms, this problem concerns the method of aggregation of individual preferences into social preferences. This problem is not unique to law and economics. It is part of a much larger methodological debate in economic philosophy and welfare economics.

Already in the late nineteenth century, Edgeworth (1881:7–8) stated the moral dilemma of social welfare analysis, observing that a moral calculus should proceed with a comparative evaluation of "the happiness of one person with the happiness of another. . . . Such comparison can no longer be shirked, if there is to be any systematic morality at all." The problem obviously arises from the fact that economists do not have any reliable method for measuring individuals' utility, let alone make inter-personal comparisons of utility.

Economic analysis generally utilizes one of the three fundamental criteria of preference aggregation.

3.1. Ordinal preferences and the Pareto criterion

The first criterion of social welfare is largely attributable to Italian economist and sociologist Vilfredo Pareto. The Pareto criterion limits the inquiry to *ordinal* preferences of the relevant

individuals. According to Pareto, an optimal allocation is one that maximizes the well-being of one individual relative to the well-being of other individuals being constant.[5] In normal situations, there are several possible solutions that would qualify for such a criterion of social optimality. For example, if the social problem is that of distributing a benefit between two parties, any hypothetical distribution would be Pareto optimal, since there is no possible alternative redistribution that would make one party better off without harming another party.

The Pareto criterion has been criticized for two main reasons: (a) it is *status quo* dependent, in that different results are achieved depending on the choice of the initial allocation; and (b) it only allows *ordinal* evaluation of preferences, since it does not contain any mechanism to induce parties or decision makers to reveal or evaluate *cardinal* preferences (i.e., the intensity of preferences). As a result of these shortcomings scholars (e.g., Calabresi, 1991), have questioned the usefulness of the Pareto criterion in its applications to law and economics.

3.2. Utilitarian tests: Bentham and Kaldor-Hicks

In the nineteenth and early twentieth century, economists and philosophers developed welfare paradigms according to which the degree of all affected individuals had to be taken into account in any comparative evaluation of different states of the world. This methodological trend, related to utilitarian philosophy, is best represented by philosophers and jurists such as Bentham (1839) and later economists such as Kaldor (1939), Hicks (1939), and Scitovsky (1941), who in different ways formulated criteria of social welfare that accounted for the *cardinal* preferences of individuals.

In *Principles of Moral and Legislation*, Bentham (1789) presents his theory of value and motivation. He suggests that mankind is governed by two masters: 'pain' and 'pleasure.' The two provide the fundamental motivation for human action. Bentham notes that not all individuals derive pleasure from the same objects or activities, and not all human sensibilities are the same.[6] Bentham's moral imperative, which has greatly influenced the methodological debate in law and economics, is that policymakers have an obligation to select rules that give "the greatest happiness to the greatest number." As pointed out by Kelly (1998:158) this formulation is quite problematic, since it identifies two maximands (i.e., degree of pleasure and number of individuals) without specifying the tradeoff between one and the other. Bentham's utilitarian approach is thus, at best, merely inspirational for policy purposes.

Later economists, including Kaldor (1939), Hicks (1939), and Scitovsky (1941), formulated more rigorous welfare paradigms which avoided the theoretical ambiguities of Bentham's proposition. However, these formulations presented a different set of difficulties in their implementation. The core idea of their approach is that state A is to be preferred to state B if those who gain from the move to A gain enough to compensate those who lose. The test is generally known as the Kaldor-Hicks test of potential compensation. It is one of "potential" compensation because the compensation of the losers is only hypothetical and does not actually need to take place.[7] In practical terms, the Kaldor-Hicks criterion requires a comparison of the gains of one group and the losses of the other group. As long as the gainers gain more than the losers lose, the move is deemed efficient. Mathematically, both the Bentham and the Kaldor-Hicks versions of efficiency are carried out by comparing the

aggregate payoffs of the various alternatives and selecting the option that maximizes such summation.

3.3. Non-linear social preferences: Nash and Rawls

Other paradigms of social welfare depart from the straight utilitarian approach, suggesting that social welfare maximization requires something more than the maximization of total payoffs for the various members of society. Societies are formed by a network of individual relations and there are some important interpersonal effects that are part of individual utility functions. Additionally, human nature is characterized by diminishing marginal utility, which gives relevance to the distribution of benefits across members of the group.

Imagine two hypothetical regimes: (a) in which all members of society eat a meal a day; and (b) in which only a random one-half of the population gets to eat a double meal while the other unlucky half remains starving. From a Kaldor-Hicks perspective, the two alternatives are not distinguishable from the point of view of efficiency because the total amount of food available remains unchanged. In a Kaldor-Hicks test, those who get a double meal have just enough to compensate the others and thus society should remain indifferent between the two allocational systems. Obviously, this indifference proposition would leave most observers unsatisfied. In the absence of actual compensation, the criterion fails to consider the diminishing marginal benefit of a second meal and the increasing marginal pain of starvation. Likewise, the randomized distribution of meals fails to consider the inter-personal effects of unfair allocations. Fortunate individuals suffer a utility loss by knowing that other individuals are starving while they enjoy a double meal. Because of the diminishing marginal utility of wealth and interpersonal utility effects, from an ex ante point of view, no individual would choose allocation system (b), even though the expected return from (b) is equal to the return from (a).

Scholars that try to evaluate the welfare implications of distributional inequalities generally do so by invoking Rawls' (1971)[8] theories of justice or by utilizing Nash's (1950)[9] framework of welfare.

The intuition underlying these criteria of welfare is relatively straightforward: the well-being of a society is judged according to the well being of its weakest members. The use of a algebraic product to aggregate individual preferences captures that intuition. Like the strength of a chain is determined by the strength of its weakest link, so the chain of products in an algebraic multiplication is heavily affected by the smallest multipliers. Indeed, at the limit, if there is a zero in the chain of products, the entire grand total will collapse to zero. This means that the entire social welfare of a group approaches zero as the utility of one of its members goes to zero.

In the law and economics tradition, these models of social welfare have not enjoyed great popularity. This is not so much for an ideological preconception but rather for a combination of several practical reasons. These reasons include the general tendency to undertake a two-step optimization in the design of policies, and the difficulties of identifying an objective criterion for assessing interpersonal utility and diminishing marginal utility effects. From a methodological point of view, distributional concerns are generally kept separate from the pursuit of efficiency in policymaking. Such separation has been rationalized on the basis

that the legal system is too costly an instrument for distribution, given the advantage of the tax system for wholesale reallocation of wealth (e.g., Kaplow-Shavell, 1994).

4. Wealth, utility and revealed preferences: The choice of maximand

There is an important methodological question that has openly engaged the attention of prominent law and economics scholars: What should the legal system try to maximize? In this debate, even strict adherents to the instrumentalist view of the law may question whether the objective of the law should be the maximization of aggregate wealth, aggregate utility, or merely provide the conditions for free individual choice.

If the scholars involved in these debates could look at the issue as neutral spectators, consensus could be reached on the idea that, the ultimate policy goal is the maximization of human happiness and well-being. But regardless of such an observation, economic analysis of law rarely uses utility-based methods of evaluation. The reason for this is, once again, mostly pragmatic. Unlike wealth (or quantities of physical resources), utility cannot be objectively measured. Furthermore, inter-personal comparisons of utility are impossible, rendering any balancing across groups or individuals largely arbitrary. These limitations make utility maximization unviable for practical policy purposes.

Given the above limitations, following Posner, several practitioners of economic analysis of law have departed from the nineteenth century utilitarian ideal of utility maximization[10] Rather, they have increasingly used a paradigm of wealth maximization. Several scholars in law and economics remain uneasy in accepting the notion of wealth maximization as an ancillary paradigm of justice. Although several of the differences prove to be largely verbal two objections continue to affect the lines of the debate.

The first objection relates to the need for specifying an initial set of individual entitlements or rights as a necessary prerequisite for operationalizing wealth maximization. In this context, one can think of the various criticisms of wealth maximization by property right advocates who perceive the social cost of adopting such criterion of adjudication as very high, given wealth maximization's instrumentalist view of individual rights and entitlements. These critics argue that rights have value that must be accounted for outside of how useful they might be to the accumulation of wealth (Buchanan, 1974; Rowley, 1989).

The second objection springs from the theoretical difficulty of defining the proper role of efficiency as an ingredient of justice, vis-à-vis other social goals. Legal scholars within the law and economics tradition (see, e.g., Calabresi, 1980) have claimed that an increase in wealth cannot constitute social improvement unless it furthers some other social goal, such as utility or equality. Denying that one can trade off efficiency against justice, these scholars argue instead that efficiency and distribution are equally essential elements of justice, which is seen as a goal of a different order than either of its constitutive elements.

The functional school of law and economics provides a third alternative by identifying individual choice and revealed preferences as the fundamental criterion for evaluation. The design of metarules that are aimed at fostering free individual choice by eliminating strategic and transactional impediments to the revelation of true preferences becomes an explicit objective of the functional school. As discussed above, the evaluation of alternative sources of law requires an appropriate analysis of the incentive structure in the originating environment

and is aimed at introducing market-like mechanisms in the creation and selection of legal rules, with an emphasis on institutional mechanism design and individual choice. The recent literature on reciprocity (Smith, McCabe and Rassenti, 1998; Fon and Parisi, 2003), social norms and customary law (Cooter, 2000; Parisi, 1998), choice of law (Romano, 1999; Parisi and Ribstein, 1998; Ribstein and O'Hara, 2000), federalism (Ribstein and Kobayashi, 2001) and freedom of contract (Trebilcock, 1994; Buckley, 1999) are examples of the growth and value of functional approaches in law and economics.

Future generations of law and economics scholars should be cognizant of the important methodological debates that have engaged their precursors, taking full advantage of the insights developed by the different methodological traditions when appraising legal rules and institutions.

Notes

1. Posner (1974) offers crime as an example. Positive law and economics can help explain and predict how various punishments will affect the behavior of criminals. It might determine that a certain sanction is more likely to deter a certain crime. While this analysis does not by itself mean that the law should be adopted, it can be used to influence normative analysis on whether the law would be beneficial to society.
2. MacKaay (2000) observes that the Yale school considers market failures to be more pervasive than Chicago scholars are willing to admit. Legal intervention is believed to be the appropriate way of correcting such failures, although it may not succeed in all circumstances.
3. For a brief intellectual history of the three approaches to law and economics, see Posner and Parisi (1998).
4. On this point, see Cooter (1994) introducing the similar idea of structural adjudication of norms.
5. As a corollary, a change to a Pareto superior alternative makes someone better off without making anyone worse off.
6. See Posner (1998) for an interesting discussion on Bentham and his influence on the law and economics movement.
7. One should note that, if actual compensation was carried out, any test satisfying the Kaldor-Hicks criterion of efficiency would also satisfy the Pareto criterion.
8. Notable scholars have considered the conditions under which principles of justice can emerge spontaneously through the voluntary interaction and exchange of individual members of a group. As in a contractarian setting, the reality of customary law formation relies on a voluntary process through which members of a community develop rules that govern their social interaction by voluntarily adhering to emerging behavioural standards. In this setting, Harsanyi (1955) suggests that optimal social norms are those that would emerge through the interaction of individual actors in a social setting with impersonal preferences. The impersonality requirement for individual preferences is satisfied if the decision makers have an equal chance of finding themselves in any one of the initial social positions and they rationally choose a set of rules to maximize their expected welfare. Rawls (1971) employs Harsanyi's model of stochastic ignorance in his theory of justice. However, the Rawlsian "veil of ignorance" introduces an element of risk aversion in the choice between alternative states of the world, thus altering the outcome achievable under Harsanyi's original model, with a bias toward equal distribution (i.e., with results that approximate the Nash criterion of social welfare). Further analysis of the spontaneous formation of norms and principles of morality can be found in Sen (1979), Ullmann-Margalit (1977), and Gauthier (1986).
9. According to the Nash criterion, social welfare is given by the product of the utility of the members of society (Nash, 1950). See Mueller (1989:379–382), attributing the multiplicative form of the social welfare function to Nash.
10. Posner is the most notable exponent of the wealth maximization paradigm. Under wealth-maximization principles, a transaction is desirable if it increases the sum of wealth for the relevant parties (where wealth is meant to include all tangible and intangible goods and services). Already Bentham (1839) challenged the

use of objective factors, such as wealth or physical resources, as a proxy for human happiness. Despite the difficulties in quantification of values such as utility or happiness, the pursuit of pleasure and happiness and the avoidance of and pain are the motivating forces of human behavior. Wealth, food and shelter are mere instruments to achieve such human goals.

References

Backhaus, J. (ed.) (2003). *The Elgar Companion to Law and Economics*, Cheltenham, UK: Edward Elgar Publishing.

Bentham, J. (1782). *Of Laws in General*.

Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation*.

Bentham, J. (1839). *A Manual of Political Economy*, New York: G.P. Putnam.

Buchanan, J. (1974). "Good Economics-Bad Law." *Virginia Law Review*. 60, 483–492.

Buchanan, J.M. (1990). "The Domain of Constitutional Economics." *Constitutional Political Economy*. 1(1), 1–18.

Buckley, F.H. (1999). *The Fall and Rise of Freedom of Contract*, Durham: Duke University Press.

Calabresi, G. (1980). "About Law and Economics: A Letter to Ronald Dworkin." *Hofstra Law Review*. 8, 553–562.

Calabresi, G. (1991). "The Pointlessness of Pareto: Carrying Coase Further." *Yale L.J.* 100, 1211.

Coase, R.H. (1960). "The Problem of Social Cost." *Journal of Law and Economics*. 3, 1–44.

Coase, R.H. (1978). "Economics and Contiguous Disciplines." *Journal of Legal Studies*. 7, 201–211.

Cooter, R. (1984). "Prices and Sanctions." *Columbia Law Review*. 84, 1523–1560.

Cooter, R.D. (1994). "Structural Adjudication and the New Law Merchant: A Model of Decentralized Law." *Int. Rev. of Law & Econ.* 14, 215.

Cooter, R.D. (2000). "Three Effects of Social Norms on Law: Expression, Deterrance, and Internalization." *Oregon Law Review*. 79(1), 1.

Edgeworth, F.Y. (1881). *Mathematical Psychics*, London: C. Kegan Paul.

Ehrlich, I. & Posner, R.A. "An Economic Analysis of Legal Rule Making." *Journal of Legal Studies*. 3, 257–186.

Fon, V. & Parisi, F. "The Limits of Reciprocity for Social Cooperation." George Mason Law & Economics Research Paper No. 03-08.

Gauthier, D. (1986). *Morals by Agreement*, Oxford: Clarendon Press.

Harsanyi, J.C. (1955). "Cardinal Welfare Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy*. 63, 315.

Hicks, J.R. (1939). "The Foundations of Welfare Economics." *Economic Journal*. 49, 696–712.

Kaldor, N. (1939). "Welfare Propositions of Economics and Inter-personal Comparisons of Utility." *Economic Journal*. 49, 549–552.

Kaplow, L. & Shavell, S. (1994). "Why the Legal System Is Less Efficient than the Income Tax in Redistributing Income." *Journal of Legal Studies*. 23, 667–681.

Kelly, P. & Bentham, J. (1998). *New Palgrave Dictionary of Economics and the Law*, vol. 1.

MacKaay, E. (2000). "History of Law and Economics." *Encyclopedia of Law and Economics*, vol. 1, Cheltenham, UK: Edward Elgar Publishing.

Mueller, D. (1989). *Public Choice II*, Cambridge, UK: Cambridge University Press.

Nash, J., (1950). "The Bargaining Problem." *Econometrica*. 18, 155.

Parisi, F. (1998). "Customary Law." *The New Palgrave Dictionary of Economics and the Law*, vol. 1.

Parisi, F. & Ribstein, L. (1998). "Choice of Law." *The New Palgrave Dictionary of Economics and the Law*, vol. 1.

Posner, R.A. (1985). "Wealth Maximization Revisited." *Notre Dame Journal of Law, Ethics, and Public Policy*. 2, 85–105.

Posner, R.A. (1988). "The Ethics of Wealth Maximization: Reply to Malloy." *Kansas Law Review*. 36, 261.

Posner, R.A. (1998). "Bentham's Influence on the Law and Economics Movement." *Current Legal Problems*. 51, 425.

Posner, R.A. & Parisi, F. (1998). "Scuole e Tendenze nella Analisi Economica del Diritto." *Biblioteca della Liberta'*. 147, 3–20.

Priest, G.L. (1977). "The Common Law Process and the Selection of Efficient Rules." *Journal of Legal Studies*. 6, 65–82.

Rawls, J. (1971). *A Theory of Justice*, Cambridge, MA: Harvard University Press.

Ribstein, L. & Kobayashi, B. (2001). "A Recipe for Cookies: State Regulation of Consumer Marketing Information." George Mason Law & Economics Research Paper No. 01-04.
Ribstein, L. & O'Hara, E.A. (2000). "From Politics to Efficiency in Choice of Law." *University of Chicago Law Review*. 67, 1151.
Romano, R. (1999). "Corporate Law as the Paradigm for Contractual Choice of Law." In F. Buckley (ed.), *The Fall and Rise of Freedom of Contract*, Durham, NC: Duke University Press.
Rowley, C.K. (1981). "Social Sciences and the Law: The Relevance of Economic Theories." *Oxford Journal of Legal Studies*. 1, 391–405.
Rowley, C.K. (1989). "The Common Law in Public Choice Perspective: A Theoretical and Institutional Critique." *Hamline Law Review*. 12, 355–383.
Rubin, P.H. (1977). "Why is the Common Law Efficient?" *Journal of Legal Studies*. 6, 51–63.
Schotter, A. (1981). *Economic Theory of Social Institutions*, Cambridge, Mass: Harvard University Press.
Scitovsky, T. (1941). "A Note on Welfare Propositions in Economics." *Review of Economic Studies*. 9(1), 77–88.
Sen, A.K. (1979). "Rational Fools: A Critique of the Behavioural Foundations of Economic Theory." In F. Hahn and M. Hollis (eds.), *Philosophy and Economic Theory*, Oxford: Clarendon Press.
Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*.
Smith, V., McCabe, K. & Rassenti, S. (1998). "Reciprocity, Trust and Payoff Privacy in Extensive Form Bargaining." *Games and Economic Behavior*. 24, 10–24.
Trebilcock, M.J. (1994). *The Limits of Freedom of Contract*, Cambridge, MA: Harvard University Press.
Ullmann-Margalit, E. (1977). *The Emergence of Norms*, Oxford: Clarendon Press.
Vanberg, V.J. (1994). *Rules and Choice in Economics*, London: Routledge.

Introduction

Astronomers have discovered a solar formation they call a "binary star." This formation consists of two suns, each in orbit around the other. Their center of gravity lies at a point in between them, and they revolve around that center of gravity. Neither star could remain where it is, or as it is, without the other. They are two separate bodies, but each is dependent on the other for its place in the universe.

The tort liability and insurance systems are very much like the two suns in a binary star, dependent on each other for their position in our legal system. For more than a century these two systems have influenced each other's course of development. Neither would be anything like what it is today if the other had not existed and developed along with it. Today the two systems constantly interact, and almost no effort to understand or reform one of them can take place without understanding the role played by the other.

There is hardly any aspect of contemporary life in the United States that goes untouched by tort law or insurance. Virtually every activity that a business or professional contemplates must be scrutinized in advance for its possible liability and insurance implications.[1] Not only are both the tort and insurance systems enormously important from both an economic and social standpoint, but neither stands alone. They interact and influence each other's development, shape, and scope. The tort system, not only as it exists on paper but also how it works in practice, is a product of the insurance system, just as the insurance system is a product of the tort system.

The Two Systems and Their Interaction

We can begin with a simple understanding of what insurance is. As I shall use the term, "insurance" is a transaction that performs two functions: risk transfer and risk distribution, or spreading. Many transactions fit this description. But in insurance the transferee of risk, the insurer, accepts the transfer of risks as part of its business and not as the incident of another transaction. In accepting the transfer of a large number of risks, the insurer diversifies and thereby reduces its overall risk, by taking advantage of the law of averages. For the insurer, the whole risk is smaller than the sum of its parts. The insurer is thus a vehicle for distributing the risk of loss among its policyholders.

In the Unites States, insurance figures prominently in both tort liability and in nontort methods of compensating the victims of injury, illness, and death. We have vast systems of private insurance that provide health, disability, life, auto, workers' compensation, liability, and property insurance to businesses, individuals, and families. All told we spend approximately $900 billion annually on these different forms of private insurance—about 7 percent of the country's gross domestic product—plus an additional $600 billion dollars each year for publicly provided "social" insurance such as Medicare, Medicaid, and Social Security Disability insurance. The total compensation that insurance pays for injury, illness, and death in this county is thus in the neighborhood of $1.5 trillion per year.

These extensive sources of insurance developed as components of the unique and partial system of social welfare that emerged in the United States over a period of more than a century.[2] Tort liability expanded at the same time the social welfare system developed, in part because that system has never provided more than incomplete protection. More important, tort liability expanded because the movement toward greater social welfare reflected the increasing socialization and spreading of risk. The expansion of tort liability was part of this trend. More tort liability helped to socialize the risk of suffering accidental personal injury. But the particular ways tort liability expanded, the particular new forms of liability that developed, and the particular shape of the current tort system were influenced by the sources of insurance that were developing at the same time as tort law. Insurance was a principal mechanism by which tort law spread risk. And so new forms of insurance developed along with the expansion of tort liability, sometimes before and sometimes after a new form of tort liability was created.

But of course tort is not only a part of our system of social welfare; it is also a part of the extensive system of health and safety regulation that emerged in the United States at the same time as our social welfare system. As we will see below, the multiple functions that tort and liability insurance are now expected to perform make it almost inevitable that in practice the way they operate will fall short of the ideal.

The tort system and the liability insurance that commonly accompanies tort thus function alongside, and arguably as part of, our larger system of compensation for the consequences of injury, illness, and death. As significant as tort may be, however, it pales in comparison to the roughly $1.5 trillion per year system of compensation of which it is a part. I estimate that for the portion of tort that addresses personal injury, illness, and death, the direct cost of the tort system is approximately $200 billion—about 13 percent of the larger system.[3] A very substantial proportion of these direct tort costs—surely 75 percent, and probably more—are covered by liability insurance. Annual tort costs today reflect a more than one-hundred-fold cost increase since 1950, when tort costs were less than $2 billion. The increase in direct tort costs during this period has outstripped economic growth by a factor of 3. It is impossible, however, to estimate the additional indirect costs of the tort system that are in incurred in the form of forgone productivity, research, and development. But it would also be extremely difficult to estimate or place a value on the benefits of the tort system that accrue from the investment we make in it. These benefits include the health and life saved from damage or destruction by the additional safety the threat of liability generates, as well as reduction of the adverse effects of illness, injury, and death that tort achieves by providing monetary compensation for at least some of the consequences of the tortiously caused harm that does occur.

The interaction of tort, liability insurance, and these other sources of insurance-based compensation is continuous and complex. I will provide just two brief examples of this phenomenon, since in a sense this entire book is comprised of examples of how, often in unseen ways, tort law and insurance interact with each other at both the structural and operational levels. Perhaps the most politically visible tort law issue these days is whether to place monetary ceilings on the amount of damages for "pain and suffering" that are available in tort suits. But participants in this debate rarely recognize or acknowledge the role that liability insurance has played in making it possible for this issue to arise at all. Only because liability insurance is so readily available to individuals and small businesses can law-

suits claiming hundreds of thousands or even millions of dollars for pain and suffering from such defendants even be imagined.

Tort liability and liability insurance have an interactive impact on each other that has resulted in a decades-long arms race. Often the sequence has begun with tort law expanding the scope of liability or permitting ever-larger recoveries. Liability insurance then responds by providing insurance against the new liabilities and greater amounts of coverage. Escalation then proceeds another rung up the ladder, as tort law targets defendants covered by the new forms of insurance and imposes even larger amounts of liability, in order to ensure that there is sufficient deterrence of now-insured conduct, and because prior increases in awards have ratcheted up what are now considered appropriate levels of compensation. In other settings the sequence of interaction takes place in reverse. Here liability insurance comes into existence first, and tort law then seeks it out by creating new forms of liability, at least partly in response to the availability of this insurance as a source of compensation. Whether it is tort liability or liability insurance that has come first, however, they have had a reciprocal, one-way ratchet effect on each other that has resulted in both increasingly large tort recoveries and increasingly large amounts of liability insurance that covers these liabilities. Indeed, in certain domains this effect has been so great that bankruptcy and other liability-limiting devices are sometimes employed to create partial shelter from these liabilities.[4]

A second prominent subject of tort reform debate is whether to modify the "collateral source" rule that governs whether tort plaintiffs are allowed to recover damages for losses that their health and disability insurance have already paid, or whether these sums should instead be deducted from tort awards. This can be a significant issue, however, only because of the proliferation of health and disability insurance that has become available to the vast majority of the U.S. population in the last 50 years. If hundreds of billions of dollars of health and disability insurance benefits were not paid each year, the proper relationship between these different sources of compensation would have little practical importance.

In addition, the ready availability of health and disability insurance to tort victims actually may have been helping to inflate the amounts that are awarded in tort, in two ways. First, in virtually all routine cases involving comparatively small amounts of damages, and even in some cases involving substantial damages, a plaintiff's medical expenses are used as a baseline for computing settlement offers. This approach is taken because in practice the

damages that juries award for pain and suffering tend to be roughly proportional to the amount of medical expenses the plaintiff has incurred on account of his or her injuries. As we will see in later chapters, the more health insurance that is available to tort victims, the higher their medical expenses will be, and consequently the larger tort settlements and awards are likely to be. Expanding health insurance may thus have a multiplier effect on tort payments. Second, health and disability insurance benefits that are paid soon after an injury occurs provide tort plaintiffs with access to what amounts to a capital market, thereby giving plaintiffs greater capacity to resist "low-ball" offers of settlement from defendants or their liability insurers. Because plaintiffs already have insurance against their out-of-pocket losses, they feel less pressure to settle their tort cases. As a result, these cases are either litigated or settled for more than they would have been in the days before health and disability insurance were as widespread as they are now. Paradoxically, then, instead of reducing the importance of tort, the spread of health and disability insurance probably has helped to fuel the expansion of tort liability.

The pain and suffering and collateral source examples reveal the way certain features of our system are so fundamental that they are almost invisible. Yet often it is precisely the hidden interactions between tort and insurance that make the tort system what it is, and that generate reform issues in the first place. Tort and insurance have become so inextricably and unavoidably intertwined that often one system is not even seen to influence features of the other that would not and could not have come into being, or would not continue to exist in their current form, without the other system. Unless these relationships are recognized and understood, proposals for tort reform may prescribe solutions that do not really address the concerns they are designed to solve.

Why This Project?

Over the years many legal scholars have thought that insurance influences tort law in important ways, but have rarely been able to say exactly how that influence operates. Insurance has been a bit like the skunk at a lawn party, present in everyone's mind but nonetheless ignored. Beginning as early as the enterprise liability scholarship of William O. Douglas, and continuing with the initial work of Fleming James, Jr., a decade or so later, legal scholars argued for the expansion of tort liability on the ground that potential

defendants can more easily insure against liability than potential victims.[5] The enterprise liability theory was, among other things, an insurance rationale for expanding tort liability.

More subtly, a later generation of scholars, led most prominently by Guido Calabresi, employed economic analysis in their study of tort law.[6] They brought rigorous concern with the deterrent effect of tort liability into tort theory. Many also thought that distributional concerns, including access to insurance on the part of both potential injurers and potential victims, could appropriately figure in the determination of liability rules. They recognized, however, that these considerations did not always argue in favor of expanding liability. But neither James nor Calabresi, nor those who have for several decades been writing in the shadow of their work, had much to say about how insurance actually figures in these determinations. The insurance rationale for tort liability has been attacked by such scholars as Richard Epstein and George Priest, on the ground that the rationale has in fact been operating not only in tort scholarship but also in judicial decisions, and that the rationale has created problems for both tort liability and insurance.[7] And that rationale has been defended by such scholars as Jon Hanson, Kyle Logue, and Gregory Keating.[8] On the other hand, as prominent a figure as Dean William Prosser, author of the leading torts treatise of the twentieth century, argued that the availability of insurance actually has only rarely figured in the shaping of tort law doctrine.[9]

It is time to try to make sense of, and where possible to disentangle, these conflicting contentions, and to explore the relationship between tort and insurance. We need to get a handle on how insurance has influenced tort law, and how tort law has influenced insurance. It is one thing to assert in a general way that the availability of insurance influences the course of tort law development, or that when tort liability expands, insurance against the new form of liability is likely to be created. It is quite another thing to understand how this has and has not actually occurred in different fields of tort liability.

The main reason that this kind of examination has never been undertaken, I think, is that there is almost no one who is interested enough in both fields to have done it. Torts scholars have stayed away from insurance because it is in a sense a subspecialty of the field of contract law, and because it is highly technical. Learning about insurance and insurance law requires a substantial investment of intellectual capital. Torts is a required subject in all law schools, and many law professors teach the subject. But in-

surance law is an elective course, and for some people insurance has an ambulance-chasing air about it that tends to undermine its respectability. Securities law, in contrast, certainly has no greater economic or legal importance than insurance law but is generally held in higher regard as a subject of scholarship. At the very least it is a more prominent subject. For these reasons, there are far fewer insurance law scholars than there are tort scholars. Conversely, outside of the law schools there are insurance specialists in schools of business and departments of economics, but they tend not to know much, if anything, about tort law. The result is that there is very little overlap between those who write about torts and those who write about insurance and insurance law.

e

Evaluating Markets versus Mandates for Environment and Energy

August 29, 2023

Coinciding with the Hoover Institution's new tagline on "Ideas Advancing Freedom", the Institution is launching a new program on "Markets vs. Mandates for Environment and Energy." Here we describe the connection between the two and the basis for a new book on the topic.

Markets vs. Mandates is based on the Institution's founding principles of generating ideas that advance healthy and free societies. The program will examine tradeoffs of market-driven and mandate approaches for promoting environmental quality and energy efficiencies. Whereas mandates rely on politics and administrative rule-making to assign objectives and command-and-control mechanisms to achieve them, markets rely on the incentives of resource owners and entrepreneurs to voluntarily discover and respond to the changing environmental demands of populations – whether this is for stable climate, cleaner air or water, or more biodiversity.

The program and book will address questions such as the following. When are markets or mandates more likely to improve actual environmental conditions? Which approach entails lower costs and fewer unwanted consequences? And how can societies achieve higher environmental quality without sacrificing freedom?

The Short Run

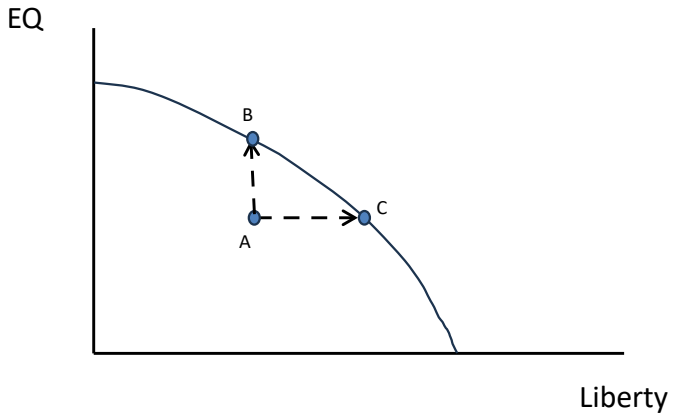
The following diagram illustrates program goals in the context of tradeoffs. On the vertical axis is a composite measure of environmental quality, be it clean air, water, or more biodiversity.[1] On the horizontal axis is a measure of liberty for, say, the median member of a society. The possibilities frontier assumes a tradeoff between freedom and environmental quality, which is usefully descriptive only in the short run. Here environmental quality might be improved in the short run by restricting human liberty to burn coal, use gas stoves, or cut trees,

[1] A composite measure would theoretically account for situations where improvements in one environmental indicator causes another to decline as discussed in later chapters. For example, wind farms may reduce CO_2 emissions but kill migratory birds. Electric vehicles may reduce CO_2 but mining for battery minerals contaminates soil and water systems. Research shows that Americans remain more concerned about water quality, soil contamination, and air quality when compared to climate change (see Keiser and Shapiro 2019, Fig. 1).

for example. If different societies have different preferences, then preferred locations along the frontier would naturally vary. The median resident of Oregon is different than that of Texas, for example, in her willingness to trade liberty for environmental quality.

Whatever the preferences, it is never ideal for land, resource, and energy use to be governed by institutions that position a society within the frontier. Any movement from point A towards B yields higher environmental quality without a loss in liberty. And movement from A towards C yields the same environmental quality with a gain in liberty. More generally any point along the frontier between B and C yields increased liberty and environmental quality. *Markets vs. Mandates* is about finding institutions and policies that do just that.[2]

Figure 1: Short-Run Tradeoff between Environment and Liberty



The Long Run

In the long run the relationship between freedom and environmental quality is likely reversed. At the very least the tradeoff is flatter, or non-existent. This is because economic freedom tends to promote economic prosperity and innovation, and both can promote clean environments and energy efficiencies.

In this regard, the Environmental Kuznets Curve (EKC) challenges those in the science and policy community who have viewed economic growth and economic freedoms as leading causes of environmental degradation. The EKC is a hypothesized inverted U-shaped relationship between income per capita (the x-axis) and degradation/emissions/pollution (the y-axis) that is

[2] Institutions that support adaptation – to climate or air or water pollution - also play a role. In the diagram, such institutions flatten the tradeoff between liberty and environmental quality (i.e., they lower the relative price of liberty).

now standard in economic texts. Its main implication – that growth will eventually improve environmental quality – has also gained traction in the policy arena. This does not mean that income growth is always a superior substitute for environmental policy but it does raise the possibility that environmental policies initially effective at reducing emissions can actually reduce long-run environmental quality if they also reduce growth.[3] Aggressive renewable energy subsidies, for example, could potentially impair more effective long-run responses to climate change by misallocating labor and capital away from their highest economic use thereby lowering growth.

The view that interventions could reduce long-run environmental quality is most strongly supported if EKC relationships are robust to different environmental indicators (e.g., SO_2 and CO_2), if income growth causes EKC-consistent changes in the environment rather than simply being correlated with them, and if the channels through which growth improves the environment are primarily market driven. Though there are challenges to evaluating all three conditions here we offer four broad-brush observations.

The first is that much of the EKC literature focuses on SO_2 , NO_2, and fine particulate matter finding strong evidence of turning points across and within countries (see Parker 2010). This bears out in obvious ways. Air quality in the U.S. was poor in many American cities during the 1950s and 1960s but it improved dramatically as incomes grew. Chinese cities are the poster child for poor air quality today, but air quality in India, Bangladesh, Pakistan and other developing economies is among the world's worst.[4]

The literature has less to say about greenhouse gas emissions which are invisible and odorless. We do know that U.S. CO_2 emissions peaked in 2007 at 6,016 metric tons and decreased to 4,970 in 2022, a 17% decline.[5] During the same period the U.S. population increased 10% and real per capita income increased 28%.[6] [More here about time and income patterns of CO_2 emissions across countries with highlights on China and India].

[3] The evidence is clear that low incomes and poverty impair climate adaptation and more generally adaptation to environmental pollution as discuss in later chapters (see, e.g., Carleton et al. 2022, Anderson 2021, Kahn 2021).

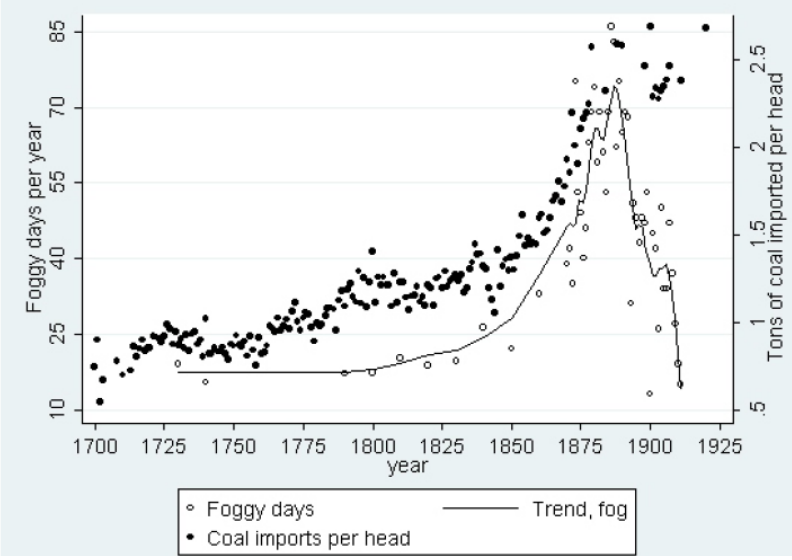
[4] See Air Quality Life Index created by the University of Chicago's Energy Policy Institute (see https://aqli.epic.uchicago.edu/the-index/).

[5] See https://www.statista.com/statistics/183943/us-carbon-dioxide-emissions-from-1999/

[6] See https://fred.stlouisfed.org/series/POPTOTUSA647NWDB and https://fred.stlouisfed.org/series/RPIPCUS

Second, there is case study evidence of the importance of both environmental regulations and pure market forces in prompting turning points in emissions but little systematic evidence of their relative importance. [7] Consider, for example, Clay and Troesken's (2010) evidence of the "world's first EKC". The find that, "if one were to plot real wages for unskilled laborers in London against foggy days per annum the typical EKC would emerge" with the turning point occurring in 1891 (p. 52) (Fig. 2). The evidence in London is consistent with regulation causing the turning point. The passage of the Public Health Act of 1891 (which fined businesses for generating excessive smoke) coincided almost exactly with the peak of annual foggy days.[8] And the U.S. Clean Air Act, described below, appears to have caused sharp decreases in SO_2 and NO_2 emissions in some areas of the country coincident with its adoption in 1970.

Figure 2: Foggy Days in London



Source: Clay and Troesken (2010, 71)

[7] It is difficult to differentiate the role of regulations from that of market driven forces beyond case studies because credible and comparable indices of environmental regulations are generally not available across countries. The theoretical models tend to be very general in this regard and therefore provide few specific hypotheses for testing. For example, in the Andreoni and Levinson (2001) model, economies of scale in pollution abatement can generate an EKC. Their reasoning is consistent with there being a fixed cost to setting up environmental regulatory mechanisms or with establishing property rights that give individuals stronger incentives to conserve. It is also consistent with there being fixed costs of switching to cleaner technologies that are not worth undertaking at small scales.

[8] The authors are careful to note that factors other than the regulation may have contributed to the improvement in air quality.

On the side of market forces, consider the time path of U.S. carbon emissions, which peaked in 2007 and then has gradually declined. The Great Recession may have initially contributed to the decline but the continued trend is due to new fracking technologies that made formerly inaccessible deposits of shale gas profitable to extract. This market-driven technology (thanks in part to high oil and gas prices) caused a sharp shift from coal to natural gas in U.S. electricity provision. This reduced SO_2 and NO_2 emissions near power plants, leading to fewer premature respiratory deaths (see Fell and Kaffine 2018, Johnsen et al. 2019). [Add other examples.]

The fourth observation is that it is difficult to know if income growth actually causes environmental improvements even though this interpretation is intuitive and consistent with economic theory and evidence suggesting that cleaner environments are normal goods, meaning demand for them increases as incomes rise. Later chapters will consider alternative explanations, such as reverse causation. For example, there is some evidence that bad air quality and heat may decrease income by decreasing labor productivity, although questions remain about implications for long-run growth and productivity net of adaptation. In the case of climate, a Biden administration white paper from his council of economic advisors projects negligible effects on U.S. growth through 2100, even under extreme warming scenarios, meaning that productivity is predicted to be almost independent of whatever shocks climate change might bring.[9]

A more important alternative casual story in our minds is that jurisdictions may tolerate pollution only until a serious environmental crisis is looming and then react strongly to promote abatement. If degradation and income are both growing over time, then the two will be positively correlated (as in the upward slope of an EKC). Once the community reacts – through individual action, new regulations, or market-based reforms like the removal of agricultural subsidies for water withdrawal or subsidies for coal-fired electricity – emissions start to drop. As long as income continues to follow a time trend, this process will produce a downward slope of an EKC. This story is consistent with some examples given by Libecap (2008), who argues that property rights and market reforms over the management of fisheries and air quality tend to be delayed until environmental crisis is imminent under the status quo regime.

[9] See https://www.whitehouse.gov/wp-content/uploads/2023/03/CEA-OMB-White-Paper.pdf

Libecap's emphasis on property rights, economic freedoms, and markets points to their enhancement as a fundamental driver behind the income-environment relationships as we emphasize throughout this book. Yandle et. al. (2004, 26) presents the logic in a concise way.

> "Property rights enforcement leads to higher income levels, which in turn generate demand for environmental quality. Strong property rights institutions also provide a legal basis for taking action against those who generate pollution that degrade property values. They also provide incentives for investing in [environmental improvements] where payoffs generally do not come for many years."

Delineating Mandates and Markets

Moving away from macroeconomic correlations and zooming in allows us to think more precisely about the tradeoffs of mandates and markets. To do so requires definitions and examples of the two. The key differences lie in how goals are set, and how they are supposed to be achieved.

On one end of the continuum is the most extreme form of mandate in which politicians and public administrators determine goals and employ command-and-control techniques in an attempt to achieve them. Early versions of the U.S. Clean Air Act and the U.S. Endangered Species Act fit this description as discussed below. So do national, state, and local bans on hunting, pesticide use, plastic bags, gas stoves, gas-powered vehicles, hydraulic fracking, wind farming, nuclear energy, and electricity from fossil fuels. In these cases the goal is zero use and the command-and-control technique is complete prohibition.

On the other end of the continuum is the free market. In it, individuals and businesses voluntarily employ resources in response to prices and contractual incentives. These actions affect environmental quality indirectly as when emissions are a byproduct of production and consumption. Market participants also affect environmental quality and energy use directly, as when resource owners and entrepreneurs contribute to environmental improvements in response to the private demands of populations. For example, private businesses provide green burial services, line-caught seafood, and energy efficient appliances (see Anderson and Parker 2013 for more examples). Environmental groups can also buy land to leave conserved as it is, and they can potentially buy coal to keep it in the ground (Harsted 2012). Through these processes, environmental quality is determined not by goals set by public officials but instead by the decentralized actions of market participants seeking profit and happiness under financial constraints.

Many real-world examples lie between the extremes of command-and-control and free markets (see Table 1). When goals are mandated, the methods for achieving those goals can involve markets as in cap-and-trade systems, which set aggregate limits on resource extraction or pollution and allow quota holders to trade with one another. Market trading improves the efficiency in which state-determined targets can be reached. For example, individual tradable quota systems have proven to be effective at reducing economic and environmental waste in ocean fisheries (Costello et al. 2008, Grainger and Parker 2013) as have SO_2 trading permits among U.S. power plants subjected to Clean Air Act regulations (Stavins 2011). Allowing trade channels production towards producers with comparative advantages in clean production, and it provides quota holders with incentives to innovate and find cleaner ways to produce.[10]

.

Table 1: Taxonomy of Mandates and Markets

Label	***Goal Setting***	***Methods for Achieving***	***Examples***
Command & Control	Mandates	Mandates	Plastic bag bans, CAFÉ standards
Cap & Trade	Mandates	Markets	ITQs in fisheries; SO_2 trading under CAA; water quality trading in some states
Taxes & Subsidies	Markets	Mandates	Renewable energy; plastic bag fees; carbon taxes; carbon credits; wetland credits
Mandatory Disclosures	Markets	Markets with Disclosure	Mandated disclosures for conflict minerals, home energy efficiency, and carbon emissions
Laissez Faire	Markets	Markets	Ben & Jerrys; buy land; buy coal; Tesla before credits

Unlike mandated caps on pollution or mandated uses of renewable energy, taxes and subsidies are pricing rather than quantity instruments. This means markets will ultimately determine the quantity of resource use, pollution, and production but politicians play a critical role by setting taxes and subsidies. A carbon tax would raise the price of coal relative to natural gas, for example, and economists expect this would reduce coal use by an amount that depends

[10] The trade in a cap-and-trade system could potentially determine the cap, making this system tilt more towards laissez faire. Consider a carbon emissions cap set far above current emission levels. If environmental groups were allowed to bid and retire CO_2 quota - thereby expressing their demand for lower carbon - the cap could eventually reflect something like a market equilibrium.

on the size of the tax (set by politics) and the elasticity of demand (determined by economics). Requirements that companies buy carbon or wetland offset credits also act as taxes; in the case of carbon on fossil fuel use and in the case of wetlands on the development of properties connected to watersheds. Subsidies for renewables are common and market responses to develop wind and solar have varied within Europe, the U.S., and elsewhere.

A lighter-handed mandate is when governments require disclosures about their supply chains, production processes, or emission outputs. For example, some jurisdictions require home sellers to provide buyers with certified audits of residential energy efficiency (see Myers et al., forthcoming). And through Section 1502 of the 2010 U.S. Dodd-Frank Act, the Security and Exchange Commission (SEC) requires companies to disclose when their supply chains include purchases from mines that fund rebel groups in Africa (see Parker et al. 2017). The SEC's humanitarian watchdog role would expand into climate if it requires U.S. corporations to disclose their carbon emissions, as advocated by some economists (see Greenstone et al. 2023). Here the government does not determine output, but it requires information that may help markets better function but with potential unwanted consequences as described in later chapters.

It is yet to be determined how some contemporary climate and energy policies will fit within the taxonomy of mandates and markets. Consider President Biden's 2021 executive order on "Tackling the Climate Crisis at Home and Abroad."[11] It announces goals to put the U.S. on a "path to achieve net-zero emissions, economy-wide, by no later than 2050" along with the so-called 30 x 30 "goal of conserving at least 30 percent of our lands and waters by 2030." Two points are in order. First, these goals appear arbitrary or at least they not set by markets or through an economic benefit-cost optimization routine. They are political and aspirational. Second, the goals are as of yet not accompanied by federal command-and-control methods to achieve them such as bans on fossil fuels by 2050 and bans on land development by 2030.

Tradeoffs between Mandates and Markets

There are many private and social choices to make about the environment such as how much emissions to tolerate, how much water to use for irrigation, and how many wolves to allow within an agricultural landscape. Each choice involves tradeoffs. Less pollution might mean

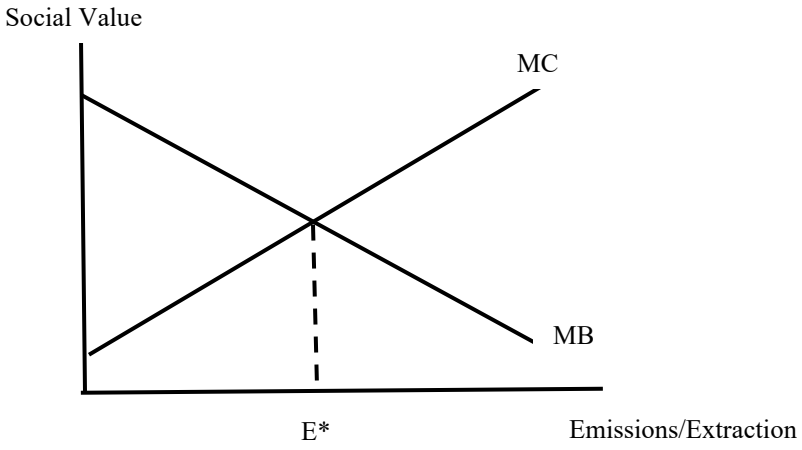
[11] https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/27/executive-order-on-tackling-the-climate-crisis-at-home-and-abroad/

closing coal mines and the reduction in local employment (Hanson 2023, Weber 2020). More water for irrigation means less is left to sustain cold temperatures for fish (Ayres et al. 2021, Brewer et al. 2007). More wolves may mean fewer deer vehicle collisions but less livestock predation (Raynor et al. 2021).

In economics these tradeoffs are theoretically balanced where the social marginal benefits of an additional unit of extraction or pollution equals its social marginal costs (Figure 3). The costs typically embed environmental damages, whether to human health, to human property, or to human enjoyment of the environment. The benefits come from the utility of consuming goods such as iPhones and services such as AirBnB, which are enabled by resource extraction and the income generated therefrom.

In theory the optimal amount of emissions/extraction is at E*. This balances the social value of resource use with the costs, at the margin. It is the economically efficient allocation because any other allocation makes aggregate wealth lower. [Explain this and why E = 0 is rarely optimal (meaning outright bans are rarely optimal) and why $E > E^*$ is also not ideal].

Figure 3: Social Benefits and Costs of Emissions



For economists at least, the question of whether markets or mandates are superior is primarily a question of which of the two gets society closer to E*. This is not a static goal. New technologies, new information, and rising incomes are constantly changing how people weigh environmental quality against other objectives and pursuits. For example, historically forests were valued solely for timber harvest, water was valued for consumption, agriculture, and

mining, and landscapes were valued for commodity production. Today, benefits from standing timber are important, rivers and lakes are valued for their ecosystem services and fish habitat, and private land is valued for its open space scenery. This is all to say that E* has likely fallen over time as incomes have grown, which is one explanation for the downward sloping section of Environmental Kuznets Curves discussed above.

With output mandates politics and administrators choose a target level of E whereas with free markets individuals, firms, and NGOs make emission decisions based on perceived gains from trade which, in aggregate, determine E. In general we expect decentralized actors to have better information about the costs and benefits of resource use and the changing values that society places on them. As the Nobel laureate economist Frederich Hayek (1945) emphasized, market prices convey decentralized information about demands and scarcity that centralized governments could rarely hope to uncover. Why, then, would a society ever favor mandates?

One answer is that free markets fail, although throughout this book we will continue to ask, as the Nobel laureate Ronald Coase always did, "compared to what"? On the economist's blackboard at least, markets fail to deliver theoretically optimal levels of environmental quality either because market participants do not bear the full marginal cost of emissions/extraction, or because they do not accrue the full marginal benefits. There is a divergence between the private and social costs and benefits that market participants are assumed to ignore unless otherwise compelled to consider. The idea is that drivers of gas guzzling vehicles ignore the effects of fossil fuel burning to others – a social cost of carbon that exceeds private costs - or that landowners paid to host wind turbines ignore the disamenity effects on neighbors – a social cost of private energy decisions. In the context of the diagram these examples mean the private marginal costs are below the social marginal cost (to the right) such that $E > E^*$ whether E represents CO_2 or NO_2 emissions or visual or noise blight from turbines.

Herein lies the main economic case for government action. Public interest theory views it as a corrective to improve on efficiency in resource allocation (see e.g., Stiglitz 1989, Pinotti 2012). The reasoning is that government, with its coercive powers, can compel an outcome closer to E* than can unregulated markets. It can do so through quantity restrictions on extraction/emissions or through pricing mechanisms such as taxes and subsidies that cause private market participants to consider the full social effects of their actions. Applied to climate,

the idea is that the social cost of carbon far exceeds the private costs, that government policy is needed to reduce its use, and that such policy will get us closer to E^* than will free markets.

Ronald Coase (1960) offered a nuanced critique to public interest theory and to the necessity and desirability of government action in his Nobel prize winning article on The Problem of Social Cost. The title has a double meaning to us. It acknowledges that the divergence of private and social cost can be an important economic problem but also argues that what economists and common law judges concluded from this divergence was deeply problematic. Coase made three points that are relevant for modern environmental challenges, including climate change.

The first is embedded in what has come to be called the Coase Theorem. It means that, if people affected by each other's actions can negotiate with one another at low costs, then they will have incentives to account for the broader, external effects of their decisions. Coase described this in the context of a rancher whose cattle trample a neighbor's crop but his general point is that negotiation will lead to convergence towards E^* absent governmental policy. Whether the farmer has the right to be free of trampling, or the rancher has the right to trample, the same remedy emerges be it fewer cattle grazed, less corn grown, or a fence erected to separate the two. Economic efficiency will prevail because the market participants in the negotiation lose wealth through any other outcome.[12]

Coase highlighted that negotiation is more difficult in complex conflicts where the transaction costs of bargaining between affected parties are high. Whereas the assignment of legal liability is inconsequential to economic efficiency when transaction costs are low, it is critically influential when they are high. The Coase Theorem makes the first point because it is immaterial whether the rancher or farmer has liability in determining the level of crop damages (the conceptual equivalent of emissions) that the parties will agree upon. To make the second point, he critiqued the logic of judges who had assigned liability to railroad operators whose sparks created fire risk to neighboring woodlands. This liability scheme, coupled with the inability to negotiate due to the high number of small plot owners and hence high transaction costs, causes woodland owners to underinvest in fire prevention by planting close to the railroad.

[12] The Coase Theorem is difficult to empirically assess because the concept of efficiency is difficult to measure (Deruyinga et al. 2021), but Coase's logic on this point won him a Nobel prize and is difficult to refute in any fundamental way. [Add more on the theoretical imprecisions such ignoring of the income effect]

Economic efficiency would suffer if the marginal value of the land is higher when used as a sparks repository instead of space for growing trees.

Coase's point about liability in high-transaction cost scenarios has received less attention than the Coase Theorem but it is arguably more relevant to real-world conflicts about the environment. It raises questions about when polluters or pollutees should be held liable. For example, should high carbon emitting countries be held liable for climate risks in low-lying coastal communities and, if yes, should they pay for adaptation costs or for mitigation? What outcome would the coastal communities choose if they had the right to be free of risk and negotiation costs were low? Does that differ from the outcome occurring if direct negotiation is impossible and instead a global forum deems high-income countries liable? As we describe in a later chapter, adaptation is like building a fence between the rancher and the farmer and the climate risk is like the railroad's sparks.

The third point in Coase (1960) that is relevant to modern environmental problems relates to his appeal for "comparative analysis" of markets versus governmental approaches. Modern environmental problems are complex and dynamic and, as Coase notes, can involve high transaction costs, at least initially. Even in these cases, however, Coase departed from public interest theory arguing that government actions can improve on the status quo or make things worse. He wrote that "governmental administrative regulation" has the best chance of improving economic efficiency "when, as is normally the case with the smoke nuisance, a large number of people is involved and when therefore the costs of handling the problem through the market or the firm may be high." But he went on to say that the "governmental administrative machine is not itself costless. It can, in fact, on occasion be extremely costly" (1960, 17-18). Coase criticized public interest theory because it too often:

> "concentrates attention on particular deficiencies in the system and tends to nourish the belief that any measure which will remove the deficiency is necessarily desirable. It diverts attention from those other changes in the system which are inevitably associated with the corrective measure, changes which may well produce more harm than the original deficiency" (p. xx).

The possibility that government policy can make economic efficiency worse, rather than better, is the topic of a later chapter on unintended consequences. Inquiry into this possibility bred entirely new fields of study as alternatives to public interest theory. The fields of public

choice and interest group theory, for example, view politicians as self-interested actors responding to constituents who have weak incentives and abilities to monitor their actions and bureaucrats as self-interested agents whose actions are also difficult to monitor and observe. In other words, they are like everyone else in economic models! In this setting, well-organized groups form to capture rents from government policies, politicians and bureaucrats sell rents, and this activity explains much of governmental policy (e.g., Tullock 1967, Stigler 1971, Peltzman 1976, Becker 1983, McChesney 1987). Throughout this book we will apply some of this thinking to environmental and climate policies.

Transaction Costs vs. Government Costs

Policy analysis that asks "compared to what" goes beyond determining whether there is a market failure as defined by blackboard economics. If we are to understand when governmental action will really improve on economic efficiency, it is useful to think of market failure as a necessary but not sufficient condition. Comparative institutional analysis requires asking questions about transaction costs and government, or collective action costs.

Transaction Costs

However one defines transaction costs they are the fundamental source of market failure whether it is modeled in textbooks as a result of externalities, public goods, or information asymmetries (see Libecap xx, Anderson xx). To Coase (1960, 15), transaction costs are those that arise to "discover who it is that one wishes to deal with, to inform people that one wishes to deal and on what terms, to conduct negotiations leading up to a bargain, to draw up the contract, to undertake the inspection needed to make sure that the terms of the contract are being observed, and so on". Another useful definition comes from Allen (1991, 4) who says that transaction costs "are the resources used to establish and maintain property rights."

This reasoning implies that externalities, local or global, are problems of incomplete and unclear property rights, formal or informal (Libecap 2014, 1989). In the absence of property rights to say, pollute or to be free from pollution, the transaction costs of market negotiations are high because it is not clear who should compensate whom. Moreover, when rights are unclear or insecure costly effort will be expended to claim and defend.

Public good problems are also transaction cost problems. Unlike private goods, public goods such as clean air are neither rivalrous nor excludable. Nonrival in consumption means that one person's enjoyment does not preclude another person from enjoying the same good. Nonexclusive in consumption means that once a good is supplied to one person, others can consume it free of charge. It is the nonexclusive characteristic that poses difficult problems for market provision because producers of nonexclusive goods cannot extract payment from nonpayers who capture some of the value. Also known as the free-rider problem, nonexclusive consumption means that the voluntary supply of public goods may fall short of what is socially optimal (e.g., E > E* in Fig. 3). The undersupply of public goods is fundamentally a transaction costs problem because, as economist Arthur Pigou put in is his book on "Social Cost", it can be difficult to exact payment" (p. 183–84) from those who would otherwise free ride. But getting free riders to pay also requires discovering "who it is that one wishes to deal with" and conducting "negotiations leading up to a bargain" (Coase 1960, 15). These transaction costs would be enormous in some settings, but it is important to emphasize that efficient supply does not require payment from all consumers.

Information problems are also transaction cost problems because it is difficult for market participants to observe and monitor manufacturing processes and environmental waste or byproducts, for example. [More here]

When and why are transaction costs high? We find it useful to differentiate natural from unnatural costs. Natural transaction costs are like transportation costs in that they exist as matter of nature, but they can be lowered by new technologies and innovations. Rights to highly migratory tuna in the open sea, for example, are difficult to enforce because the resource is remote and inaccessible. And rights to clean water are also difficult to define and enforce when water pollution come from non-point sources such as farms and lawns. Demsetz (1967) hypothesized that the property rights to resources such as these will become better defined as the marginal costs of enforcing property rights fall and the marginal benefits increase.[13]

This implies that transaction costs – and the possibilities for market failure -- are dynamic because property rights to environmental assets are always evolving. Evolving technologies can

[13] Demsetz showed that informal property rights to hunting grounds were better defined and enforced by Native Americans when the value of furs, especially beaver pelts, rose. More generally, he posited that property rights become better defined when the marginal benefits of more complete ownership exceed the marginal costs.

also play an important role in overcoming some of the technical difficulties in creating property rights and excluding free riders. Just as barbed wire revolutionized fencing in the 19th century, satellite monitoring, lasers, and remote sensors make it feasible to record and charge for use. Fisheries can now equip authorized boats with transponders and use real-time satellite monitors to detect trespassers. And tracers can be introduced into air and water emissions to identify and hold accountable owners of the sources. The improved enforcement of property rights lowers the natural transaction costs of market solutions to environmental problems and the likelihood that free market outcomes will move towards E*.

[Add paragraph describing examples of changing technologies for CO_2 detection? What are they?]

In many cases, however, the issue is not that property rights are innately costly to define and enforce but rather that governments refuse to define or enforce them for political reasons. Consider demands for new property rights that come from increased values that society places on pristine environments. For example, historically water was valued for consumption, agriculture, and mining and coal was valued only for energy production. Today, rivers and lakes are valued for their ecosystem services and coal is valued as carbon storage. These changes in demands have led to growing interest among willing buyers to pay farmers to keep water in stream and to coal owners to keep it in the ground.

But governments unnaturally raise the transaction costs of market agreements when they create permitting and red tape constraints on voluntary agreements. Governments also raise transaction costs when they define rights to water and minerals in incomplete and insecure ways. Laws in many countries include “use it or lose it” constraints on water and mineral ownership. Some also require the owner to put the resource to “beneficial use”, as defined by the state. These constraints historically promoted commodity production from land, water, and minerals, but today they are a barrier to environmental conservation (see Anderson and Parker 2013, Leonard et al. 2021).

These examples highlights the role the state can play in lowering or raising transaction costs for market solutions to environmental problems. On the positive, it can serve as a third-party enforcer of property rights by creating laws and courts that clarify liability and, potentially, by providing credible information about risks and pollution. It can also create and enforce tradeable property rights to environmental assets, as with ITQs and emissions credits in tradable

pollution schemes. On the negative side, the state can grant incomplete rights to environmental assets and make them uncertain by threatening to revoke them. [Add example]

Government Costs

[Discuss types and sources of government costs. Citizens have to monitor government agents and outputs against corruption or shirking. Red tape is a costly solution to monitoring problems but at the cost of making low-powered incentives for bureaucrats. Governments lack local information that private actors have. One-size-fits-all standards mean that satisfying the median constituent will not satisfy many. These drawbacks tend to be higher the larger the node of government. Larger nodes of governments, however, are arguably better suited to address large scale environmental problems given their geographic scope of coverage. Discuss tradeoffs in the context of environmental federalism.]

Case Studies

To offer concrete examples of tradeoffs, successes, and failures, *Markets vs. Mandates* will provide case studies of major environmental policies and markets, historical and modern. Here we summarize four key state and federal actions – the first two were more necessary and successful than the latter two by our criteria - and the lessons contained therein.

America's First Environmental Agencies[14]

Most people associate American environmental policy with the federal government, where agencies such as the Department of Interior, Environmental Protection Agency, and Department of Energy control vast resources and regulate private sector activities. But environmental and natural resource agencies in the 50 states have a much long history. Today all states have agencies to regulate agriculture, air quality, fish and wildlife, forests, minerals and energy resources, parks, state lands, water, and air emissions.

Consider the nation's oldest environmental bureau, state wildlife agencies. They were formed during 1878 through 1932 in the wake of severe depletion of many native wildlife stocks and some cases of extinction. Even the once ubiquitous white-tail deer dwindled from estimates

[14] This example is drawn from Lueck and Parker (2023).

of 34 million in 1700 to 500,000 in 1900. By authorizing state agencies to make commercial hunting illegal, regulate and charge for recreational hunting, and hire wardens for enforcement, wildlife agencies consolidated authority over wildlife making it a public asset in contrast to privately owned wildlife elsewhere in the world.

Viewed through the lens of market failure, the creation of wildlife agencies appears a successful intervention. American wildlife were depleted during the 1800s "age of extermination" when commercial markets were widespread. The rapid decline resulted primarily from exploitation by market hunters operating in a *de facto* open access regime. The transaction costs of private regulation at that time were high because private landholdings traversed by wildlife such as deer and elk tended to be small and interspersed with unclaimed parcels. Moreover, "unnatural" transaction costs were high because state legislators and judges refused to enforce landowner claims against trespass, especially on unenclosed and uncultivated land. This raised the costs of private regulation and the demand for state regulatory management. Empirical evidence suggests that wildlife agencies emerged first in states where transaction costs were highest, and that they helped reverse the decline in wildlife stocks. Today there are over 34 million whitetail deer again in the continental U.S.

Some qualifications are in order before declaring the agency solution an unmitigated success. First, there was substantial private action prior to wildlife agencies, especially in areas with large landholdings. Recreational hunting clubs in the 19th century provided management, stocking, rules to limit harvest, and law enforcement. Agencies may have crowded out long-term private conservation by monopolizing regulatory authority. Second, the need for agencies to regulate wildlife on private land was partly due to the refusal of states and judges to create and enforce trespass laws, and because governments claimed land within and adjacent to private land making private control difficult. Transaction costs have fallen over time as rural landholdings grew, ownership voids were filled, and states began enforcing trespassing laws, weakening the rationale for modern state regulatory involvement except areas where public lands still dominate. Third, the ban on commercial markets of wildlife products now frustrates the management of overabundant species such as whitetail deer. Public trusteeship means there is no owner liable for wildlife damages such as those from deer-vehicle collisions.

The Clean Air Act

Federal involvement in environmental policy began in earnest with landmark legislation in the 1960s and 1970s. The National Environmental Protection Act (NEPA) created the EPA. Other major legislation includes the Clean Air Act (1970), the Clean Water Act (1972), and the Endangered Species Act (1967 and 1973).

Although states already had air pollution agencies, the CAA created federal rules to govern emissions across the country. Air pollution at that time was a significant risk to human health in many American cities due especially to power plant and vehicle emissions. The CAA initially employed command-and-control mandates such as requiring plants to install scrubbers in smokestacks, but 1990 amendments implemented emission standards rather than specific technology requirements. Importantly, the amendments created a cap-and-trade emissions systems and it expanded allowance of emission offsets (Stavins 2011, Shapiro and Walker 2023).

Viewed through the lens of market failure, the CAA appears a successful intervention. The transaction costs of market-driven improvements were high because air pollution affects large numbers, and it can be difficult to assign legal liability. Moreover, wind currents move dirty air across jurisdictions making it difficult for states to control emission sources and weakening the incentives of upwind states to voluntarily regulate. According to one assessment, the CAA delivered benefits over ten times the regulatory cost between 1992 to 2017 with most of the benefits resulting from decreased risk of premature death (Keiser and Shapiro 2019, Table 1).[15]

This is not to suggest the CAA has been free of mischief and unwanted consequences. Due to lobbying pressure from industry, the Act initially forced new plants to adopt a particular abatement technology (smokestack scrubbers) even though alternatives to getting cleaner air existed such as burning low-sulphur coal. The requirement caused existing plants to delay upgrading and to forgo cleaner coal, thereby decreasing air quality in some locales (Ackerman and Hassler 1981). Today, CAA requirements can limit prescribed woodland burns that are effective at reducing wildlife risk, thereby increasing the likelihood of future air quality impairments due to wildfire smoke (cite). And enforcement of the CAA relies on measurements

[15] One study, conducted by the University of Chicago's Energy Policy Institute, identifies air pollution as the world's top threat to public health, responsible for reducing average life expectancy by 2.3 years worldwide. The net effects of air pollution on mortality depends of course on people's ability to adapt. Higher incomes allow easier adaptation as we discuss later in this book.

from air pollution monitors in locations chosen for political reasons (Grainger et al. 2019). This means that air quality measurement - and hence the location of regulatory restrictions - is politically maneuvered.

The Clean Water Act

- All states had water pollution control agencies before 1972, but the federal CWA offered grants and subsidized loans to cities for wastewater treatment improvements. It also required that facilities discharging pollution from a fixed point receive permits from the EPA to meet water quality standards. Water quality, at least measured by the proportion of fishable surface waters, was low by today's standards but already improving prior to 1972. Trends in water quality continued to improve after 1972.
- The CWA does not allow water quality trading between permitted facilities.
- Though there is evidence that the CWA did improve surface water quality, most analyses suggest those benefits have been less than the costs (see Keiser and Shapiro 2019). There is limited evidence to suggest the CWA has delivered major human health benefits, and its compliance costs are higher than if cap-and-trading were allowed.
- Viewed through the lens of *ex post* analysis, the CWA does not appear successful. Private (albeit subsidized) approaches to improving surface water quality are emerging through conservation easements and watershed organizations who pay landowners to change management practices. Some state agencies, such as Wisconsin's, allow water trading [More…talk to Dan Phanuef]

The Endangered Species Act

- The ESA, passed in 1967 and substantively amended in 1973, is supposed to "provide a means whereby the ecosystems upon which endangered and threatened species depend may be conserved." It authorizes the U.S. Fish and Wildlife Service to list species to prevent their killing and habitat modification, even on private land. Some regard the ESA as America's most powerful environmental legislation given this regulatory reach.
- The market failure rationale is that endangered species are a public good and, as such, free markets will provide very little habitat or protection, especially for species that are not charismatic, not marketable, and not valued as game.

- The ESA requires the USFWS to determine whether or not a species warrants a listing based solely on the science, without consideration of economic impacts. Evidence, however, indicates that listing decisions were quite political, at least during the first two decades of implementation. Charismatic animals were listed sooner, as were animals whose listing would impose higher costs on private landowners (Ando 1999, xxx).
- ESA regulations reduce private land values (Auffhammer et al. 2023), and hence create perverse incentives for landowners to preemptively destroy species habitat to avoid regulation, as in the case of early tree harvesting to avoid hosting Red Cockaded Woodpeckers (Lueck and Michaels 2003). Market reforms to mitigate these effects, such as safe harbor agreements that compensate affected landowners and habitat offsets can in theory reduce costs and increase benefits.
- We are not aware of benefit-cost studies of the ESA (which are prohibited as a regulatory consideration) but the record of success looks bleak. The ESA has arguably prevented extinction but it has proven less successful in recovering endangered species. Only one percent of listed species have gone extinct, while less than two percent have been delisted. And there are 287 species that remain listed that the USFWS projected should have recovered by now (Wright and Regan 2023).

Conclusion

On Earth Day in 2005 *The Economist* magazine ran an article with the title "Market forces could prove to be the environment's best friend – if only greens could learn to love them." This book will explore when and how markets can work, compared to mandates, for greens who measure performance mainly by improvements in environmental conditions. It will approach this analysis with the skepticism of economist Thomas Sowell (2010, 45), who reminds us that "policies need to be analyzed in terms of the incentives they create, rather than the hopes that inspire them."

But we will also continue to remind readers that environmental quality, while important, is not the only good that makes people happy. People also want things reliable cars for transport, international travel, and safe homes with secure energy. Most people also value freedom – personal, economic, and political - as an intrinsic good that, to some, is more important than environmental quality. When there are tradeoffs, as there always are, we will rely throughout this

book on people's actual decisions in markets to learn what is most important to them, and about what markets or mandates can offer.

An equally important reminder throughout is that people value different aspects of "the environment". Improving one aspect can sometimes come at the expense of another. Though discussions of climate change dominate mainstream media, Americans have since 1989 remained more concerned about water quality, soil contamination, and air quality (Keiser and Shapiro 2019, Fig. 1). Though we have not seen global survey data, we doubt that citizens in poorer countries put more weight on climate than environmental conditions more immediately impactful to day-to-day life, such as access to safe drinking water. This is important because, for example, mining in Indonesia for minerals for electric vehicle battery manufacturing in China degrades water quality and soils in Indonesia.

Markets vs. Mandates seeks an audience of scientists, social scientists, and policy analysts who are well trained at identifying market failures but less acquainted with government failure. Perhaps ironically, the emphasis on market failure continues to emanate from economics itself. This includes the University of Chicago's modern economics department, which had built a legacy of intellectual support for free market economics during much of the 20th century with economists such as Gary Becker, Ronald Coase, Frank Knight, Milton Friedman, and George Stigler (see Banzaf 2023). As a *Wall Street Journal* visiting journalist at the University of Chicago recently put it, "my impression was that today's professors emphasize the drawbacks of markets as much as the benefits." The problem, as explained to the journalist, is that a free market is the best way to run an economy only when it is "fully competitive, has no unpriced side effects, or 'externalities', such as carbon emissions, and where contracts cover every eventuality". The conclusion, of course, is that "these conditions are never met" (Mackintosh 2023).

Markets vs. Mandates pushes back, as Ronald Coase would, by relentlessly asking "compared to what". Mandates and government interventions can do better than free markets, but this requires more than contracts failing to "cover every eventuality" which is true of every real-world contract. The real world offers imperfect solutions to environmental problems, and this book will assess which approaches deliver people more of what they want – including environmental quality and individual freedom – while giving up less to get it.

And as this book will incessantly highlight, the economist's toolkit is critical here, especially in offering benefit-cost assessments of policies relative to realistic alternatives. Equally critical is an appreciation for human adaptation – to environmental conditions including climate – and the role that markets and mandates can play in facilitating or stunting it over the long run.

References

Ackerman, Bruce and William T. Hassler. 1981. *Clean Coal, Dirty Air*. Yale University Press.

Aldy, Joseph E. 2005. An Environmental Kuznets Curve Analysis of US State Level Carbon Dioxide Emissions. *Journal of Environment and Development* 14: 48-72.

Anderson, Terry L. 2021. *Adapt and be Adept: Market Responses to Climate Change*. Hoover Institution Press.

Anderson, Terry L. and Dominic P. Parker. 2013. Transaction Costs and Environmental Markets: The Role of Entrepreneurs. *Review of Environmental Economics and Policy* 7(2): 259-275.

Anderson, Terry and Robert McCormick. 2007. More Inconvenient Truths. *Hoover Digest* No. 2.

Ando, Amy W. 1999. Waiting to be Protected Under the Endangered Species Act: The Political Economy of Regulatory Delay. *Journal of Law and Economics* 42(1): 29-60.

Andreoni, James and Arik Levinson. The Simple Analytics of the Environmental Kuznets Curve. *Journal of Public Economics* 80: 269-286.

Auffhammer, Maximilian, Eyal Frank, David W. McLaughlin, Beia Spiller and David Sunding. 2023. The Cost of Species Protection: The Land Market Impacts of the Endangered Species Act. Working Paper.

Ayres, Andrew B, Kyle C. Meng and Andrew J. Plantinga. 2021. Do Environmental Markets Improve on Open Access? Evidence from California Groundwater Rights. *Journal of Political Economy* 129(10): 2817-2860.

Banzhaf, H. Spencer. 2023. *Pricing the Priceless: A History of Environmental Economics*. Cambridge University Press.

Brewer, Jedidiah, Robert Glennon, Alan Ker and Gary D. Libecap. 2007. Water Markets in the West: Prices, Trading, and Contractual Forms. NBER Working Paper 13002.

Carleton, Tamma et al. 2022. Valuing the Global Mortality Consequences of Climate Change Accounting for Adaptation Costs and Benefits. *Quarterly Journal of Economics* 137(4): 2037-2105.

Clay, Karen and Werner Troesken. 2010. Did Frederick Brodie Discover the World's First Environmental Kuznets Curve? Coal, Smoke and the Rise and Fall of the London Fog. NBER Working Paper 15669.

Costello, Christopher, Steven D. Gains and John Lynham. 2008. Can Catch Shares Prevent Fisheries Collapse? *Science* 321(19): 1678-1681.

Fell, Harrison and Daniel T. Kaffine. 2018. The Fall of Coal: Joint Impacts of Fuel Prices and Renewables on Generation and Emissions. *American Economic Journal: Economic Policy* 10(2): 90-116.

Grainger, Corbett and Dominic P. Parker. 2013. The Political Economy of Fishery Reform. *Annual Review of Resource Economics* 5: 369-386.

Grainger, Corbett, Andrew Schreiber and Wonjun Chang. 2019. Do Regulators Strategically Avoid Hotspots when siting Monitors? Evidence from Remote Sensing of Air Pollution. Working Paper.

Greenstone, Michael, Christian Leuz, and Patricia Breuer. 2023. Mandatory Disclosures would Reveal Corporate Climate Damages. *Science* 381(6660): 837-840.

Hanson, Gordon H. 2023. Local Labor Market Impacts of the Energy Transition: Prospects and Policies. NBER Working Paper.

Harsted, Bard. 2012. Buy Coal! A Case for Supply-Side Environmental Policy. *Journal of Political Economy* 120(1): 77-115.

Hayek, Frederich A. 1945. The Use of Knowledge in a Society. *American Economic Review* 35(4): 519 – 530.

Johnsen, Reid, Jacob LaRiviere, and Hendrik Wolff. 2019. Fracking, Coal, and Air Quality. *Journal of the Association of Environmental and Resource Economists* 6(5): 1001-37.

Kahn, Matthew E. 2021. *Adapting to Climate Change: Markets and the Management of an Uncertain Future*. Yale University Press.

Libecap, Gary D. 1994. *Contracting for Property Rights*. Cambridge University Press

Libecap, Gary D. 2008. Open-Access Losses and Delay in the Assignment of Property Rights. *Arizona Law Review* 50: 379 – 408.

Libecap, Gary D. 2014. Addressing Global Environmental Externalities: Transaction Costs Considerations. *Journal of Economic Literature* 52(2): 424-479.

Lueck, Dean and Jeffrey A. Michael. 2003. Preemptive Habitat Destruction under the Endangered Species Act. *Journal of Law and Economics* 46(1): 27-60.

Mackintosh, James. 2023. What I Learned about 'Woke' Capital and Milton Friedman at the University of Chicago. *Wall Street Journal*, June 9.

Millimet, Daniel L., John A. List and Thanasis Stengos. 2003. The Environmental Kuznets Curve: Real Progress or Misspecified Models? *Review of Economics and Statistics* 85(4): 1038-47.

Myers, Erica, Steven Puller, and Jeremy West. Forthcoming. Mandatory Energy Efficiency Disclosure in Housing Markets. *American Economic Journal: Economic Policy*.

Parker, Dominic P. 2010. Does Evidence on the Environmental Kuznets Curve Support the Case for Market-Based Environmentalism? Paper prepared for PERC Conference on "Tough Questions for Free-Market Environmentalism", July 2010, Bozeman Montana.

Parker, Dominic P., Jeremy Foltz, and David Elsea. 2016. Unintended Consequences of Sanctions for Human Rights: Conflict Minerals and Infant Mortality. *Journal of Law & Economics* 59(4): 731-774.

Phaneuf, Daniel and Till Requate. 2017. *A Course in Environmental Economics: Theory, Policy, and Practice*. Cambridge University Press.

Raynor, Jennifer, Corbett Grainger, and Dominic P. Parker. 2021. Wolves Make Roadways Safer, Generating Large Economic Returns to Predator Conservation. *Proceedings of the National Academy of Sciences* 118(22): 1-10.

Selden, Thomas M. and Daqing Song. 1994. Environmental Quality and Development: Is There a Kuznets Curve for Air Pollution Emissions? *Journal of Environmental Economics and Management* 27: 147-62.

Shapiro, Joseph S. and Reed Walker. 2023. Is Air Pollution Regulation too Stringent? Evidence from U.S. Offset Markets. Working Paper.

Sowell, Thomas. 2010. *Basic Economics, 4th Edition: A Common Sense Guide to the Economy*. Basic Books. New York.

Stavins, Robert N. 2011. The Problem of the Commons: Still Unsettled after 100 Years. *American Economic Review* 101: 81-108.

Stern, David I. 1998. Progress on the Environmental Kuznets Curve? *Environment and Development Economics 3*(2): 175-198.

Weber, Jeremy G. 2020. How Should we Think about Environmental Policy and Jobs? An Analogy with Trade Policy and an Illustration from U.S. Coal Mining. *Review of Environmental Economics and Policy* 14(1): 44 – 66.

Wright, Katherine and Shawn Regan. 2023. Missing the Mark: How the Endangered Species Act Falls Short of its Own Recovery Goals. PERC Report, Bozeman Montana.

Yandle, Bruce, Madhusudan Bhattarai and Maya Vijayaraghavan. 2004. Environmental Kuznets Curves: A Review of Findings, Methods, and Policy Implications. PERC Research Study 02-1, April.

The Nature of the Firm

ECONOMIC theory has suffered in the past from a failure to state clearly its assumptions. Economists in building up a theory have often omitted to examine the foundations on which it was erected. This examination is, however, essential not only to prevent the misunderstanding and needless controversy which arise from a lack of knowledge of the assumptions on which a theory is based, but also because of the extreme importance for economics of good judgment in choosing between rival sets of assumptions. For instance, it is suggested that the use of the word "firm" in economics may be different from the use of the term by the "plain man."[1] Since there is apparently a trend in economic theory towards starting analysis with the individual firm and not with the industry,[2] it is all the more necessary not only that a clear definition of the word "firm" should be given but that its difference from a firm in the "real world," if it exists, should be made clear. Mrs. Robinson has said that "the two questions to be asked of a set of assumptions in economics are: Are they tractable? and: Do they correspond with the real world?"[3] Though, as Mrs. Robinson points out, "more often one set will be manageable and the other realistic," yet there may well be branches of theory where assumptions may be both manageable and realistic. It is hoped to show in the following paper that a definition of a firm may be obtained which is not only realistic in that it corresponds to what is meant by a firm in the real world, but is tractable by two of the most powerful instruments of economic analysis developed by Marshall, the idea of the margin and that of substitution, together giving the idea of substitution at

[1] Joan Robinson, *Economics is a Serious Subject*, p. 12.
[2] See N. Kaldor, "The Equilibrium of the Firm," *Economic Journal*, March, 1934.
[3] Op. cit., p. 6.

the margin.[1] Our definition must, of course, "relate to formal relations which are capable of being *conceived* exactly."[2]

I

It is convenient if, in searching for a definition of a firm, we first consider the economic system as it is normally treated by the economist. Let us consider the description of the economic system given by Sir Arthur Salter.[3] "The normal economic system works itself. For its current operation it is under no central control, it needs no central survey. Over the whole range of human activity and human need, supply is adjusted to demand, and production to consumption, by a process that is automatic, elastic and responsive." An economist thinks of the economic system as being co-ordinated by the price mechanism and society becomes not an organisation but an organism.[4] The economic system "works itself." This does not mean that there is no planning by individuals. These exercise foresight and choose between alternatives. This is necessarily so if there is to be order in the system. But this theory assumes that the direction of resources is dependent directly on the price mechanism. Indeed, it is often considered to be an objection to economic planning that it merely tries to do what is already done by the price mechanism.[5] Sir Arthur Salter's description, however, gives a very incomplete picture of our economic system. Within a firm, the description does not fit at all. For instance, in economic theory we find that the allocation of factors of production between different uses is determined by the price mechanism. The price of factor A becomes higher in X than in Y. As a result, A moves from Y to X until the difference between the prices in X and Y, except in so far as it compensates for other differential advantages, disappears. Yet in the real world, we find that there are many areas where this does not apply. If a workman moves from department Y to department X, he does not go because of a change in relative prices, but because he is ordered to do so. Those who

[1] J. M. Keynes, *Essays in Biography*, pp. 223–4.
[2] L. Robbins, *Nature and Significance of Economic Science*, p. 63.
[3] This description is quoted with approval by D. H. Robertson, *Control of Industry*, p. 85, and by Professor Arnold Plant, "Trends in Business Administration," ECONOMICA, February, 1932. It appears in *Allied Shipping Control*, pp. 16-17.
[4] See F. A. Hayek, "The Trend of Economic Thinking," ECONOMICA, May, 1933.
[5] See F. A. Hayek, op. cit.

object to economic planning on the grounds that the problem is solved by price movements can be answered by pointing out that there is planning within our economic system which is quite different from the individual planning mentioned above and which is akin to what is normally called economic planning. The example given above is typical of a large sphere in our modern economic system. Of course, this fact has not been ignored by economists. Marshall introduces organisation as a fourth factor of production; J. B. Clark gives the co-ordinating function to the entrepreneur; Professor Knight introduces managers who co-ordinate. As D. H. Robertson points out, we find "islands of conscious power in this ocean of unconscious co-operation like lumps of butter coagulating in a pail of buttermilk."[1] But in view of the fact that it is usually argued that co-ordination will be done by the price mechanism, why is such organisation necessary? Why are there these "islands of conscious power"? Outside the firm, price movements direct production, which is co-ordinated through a series of exchange transactions on the market. Within a firm, these market transactions are eliminated and in place of the complicated market structure with exchange transactions is substituted the entrepreneur-co-ordinator, who directs production.[2] It is clear that these are alternative methods of co-ordinating production. Yet, having regard to the fact that if production is regulated by price movements, production could be carried on without any organisation at all, well might we ask, why is there any organisation?

Of course, the degree to which the price mechanism is superseded varies greatly. In a department store, the allocation of the different sections to the various locations in the building may be done by the controlling authority or it may be the result of competitive price bidding for space. In the Lancashire cotton industry, a weaver can rent power and shop-room and can obtain looms and yarn on credit.[3] This co-ordination of the various factors of production is, however, normally carried out without the intervention of the price mechanism. As is evident, the amount of "vertical" integration, involving as it does

[1] Op. cit., p. 85.

[2] In the rest of this paper I shall use the term entrepreneur to refer to the person or persons who, in a competitive system, take the place of the price mechanism in the direction of resources.

[3] *Survey of Textile Industries*, p. 26.

the supersession of the price mechanism, varies greatly from industry to industry and from firm to firm.

It can, I think, be assumed that the distinguishing mark of the firm is the supersession of the price mechanism. It is, of course, as Professor Robbins points out, "related to an outside network of relative prices and costs,"[1] but it is important to discover the exact nature of this relationship. This distinction between the allocation of resources in a firm and the allocation in the economic system has been very vividly described by Mr. Maurice Dobb when discussing Adam Smith's conception of the capitalist: "It began to be seen that there was something more important than the relations inside each factory or unit captained by an undertaker; there were the relations of the undertaker with the rest of the economic world outside his immediate sphere . . . . the undertaker busies himself with the division of labour inside each firm and he plans and organises consciously," but "he is related to the much larger economic specialisation, of which he himself is merely one specialised unit. Here, he plays his part as a single cell in a larger organism, mainly unconscious of the wider rôle he fills."[2]

In view of the fact that while economists treat the price mechanism as a co-ordinating instrument, they also admit the co-ordinating function of the "entrepreneur," it is surely important to enquire why co-ordination is the work of the price mechanism in one case and of the entrepreneur in another. The purpose of this paper is to bridge what appears to be a gap in economic theory between the assumption (made for some purposes) that resources are allocated by means of the price mechanism and the assumption (made for other purposes) that this allocation is dependent on the entrepreneur-co-ordinator. We have to explain the basis on which, in practice, this choice between alternatives is effected.[3]

[1] Op. cit., p. 71.

[2] *Capitalist Enterprise and Social Progress*, p. 20. Cf., also, Henderson, *Supply and Demand*, pp. 3–5.

[3] It is easy to see when the State takes over the direction of an industry that, in planning it, it is doing something which was previously done by the price mechanism. What is usually not realised is that any business man in organising the relations between his departments is also doing something which could be organised through the price mechanism. There is therefore point in Mr. Durbin's answer to those who emphasise the problems involved in economic planning that the same problems have to be solved by business men in the competitive system. (See "Economic Calculus in a Planned Economy," *Economic Journal*, December, 1936.) The important difference between these two cases is that economic planning is imposed on industry while firms arise voluntarily because they represent a more efficient method of organising production. In a competitive system, there is an "optimum" amount of planning!

II

Our task is to attempt to discover why a firm emerges at all in a specialised exchange economy. The price mechanism (considered purely from the side of the direction of resources) might be superseded if the relationship which replaced it was desired for its own sake. This would be the case, for example, if some people preferred to work under the direction of some other person. Such individuals would accept less in order to work under someone, and firms would arise naturally from this. But it would appear that this cannot be a very important reason, for it would rather seem that the opposite tendency is operating if one judges from the stress normally laid on the advantage of "being one's own master."[1] Of course, if the desire was not to be controlled but to control, to exercise power over others, then people might be willing to give up something in order to direct others; that is, they would be willing to pay others more than they could get under the price mechanism in order to be able to direct them. But this implies that those who direct pay in order to be able to do this and are not paid to direct, which is clearly not true in the majority of cases.[2] Firms might also exist if purchasers preferred commodities which are produced by firms to those not so produced; but even in spheres where one would expect such preferences (if they exist) to be of negligible importance, firms are to be found in the real world.[3] Therefore there must be other elements involved.

The main reason why it is profitable to establish a firm would seem to be that there is a cost of using the price mechanism. The most obvious cost of "organising" production through the price mechanism is that of discovering what the relevant prices are.[4] This cost may be reduced but it will not be eliminated by the emergence of specialists who will sell this information. The costs of negotiating and

[1] Cf. Harry Dawes, "Labour Mobility in the Steel Industry," *Economic Journal*, March, 1934, who instances "the trek to retail shopkeeping and insurance work by the better paid of skilled men due to the desire (often the main aim in life of a worker) to be independent" (p. 86).

[2] None the less, this is not altogether fanciful. Some small shopkeepers are said to earn less than their assistants.

[3] G. F. Shove, "The Imperfection of the Market: a Further Note," *Economic Journal*, March, 1933, p. 116, note 1, points out that such preferences may exist, although the example he gives is almost the reverse of the instance given in the text.

[4] According to N. Kaldor, "A Classificatory Note of the Determinateness of Equilibrium," *Review of Economic Studies*, February, 1934, it is one of the assumptions of static theory that "All the relevant prices are known to all individuals." But this is clearly not true of the real world.

concluding a separate contract for each exchange transaction which takes place on a market must also be taken into account.[1] Again, in certain markets, e.g., produce exchanges, a technique is devised for minimising these contract costs; but they are not eliminated. It is true that contracts are not eliminated when there is a firm but they are greatly reduced. A factor of production (or the owner thereof) does not have to make a series of contracts with the factors with whom he is co-operating within the firm, as would be necessary, of course, if this co-operation were as a direct result of the working of the price mechanism. For this series of contracts is substituted one. At this stage, it is important to note the character of the contract into which a factor enters that is employed within a firm. The contract is one whereby the factor, for a certain remuneration (which may be fixed or fluctuating), agrees to obey the directions of an entrepreneur *within certain limits*.[2] The essence of the contract is that it should only state the limits to the powers of the entrepreneur. Within these limits, he can therefore direct the other factors of production.

There are, however, other disadvantages—or costs—of using the price mechanism. It may be desired to make a long-term contract for the supply of some article or service. This may be due to the fact that if one contract is made for a longer period, instead of several shorter ones, then certain costs of making each contract will be avoided. Or, owing to the risk attitude of the people concerned, they may prefer to make a long rather than a short-term contract. Now, owing to the difficulty of forecasting, the longer the period of the contract is for the supply of the commodity or service, the less possible, and indeed, the less desirable it is for the person purchasing to specify what the other contracting party is expected to do. It may well be a matter of indifference to the person supplying the service or commodity which of several courses of action is taken, but not to the purchaser of that service or commodity. But the purchaser will not know which of these several courses he will want the supplier to take. Therefore,

[1] This influence was noted by Professor Usher when discussing the development of capitalism. He says: "The successive buying and selling of partly finished products were sheer waste of energy." (*Introduction to the Industrial History of England*, p. 13). But he does not develop the idea nor consider why it is that buying and selling operations still exist.

[2] It would be possible for no limits to the powers of the entrepreneur to be fixed. This would be voluntary slavery. According to Professor Batt, *The Law of Master and Servant*, p. 18, such a contract would be void and unenforceable.

the service which is being provided is expressed in general terms, the exact details being left until a later date. All that is stated in the contract is the limits to what the persons supplying the commodity or service is expected to do. The details of what the supplier is expected to do is not stated in the contract but is decided later by the purchaser. When the direction of resources (within the limits of the contract) becomes dependent on the buyer in this way, that relationship which I term a " firm " may be obtained.[1] A firm is likely therefore to emerge in those cases where a very short term contract would be unsatisfactory. It is obviously of more importance in the case of services—labour—than it is in the case of the buying of commodities. In the case of commodities, the main items can be stated in advance and the details which will be decided later will be of minor significance.

We may sum up this section of the argument by saying that the operation of a market costs something and by forming an organisation and allowing some authority (an " entrepreneur ") to direct the resources, certain marketing costs are saved. The entrepreneur has to carry out his function at less cost, taking into account the fact that he may get factors of production at a lower price than the market transactions which he supersedes, because it is always possible to revert to the open market if he fails to do this.

The question of uncertainty is one which is often considered to be very relevant to the study of the equilibrium of the firm. It seems improbable that a firm would emerge without the existence of uncertainty. But those, for instance, Professor Knight, who make the *mode of payment* the distinguishing mark of the firm—fixed incomes being guaranteed to some of those engaged in production by a person who takes the residual, and fluctuating, income—would appear to be introducing a point which is irrelevant to the problem we are considering. One entrepreneur may sell his services to another for a certain sum of money, while the payment to his employees may be mainly or wholly a share in profits.[2] The significant question would

[1] Of course, it is not possible to draw a hard and fast line which determines whether there is a firm or not. There may be more or less direction. It is similar to the legal question of whether there is the relationship of master and servant or principal and agent. See the discussion of this problem below.

[2] The views of Professor Knight are examined below in more detail.

appear to be why the allocation of resources is not done directly by the price mechanism.

Another factor that should be noted is that exchange transactions on a market and the same transactions organised within a firm are often treated differently by Governments or other bodies with regulatory powers. If we consider the operation of a sales tax, it is clear that it is a tax on market transactions and not on the same transactions organised within the firm. Now since these are alternative methods of "organisation"—by the price mechanism or by the entrepreneur—such a regulation would bring into existence firms which otherwise would have no *raison d'être*. It would furnish a reason for the emergence of a firm in a specialised exchange economy. Of course, to the extent that firms already exist, such a measure as a sales tax would merely tend to make them larger than they would otherwise be. Similarly, quota schemes, and methods of price control which imply that there is rationing, and which do not apply to firms producing such products for themselves, by allowing advantages to those who organise within the firm and not through the market, necessarily encourage the growth of firms. But it is difficult to believe that it is measures such as have been mentioned in this paragraph which have brought firms into existence. Such measures would, however, tend to have this result if they did not exist for other reasons.

These, then, are the reasons why organisations such as firms exist in a specialised exchange economy in which it is generally assumed that the distribution of resources is "organised" by the price mechanism. A firm, therefore, consists of the system of relationships which comes into existence when the direction of resources is dependent on an entrepreneur.

The approach which has just been sketched would appear to offer an advantage in that it is possible to give a scientific meaning to what is meant by saying that a firm gets larger or smaller. A firm becomes larger as additional transactions (which could be exchange transactions co-ordinated through the price mechanism) are organised by the entrepreneur and becomes smaller as he abandons the organisation of such transactions. The question which arises is whether it is possible to study the forces which determine the size of the firm. Why does the entrepreneur not organise one

less transaction or one more? It is interesting to note that Professor Knight considers that:

> "the relation between efficiency and size is one of the most serious problems of theory, being, in contrast with the relation for a plant, largely a matter of personality and historical accident rather than of intelligible general principles. But the question is peculiarly vital because the possibility of monopoly gain offers a powerful incentive to *continuous and unlimited* expansion of the firm, which force must be offset by some equally powerful one making for decreased efficiency (in the production of money income) with growth in size, if even boundary competition is to exist."[1]

Professor Knight would appear to consider that it is impossible to treat scientifically the determinants of the size of the firm. On the basis of the concept of the firm developed above, this task will now be attempted.

It was suggested that the introduction of the firm was due primarily to the existence of marketing costs. A pertinent question to ask would appear to be (quite apart from the monopoly considerations raised by Professor Knight), why, if by organising one can eliminate certain costs and in fact reduce the cost of production, are there any market transactions at all?[2] Why is not all production carried on by one big firm? There would appear to be certain possible explanations.

First, as a firm gets larger, there may be decreasing returns to the entrepreneur function, that is, the costs of organising additional transactions within the firm may rise.[3] Naturally, a point must be reached where the costs of organising an extra transaction within the firm are equal to the costs involved in carrying out the transaction in the open market, or, to the costs of organising by another entrepreneur. Secondly, it may be that as the transactions which are organised increase, the entrepreneur fails to place the factors of production in the uses where their value

[1] *Risk, Uncertainty and Profit*, Preface to the Re-issue, London School of Economics Series of Reprints, No. 16, 1933.

[2] There are certain marketing costs which could only be eliminated by the abolition of "consumers' choice" and these are the costs of retailing. It is conceivable that these costs might be so high that people would be willing to accept rations because the extra product obtained was worth the loss of their choice.

[3] This argument assumes that exchange transactions on a market can be considered as homogeneous; which is clearly untrue in fact. This complication is taken into account below.

is greatest, that is, fails to make the best use of the factors of production. Again, a point must be reached where the loss through the waste of resources is equal to the marketing costs of the exchange transaction in the open market or to the loss if the transaction was organised by another entrepreneur. Finally, the supply price of one or more of the factors of production may rise, because the "other advantages" of a small firm are greater than those of a large firm.[1] Of course, the actual point where the expansion of the firm ceases might be determined by a combination of the factors mentioned above. The first two reasons given most probably correspond to the economists' phrase of "diminishing returns to management."[2]

The point has been made in the previous paragraph that a firm will tend to expand until the costs of organising an extra transaction within the firm become equal to the costs of carrying out the same transaction by means of an exchange on the open market or the costs of organising in another firm. But if the firm stops its expansion at a point below the costs of marketing in the open market and at a point equal to the costs of organising in another firm, in most cases (excluding the case of "combination"[3]), this will imply that there is a market transaction between these two producers, each of whom could organise it at less than the actual marketing costs. How is the paradox to be resolved? If we consider an example the reason for this will become clear. Suppose *A* is buying a product from *B* and that both *A* and *B* could organise this marketing transaction at less than its present cost. *B*, we can assume, is not organising one process or stage of production, but several. If *A* therefore wishes to avoid a market transaction, he will have to take over all the processes of production controlled by *B*. Unless *A* takes over all the processes of

[1] For a discussion of the variation of the supply price of factors of production to firms of varying size, see E. A. G. Robinson, *The Structure of Competitive Industry*. It is sometimes said that the supply price of organising ability increases as the size of the firm increases because men prefer to be the heads of small independent businesses rather than the heads of departments in a large business. See Jones, *The Trust Problem*, p. 531, and Macgregor, *Industrial Combination*, p. 63. This is a common argument of those who advocate Rationalisation. It is said that larger units would be more efficient, but owing to the individualistic spirit of the smaller entrepreneurs, they prefer to remain independent, apparently in spite of the higher income which their increased efficiency under Rationalisation makes possible.

[2] This discussion is, of course, brief and incomplete. For a more thorough discussion of this particular problem, see N. Kaldor, "The Equilibrium of the Firm," *Economic Journal*, March, 1934, and E. A. G. Robinson, "The Problem of Management and the Size of the Firm," *Economic Journal*, June, 1934.

[3] A definition of this term is given below.

production, a market transaction will still remain, although it is a different product that is bought. But we have previously assumed that as each producer expands he becomes less efficient; the additional costs of organising extra transactions increase. It is probable that *A*'s cost of organising the transactions previously organised by *B* will be greater than *B*'s cost of doing the same thing. *A* therefore will take over the whole of *B*'s organisation only if his cost of organising *B*'s work is not greater than *B*'s cost by an amount equal to the costs of carrying out an exchange transaction on the open market. But once it becomes economical to have a market transaction, it also pays to divide production in such a way that the cost of organising an extra transaction in each firm is the same.

Up to now it has been assumed that the exchange transactions which take place through the price mechanism are homogeneous. In fact, nothing could be more diverse than the actual transactions which take place in our modern world. This would seem to imply that the costs of carrying out exchange transactions through the price mechanism will vary considerably as will also the costs of organising these transactions within the firm. It seems therefore possible that quite apart from the question of diminishing returns the costs of organising certain transactions within the firm may be greater than the costs of carrying out the exchange transactions in the open market. This would necessarily imply that there were exchange transactions carried out through the price mechanism, but would it mean that there would have to be more than one firm? Clearly not, for all those areas in the economic system where the direction of resources was not dependent directly on the price mechanism could be organised within one firm. The factors which were discussed earlier would seem to be the important ones, though it is difficult to say whether "diminishing returns to management" or the rising supply price of factors is likely to be the more important.

Other things being equal, therefore, a firm will tend to be larger:

(*a*) the less the costs of organising and the slower these costs rise with an increase in the transactions organised.

(*b*) the less likely the entrepreneur is to make mistakes and the smaller the increase in mistakes with an increase in the transactions organised.

The Quiet Resignation: Why Do So Many Female Lawyers Abandon Their Careers?

Thirty percent of female lawyers leave their careers. The same is true for female doctors. Over time, an increasing number of married professionals have recreated traditional gender roles, and society has lost a tremendous amount of training and well-honed talent as a result. Neither workplace discrimination nor family obligations can fully and satisfactorily explain the trend. Both of those theories assume that women take a more dependent and vulnerable position in the household because of constraints, but in one important respect, men are more constrained than women, and they are better off for it: to maintain social status, men have to work. Women do not.

This Article advances a theory and corroborating evidence that the cultural acceptance of female underemployment is a privilege in the abstract, but a curse in practice. Even under the best conditions, the early stages of professional careers involve mistakes, mismatches, and disappointments. An opportunity to escape the stress of the public sphere by focusing on the family may have great appeal in the short run even though the long-run consequences are severe. Asymmetric cultural acceptance creates an easy off-ramp for females, to nearly everybody's detriment.

INTRODUCTION

Women have made impressive gains in the labor market since World War II.[1] Before 1947, very few married women worked fulltime.[2] Today, the majority

1. Female fulltime employment grew from 33.9% in 1950 to 57% in 2013. Cynthia Hess et al., *The Status of Women in the States: 2015*, INST. FOR WOMEN'S POL'Y RES. 53 (2015), http://statusofwomendata.org/wp-content/uploads/2015/02/Status-of-Women-in-the-States-2015-Full-National-Report.pdf [https://perma.cc/G6W4-LU99]. And this trend was *most strong* for the most educated women and the women who married high-earning husbands. Ellen R. McGrattan & Richard Rogerson, *Changes in the Distribution of Family Hours Worked Since 1950*, *in* FRONTIERS OF FAMILY ECON. 115 (Peter Rupert ed., 2008); Francine D. Blau & Lawrence M. Kahn, *The Gender Wage Gap: Extent, Trends, and Explanations* 12 (Nat'l Bureau of Econ. Research, Working Paper No. 21913, 2016) [hereinafter Blau & Kahn, *The Gender Wage Gap*], http://www.nber.org/papers/w21913 [https://perma.cc/R5N9-DK2W].

2. Blau & Kahn, *The Gender Wage Gap*, *supra* note 1.

do.[3] American families seemed to be on an irreversible cultural path toward gender parity in which men and women both invested in their own careers. Claudia Goldin called it "the quiet revolution."[4]

The revolution was led by educated women who married high-earning men.[5] Elites, it seems, were the most interested and the most capable of striving for a two-earner marital partnership.[6] This is not surprising; well-educated women are, after all, in the best position to claim high salaries, secure high-status jobs, and to marry successful partners. These power couples can easily outsource a lot of the cooking, housekeeping, and childcare work that could cause tension and zero-sum bargaining in other households. But things have started to change.

Today, we are in the midst of a "quiet resignation." The cultural pressure that drove many women to proudly invest in their careers has subsided.[7] The most talented and privileged women are returning to the home in staggering numbers after short stints in the white-collar workforce. When the most privileged women leave the public sphere, it causes a cascade of other social problems.[8] The wage gap has stopped closing.[9] Desegregation within college majors (where men tend to focus on the more lucrative STEM fields while women tend to major in the liberal

3. Female fulltime employment grew from 33.9% in 1950 to 57% in 2013. Hess et al., *supra* note 1.

4. Claudia Goldin, *The Quiet Revolution that Transformed Women's Employment, Education, and Family*, (Nat'l Bureau of Econ. Research, Working Paper No. 11953, 2006) [hereinafter Goldin, *Quiet Revolution*], http://www.nber.org/papers/w11953 [https://perma.cc/NZ65-L9NH], *reprinted in* 96 AM. ECON. ASS'N PAPERS & PROC. 1 (2006); *see also* Claudia Goldin, *Career and Family: College Women Look to the Past*, *in* GENDER AND FAMILY ISSUES IN THE WORKPLACE 20 (Francine D. Blau & Ronald G. Ehrenberg eds., 1997); CLAUDIA GOLDIN, UNDERSTANDING THE GENDER GAP: AN ECONOMIC HISTORY OF AMERICAN WOMEN (Robert W. Fogel & Clayne L. Pope eds., 1990). Many readers will be familiar with Claudia Goldin's empirical work on discrimination including the now famous orchestra study demonstrating that blind orchestra auditions—which forced musicians to audition behind a screen—helped women improve their outcomes. Claudia Goldin & Cecilia Rouse, *Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians*, 90 AM. ECON. REV. 715 (2000).

5. McGrattan & Rogerson, *supra* note 1, at 115; Blau & Kahn, *The Gender Wage Gap*, *supra* note 1, at 12.

6. Blau & Kahn, *The Gender Wage Gap*, *supra* note 1.

7. Claudia Goldin herself considered the evidence available ten years ago to see whether labor participation had stalled or reached some sort of natural equilibrium. She found some evidence that the revolution was continuing, though, since women in their thirties were more likely to have young kids than women in earlier generations (whose kids would have been older), and therefore what looked like a plateau may actually be a shift in the timing of family-related disruption. Goldin, *Quiet Revolution*, *supra* note 4, at 24–26, 33. ("[W]as Ely correct in asserting that 'Revolutions do not go backward?' There is little evidence that this one has, at least for now.").

8. *See infra* Part IV.

9. Casey B. Mulligan & Yona Rubinstein, *Selection, Investment, and Women's Relative Wages Over Time*, 123 Q.J. ECON. 1061, 1062–63 (2008). Wage inequality within gender caused talented women, with greater earnings potential, to enter the workforce in greater numbers. Thus, the shrinking of the wage gap between 1970s and 1990s may be illusory, as the ranks of working women were joined by more talented and ambitious women during this time.

arts) has not improved in decades.[10] And boardrooms and law firm partnerships, suffering from an alleged "pipeline problem," are still dominated by men.[11]

This Article grounds the discussion of professional career abandonment in hard facts, and then explains the phenomenon. We present a mix of new and existing evidence showing that women in the best position to advance gender equality—who are trained in law, medicine, and business—are leaving the fulltime workforce at very high rates. About 30% of female lawyers are unemployed or under-employed[12] during prime career-building years—ages 35–40.[13] These are the worst years to drop out of the labor market, just when the fruits of all the training and hard work are nearly ready for harvest. Despite the time and effort spent completing exhausting educational programs, these women wind up in households that conform to all the hallmarks of traditional gender roles.

Two theories dominate the literature on gender gaps in the legal field.[14] First, discrimination may explain some of the gender gap in labor participation. Since women experience and expect unfair treatment at hiring, promotion, salary negotiations,[15] and case assignment, they may value their careers less than similarly qualified men. This theory puts the blame on employers for failing to correct for explicit and implicit biases that have long plagued the workplace. The second explanation is self-selection.[16] For cultural, biological, and path-dependent

10. Blau & Kahn, *The Gender Wage Gap*, *supra* note 1, at 15–16.

11. Lauren Stiller Rikleen, *Solving the Law Firm Gender Gap Problem*, HARV. BUS. REV. (Aug. 20, 2013), https://hbr.org/2013/08/solving-the-law-firm-gender-ga [https://perma.cc/XXL6-WBAN] (rejecting the theory that the gap in law firm partnership is explained by women voluntarily leaving the pipeline).

12. We use the term "underemployed" to mean working fewer than twenty hours per week. We justify our focus on this metric *infra* in Part I.

13. We focus on lawyers because the data is more suitable for our analyses. The performance of lawyers can be quantified to some degree (hours billed, revenue generated) and, therefore, it allows us to better identify how much of the gender gaps in earnings or career trajectory can be attributed to performance gaps, and how much cannot. *See* Ghazala Azmat & Rosa Ferrer, *Gender Gaps in Performance: Evidence from Young Lawyers*, 125 J. POL. ECON. 1306 (2017).

14. Other theories that get much less attention in the context of lawyers include differences in performance, differences in the taste for competition, *see* Uri Gneezy et al., *Performance in Competitive Environments: Gender Differences*, 118 Q.J. ECON. 1049 (2003); Muriel Niederle, *Gender* (Nat'l Bureau of Econ. Research, Working Paper No. 20788, 2014), http://www.nber.org/papers/w20788 [https://perma.cc/58GX-4T8X], *reprinted in* 2 HANDBOOK OF EXPERIMENTAL ECONOMICS 483 (John Kagel & Alvin E. Roth eds., 2015), and long-term cultural commitments to gender equality, *see* Esther Duflo, *Women Empowerment and Economic Development*, 50 J. ECON. LITERATURE 1051, 1051 (2012).

15. There is, of course, a gender dimension to negotiations. Women do not negotiate as much as men, partly because women are by nature or socialized to be more egalitarian.
Catherine Eckel et al., *Gender and Negotiation in the Small: Are Women (Perceived to Be) More Cooperative than Men?*, 24 NEGOT. J. 429 (2008).

16. Also known as the "Opting Out" phenomenon. Nicole M. Fortin, *Gender Role Attitudes and Women's Labor Market Participation: Opting-Out, AIDS, and the Persistent Appeal of Housewifery*, 117 ANNALS ECON. & STAT. 379 (2015).

reasons, women simply prefer to leave their careers and to use their talents raising children and performing volunteer services.[17]

Both discrimination and self-selection theories have empirical support, and we think it's obvious that both are operating to some degree. The relative effects of each are hard to determine, and even the best data sources can be cherry-picked to promote one explanation or the other. The law and policy literature has compulsively focused on untangling the proportional share of responsibility between discrimination and self-selection so that an appropriate level of blame and legal intervention can be directed at employers. For the most part, scholars who study gender gaps treat voluntary decisions to leave the labor market as a matter of each person's prerogative.[18] Self-selection, in other words, is treated as a non-problem so long as women leave their careers for some reason —any reason—other than discriminatory treatment at work. But there is good reason to believe that attempts to settle the debate between discrimination and self-selection shroud a more direct, more powerful influence on female labor participation.

This Article highlights the power of asymmetrical cultural pressures that shape the trajectories of men and women. Single professionals must invest in work for practical reasons—unless they are independently wealthy, they have no other option to achieve financial security and social status. But married men and women make decisions under very different social pressure. Men are expected to participate in the labor market for most of their lives, and the pressure comes from outside and inside the home. Women, by contrast, are given much more leniency to divest from their careers. Progressives and conservatives alike give women the authority to opt out of the workforce, and encourage their spouses to respect and support their choices.[19]

This cultural acceptance for female underemployment is meant to be a privilege—something that expands the option set for women. But in practice, it is a trap. Culture interacts with other layers of pressure and influence to create a one-way valve that takes women out of the workforce. Consider how this cultural asymmetry combines with marriage, work experience, and family.

---

17. We refer to this as self-selection rather than family or child-rearing responsibility because, except in unusual cases, there is no necessity for a parent to stay home to care for children when the mother is professionally trained.

18. *See* Azmat & Ferrer, *supra* note 13.

19. For example, Aishwarya Rai, the former Miss World, one of the leading actors of Bollywood, and a goodwill ambassador of the United Nations, stated: "The housewives are the biggest CEOs in the country and they should be given utmost respect and appreciation in the country. I doff my hat off with full respect and appreciation to all the housewives of the country and the world[.]" Pranita Chaubey, *Aishwarya Rai Bachchan Says 'Housewives Are the Biggest CEOs in India'*, NDTV.COM (Aug. 04, 2018), https://www.ndtv.com/entertainment/aishwarya-rai-bachchan-says-housewives-are-the-biggest-ceos-in-india-1895040?pfrom=home-lateststories [https://perma.cc/8FGF-4UCQ]. The mores are different in matrilineal societies, where female underemployment would presumably be less tolerated. Uri Gneezy et al., *Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society*, 77 ECONOMETRICA 1637 (2009).

Marriage creates the opportunity for career divestment. Professionally trained women are typically married to men who also have high earnings potential.[20] Assortative mating creates an opportunity for either one of the partners to stop working for a prolonged period of time while still enjoying a high quality of life.[21]

Meanwhile, the labor market is a slog. In due time it will be kind to most professionals—both lucrative and personally rewarding. But professional careers often begin with some challenging years, requiring newcomers to manage stress and disappointment as they figure out their niche.[22] That's the best-case scenario. In addition to the strain of mismatched jobs, inconsistent performance, abusive bosses, and random bad luck that all young professionals have to navigate, young female professionals have additional stress and discouragement from discriminatory treatment. And of course, they are also biologically destined to manage the mother of all career setbacks—pregnancy and early infant care—which affects the productivity of even the most devoted careerist. Thus, while many young professionals face some early career turmoil that causes them to question their path and devotion to working, women face disproportionately more of these moments.

Finally, the influence of family. When young children are in a household, the domestic sphere is much more demanding on professional parents' time—usually more demanding than either of the parents had anticipated.[23] Even with the aid of daycare, nannies, housekeepers, and other paid help, the household still requires active management of one of the parents. All this typically hits at a time when professionals are in the middle of their career marathons, and when success in the labor market rewards the people who can double-down their investment in work. How will professional couples respond to the domestic demands, to the vacuum at home? One parent cannot divest from their careers without losing their status in society and, often, the respect of their spouse. The other can.

---

20. Throughout this article, we focus on opposite-sex relationships. In the future, we may have enough data to be able to reexamine these trends with more nuance by including studies of same-sex couplings and transgender professionals. We suspect that the results we present *infra* in Part I are *underestimates* of the trends we discuss because an unknown portion of female graduates are in same-sex relationships where home-market divisions of labor may be less likely to occur.

21. This Article focuses on heterosexual couples partly out of necessity (many of the datasets we rely on do not have sufficient numbers of same-sex-professional couples to make comparative analyses). Moreover, since the majority of female JD holders are heterosexual, the dynamics of their households will have greater explanatory power for the statistical trends we are trying to explain. Still, we hope in future work we can work with more comprehensive data and provide more detailed analyses that do not rely on these heteronormative assumptions.

22. Job satisfaction increases with time. *See, e.g.*, Arne L. Kalleberg & Karyn A. Loscocco, *Aging, Values, and Rewards: Explaining Age Differences in Job Satisfaction*, 48 AM. SOC. REV. 78 (1983).

23. Iyana Kuziemko et al., *The Mommy Effect: Do Women Anticipate the Employment Effects of Motherhood?* 1 (Nat'l Bureau of Econ. Research, Working Paper No. 24740, 2018).

While it's true that workplace gender discrimination and pregnancy have differential effects on men and women, obsessive focus on both of those factors obscures the profound impact of culture. Highly trained professionals of both genders will frequently have an unpleasant period at work, and will also recognize that their young children require a good deal of time and attention. Women will tend to approach marital bargains and household decision-making with a very different set of preferences than their husbands because there is no cultural counterpressure to rebut the push out of the workforce and the pull into the home.

While societal expectations will push men to meet the challenges of early career bobbles with resilience and problem-solving, married women will have too many options. They will have an all-too-appealing opportunity to leave the labor market. This option affords women short-term relief from a bad work situation and medium-term satisfaction as they nurture young children, but it deprives them from fully realizing their long-term career potential.[24] It also causes externalities to their children and coworkers, and it perpetuates the (accurate) stereotype that women are less likely to remain in their jobs.[25] While economists describe gender gaps in terms of constrained options for women,[26] in one important respect, men are more constrained than women, and they are better off for it. To maintain social status, men have to work. Women do not.

The career trajectory of one of the authors, Jane Bambauer, illustrates our theory. One year out of law school, Jane was unhappy with her work at a boutique litigation firm. The hours were long and often spent concentrating on problems that were dull (albeit important to the case). Just when the dissatisfaction became particularly acute, Jane also experienced sexual harassment. If Jane were married to another professional, she would have quit and turned her attention to family. But for being single, she would have become part of the 30% of female JD-holders who leave full-time employment. Lacking that option, Jane regrouped for a few weeks and redirected her energy toward a job that better matched her interests and values.

Thus, it is with the sincerity that comes from personal experience that we assure readers our theory is not intended to place fault on any individual woman or family that winds up recreating traditional gender roles. Work can be unfair and inhospitable. But the amount of talent lost to the culturally-assisted exit ramp is unacceptable. Future generations of lawyers should be making work-related decisions and spousal bargains under a different set of norms in which parents understand that the public and the family lose out when a mother with marketable skills leaves the workforce.

---

24. *See infra* Part IV.C.

25. The externalities and "internalities" are considered at length *infra* Part IV.

26. *See, e.g.*, Kai Liu, *Explaining the Gender Wage Gap: Estimates from a Dynamic Model of Job Changes and Hours Changes*, 7 QUANTITATIVE ECON. 411 (2016) (analyzing constraints that limit women's hourly wages).

Our cultural acceptance theory diverges from most of the existing literature in three ways. First, it is unabashedly pro-work. We define gender gaps in labor participation (particularly in the high-paying, high status professions) as a social problem that deserves a remedy. This is not to say that every person should try to maximize earnings. To the contrary, the most satisfying careers are often the ones that make conscious trade-offs between income and interests, and between work and leisure/family time. But for reasons we elaborate, leaving the labor force is usually a bad deal for the professionals who do so, for the firms that trained them, and for the children who are the intended beneficiaries of the sacrifice.

Second, and relatedly, the cultural acceptance theory merges the discrimination and self-selection theories by reducing the importance of whether exit was voluntary. A commitment to gender parity demands some scrutiny when women routinely volunteer to take a more dependent, vulnerable position in a household after attaining advanced degrees and on-the-job professional training.

Third, the perverse cultural acceptance theory calls for significantly different policy interventions than the discrimination and self-selection theories. Employment law that minimizes discrimination and reduces the disruption from childbearing is valuable, but employers are not wholly responsible for attitudes about labor supply and childcare. We propose some changes to tax and disability law that can help promote an expectation for both parents in a two-parent household to maintain their careers. But lasting change will require the help of cultural leaders and role models as much or more than public law. We offer some recommendations for law schools and firms to help generate discussion, awareness, and more informed decision-making among junior attorneys.

The Article proceeds in five parts. Part I presents the data on female underemployment across multiple professions, with particular focus on lawyers. Part II describes and critiques the standard explanations for female career detachment—discrimination and self-selection. Part III describes our theory of perverse cultural acceptance and provides the existing evidence. Part IV makes the normative case that underemployment by highly trained married women is problematic not merely for aesthetic or symbolic reasons but because it has repercussions for women, for their families, and for society. Part V sketches some pragmatic policy proposals (both legal and non-legal) that can help balance out the asymmetric cultural pressure placed on male and female professionals.

I. FEMALE UNDEREMPLOYMENT

Over the last fifty years, women have been closing the gap between their earnings and their husbands'.[27] This work was spearheaded by well-educated women who were drawn to the challenges and rewards of a professional work

27. Casey B. Mulligan & Yona Rubinstein, *The Closing of the Gender Gap as a Roy Model Illusion* 3 fig.1 (Nat'l Bureau of Econ. Research, Working Paper No. 10892, 2004).

life.[28] But today, married women with professional degrees are *less* invested in work than their less educated counterparts. These professional women overcame whatever obstacles were in the way and made it to the launching pad for financial independence and professional status only to abandon the project at the proof-of-concept stage.

This Part presents evidence of the quiet trend among elite women to leave their careers. It also discusses why we focus on under-employed lawyers as key indicators. Later parts will go on to explain the trend.

A. Description

A large number of highly skilled females are underemployed—that is, they work fewer than twenty hours or are not working at all. As Figures 1–3 show, the rate of underemployment is consistent across disciplines (law, medicine, and business) and has been increasing since 1993. We recommend particular focus on the 36–40 age bands. These are prime years in which career investments start to mature, but many professional women in this age range are not fully participating in the labor market.[29] For lawyers, 28.5%–30.1% of women in this age band are underemployed compared to less than 4% of men.[30]

Figure 1.
Proportion of Female JDs Who Are Underemployed
(unemployed or working fewer than twenty hours)

	ALL	< 30	31–35	36–40	41–45	46–50	51–60
1993 (874)	25.3%	18.7%	26.4%	29.7%	21.3%	18.2%	21.1%
2003 (997)	24.3%	16.6%	21.3%	30.1%	27.2%	21.6%	19.1%
2010 (1,079)	29.1%	24.4%	17.8%	28.5%	36.9%	31.2%	26.9%

Source: National Survey of College Graduates 1993, 2003, and 2010

Figure 2.
Proportion of Female MDs Who Are Underemployed
(unemployed or working fewer than twenty hours)

	ALL	< 30	31–35	36–40	41–45	46–50	51–60
1993 (1,448)	24.2%	12.4%	19.2%	27.3%	21.9%	16.1%	18.8%
2003 (1,263)	26.8%	17.0%	17.3%	29.9%	25.2%	25.9%	22.0%
2010 (824)	31.4%	4.9%	24.0%	24.4%	38.1%	24.2%	34.3%

Source: National Survey of College Graduates

28. *Id.*

29. For the purposes of this paper, we use "professional" to mean graduates of a professional-postbaccalaureate-degree program holding a JD, MD, or MBA.

30. Across the three survey years, the percentage of underemployed men with JDs ranged from 3.4%–3.7%. Men with MDs and MBAs in the ages 36–40 were somewhat more likely to be underemployed than lawyers. The range of underemployment for male MDs and MBAs was 2.8%–12% and 5.4%–10.1%, respectively.

Figure 3.
Proportion of Female MBAs Who Are Underemployed
(unemployed or working fewer than twenty hours)

	ALL	< 30	31–35	36–40	41–45	46–50	51–60
1993 (1,470)	24.1%	13.5%	22.8%	28.5%	19.0%	20.9%	25.8%
2003 (1,622)	26.0%	15.1%	22.0%	28.7%	26.9%	23.9%	23.7%
2010 (1,497)	28.2%	17.0%	17.3%	25.2%	32.9%	27.8%	29.5%

Source: National Survey of College Graduates

Other, more specialized data sets corroborate these findings. In fact, the rate of underemployment may be higher, ironically, for women who attend the most prestigious graduate programs. Figure 4 shows the results of an alumni survey from Harvard Law School. Less than 60% of female alumnae who were fifteen years out of law school were working in full-time positions.[31]

Figure 4.
Proportion of Harvard Law School Alumni Who Are Underemployed

Graduating Year	Men	Women	% Respondents Who Are Female
1975 (35 years out)	22.4%	38.1%	16%
1985 (25 years out)	9.4%	33.2%	43%
1995 (15 years out)	5.1%	40.3%	46%
2000 (10 years out)	5.7%	25%	53%

Source: 2010 Survey of Harvard Law School graduates reporting either working part-time or not working at all.

Data from another top-ranked law school, the University of Michigan, shows similar patterns. Ken Dau-Schmidt's research on Michigan law alumni has found that more than a third of female alums take time off of full-time work for childcare (beyond maternity leave) and that the *average* length of the leave is nearly five years.[32] This average length of leave has increased over time.[33] (By

31. DAVID B. WILKINS ET AL., THE WOMEN AND MEN OF HARVARD LAW SCHOOL: PRELIMINARY RESULTS FROM THE HLS CAREER STUDY 24 tbl.5.1 (2015). Results from a class of graduates from the University of Virginia School of Law are very similar. *See* John Monahan & Jeffrey Swanson, *Lawyers at Mid-Career: A 20-Year Longitudinal Study of Job and Life Satisfaction*, 6 J. EMPIRICAL LEGAL STUD. 451 (2009).

32. Kenneth Dau-Schmidt et al., *Men and Women of the Bar: The Impact of Gender on Legal Careers*, 16 MICH. J. GENDER & L. 49, 72 (2009) [hereinafter Dau-Schmidt et al., *Men and Women of*

The Law is Not the Contract:[1] A Reflection on *ProCd v. Zeidenberg*

Working Paper – Basis for Manne Madness Lecture "Judge Easterbrook was Right in *ProCD*, and Online Contracting is (Better Than) Fine"

In his controversial decision in *ProCD, Inc. v. Zeidenberg*, U.S. Circuit Judge Frank Easterbrook enforced a software license restriction contained inside a box of compact discs.[2] The software at issue was a nationwide telephone directory that, though costly to compile, maintain, and update, was not copyrightable.[3] To protect its investment, ProCD, the software's maker, created two types of licenses: a relatively expensive one for business users who would be willing to pay high dollar for the contact information and a relatively less expensive one for individual users who would not find the commercial license worth the commercial price, but who might pay something for the usefulness of a digital, searchable nationwide phonebook.[4] The license for the consumer version of the software restricted reproduction of the information, in an effort to prevent a sharp arbitrageur from purchasing the consumer version of the software, copying the discs or contents, and reselling the information in some way.[5] Matthew Zeidenberg, though, ignored the license restriction and did exactly that, creating a website, posting the entirety of ProCD's database, charging his own customers and selling ads to profit from web traffic to his site where users could use it without buying it from ProCD.[6]

[1] See generally Alfred Korzybski, A Non-Aristotelian System and Its Necessity for Rigour in Mathematics and Physics, *Science and Sanity*, 747, 750 (1933) ("the map is *not* the territory").
[2] 86 F.3d 1447, 1449 (7th Cir. 1996).
[3] Id.
[4] Id. at 1449–50.
[5] Id. at 1449–50.
[6] Id. at 1450; Emma Peaslee, Planet Money looks at how hidden contracts took over the internet, Planet Money, Morning Edition, National Public Radio (July 21, 2023).

Because the printed terms designed to bind Zeidenberg contractually from doing this could not be accessed prior to opening the box, he could not have known the full terms of the license restriction until after buying the box of discs and opening it.[7] The district court held the terms did not bind him because they were inside the box and, thus, unavailable prior to payment and transfer of the physical vessel.[8] The district court concluded Zeidenberg could not be bound by terms of which he was unaware and over which he could not bargain at the time money and discs changed hands.[9]

By reversing this, Judge Easterbrook offended an understanding of contract law that apparently measures legal validity by the extent of the actual bargaining between the parties.[10] That view of contract law is tantalizing in its intuitiveness and simplicity. Why should courts enforce bargains with no actual bargaining, agreements with no conscious agreement, and contracts with no meaningful contracting? Moreover, refusing to enforce such deals would protect parties with lower bargaining power from predation, avoid sticky questions about whether the terms were substantively unfair, deter more powerful parties from sneaking terms into contracts through interminable boilerplate, encourage bargaining instead of disputes, and, at least in the case of intellectual property, protect a valuable federal role in regulating economic assets and organizing property rights.[11] All this, while also providing simple and predictable rules for judges – what's not to like?

---

[7] Id.
[8] *ProCD, Inc. v. Zeidenberg*, 908 F.Supp 640, 655 (W.D. Wis. 1996).
[9] Id.
[10] *See* Roger Schecter, The Unfairness of Click-on Licenses, 46 Wayne L. R. 1735, 1739 (2000) (describing commentary concluding that such licenses should be unenforceable for lacking "a bargained-for exchange").
[11] *See generally* Margaret Jane Radin, Regime Change: Superseding the Law of the State with the "Law" of the Firm, 1 U. Ottawa L. & Tech. J. 173, 185 (proposing distinguishing *ProCD* in "circumstances where the traditional ideological picture of contract has become clearly inapposite").

As Judge Easterbrook recognized, there are a few missing puzzle pieces. First, what about the real costs of bargaining individually over terms?[12] If the law demands costly, time-consuming, and legible "bargaining" to suit this vision of the law, what sorts of deals will be lost? Some may never occur at all, while others will occur outside the legal system at greater cost to the parties. Second, what about the costs to purchasers who would readily agree to the terms if they knew them?[13] Matthew Zeidenberg was an arbitrageur at best, a sharp dealer at worst, profiting from the labors and investment of ProCD and exploiting a gray area in copyright law. Why should other purchasers bear the cost of his actions either by being unable to obtain ProCD's consumer product at a reasonable price or at all?[14] Finally, standardization benefits the seller and the buyer, so long as the buyer is always ultimately getting what she paid for, even if it isn't exactly what she would want in a world where transacting itself was costless.

Judge Easterbrook valiantly supplied a countering social policy explanation to challenge the contrary social policy views about contract law and actual bargaining.[15] This Article argues that, while he has the better of that argument, the real problem is the chronically unclear purpose and role contract law plays in transacting. Lawyers, courts, and legal scholars often focus on the law and the broad public policy it must convey. But, in the process, they may be tempted to place the law, courts, and optimal social policy at the center of the analysis – that is, they implicitly view contract law as a regulatory device.[16] This Article argues that this view is

[12] *ProCD*, 86 F.3d at 1451.
[13] Id. at 1450–51.
[14] Id. at 1449–51.
[15] Id. at 1451.
[16] *Compare* William Shugart, On the Virginia school of antitrust: Competition policy, law & economics and public choice, 191 Pub. Choice 1, 16 (2022) (describing the application of antitrust law as an instance of economic regulation more broadly).

incomplete. Instead, the role and function of contract *law* should be to facilitate and enforce private *transactions* in identifiable, oft-repeated circumstances where private facilitation and enforcement *actually* fail.[17] *ProCD* was not one of those situations, and neither are the vast run of online contracting scenarios. Judge Easterbrook's approach in *ProCD* undermined nothing except for an incomplete vision of the role and purpose of contract law in the economics of transactions.

Consequently, *ProCD* is not the "case that kind of broke contract law."[18] Instead, *ProCD* held implicitly, and correctly, that contract law does not exist for its own sake, it exists as just one institution for providing economic actors with assurance for their mutually-beneficial transactions. The district court, citing judicial decisions and law review articles skeptical of these so-called "shrinkwrap" terms, decided that the terms would not be a part of the contract because Zeidenberg:

> "did not receive the opportunity to inspect or consider those terms. Mere reference to the terms at the time of initial contract formation does not present buyers an adequate opportunity to decide whether they are acceptable. They must be able to read and consider the terms in their entirety."[19]

What the district court ignored is what might have happened had the full terms been available to Zeidenberg pre-purchase. Perhaps, with hindsight, one could argue that Zeidenberg simply would not have purchased the discs, never posted the information, and never collected his profits. But, given that Zeidenberg did post the information, collect the

---

[17] *See* Barak Richman, Firms, Courts, and Reputation Mechanisms: Toward a Positive Theory of Private Ordering, 104 Colum. L. Rev. 2229, 2230 (2004) ("Businesspeople need *transactional* assurance") (emphasis original).
[18] Emma Peaslee, Planet Money looks at how hidden contracts took over the internet, Planet Money, Morning Edition, National Public Radio (July 21, 2023).
[19] Id. at 654. Notably, the court acknowledged that a reference was made to the terms' existence on the outside of the box. Id.

profits, and purchase two more updates to the software *after* learning of the terms, this seems doubtful. More critically, ProCD made abundantly clear by its licensing terms that it would not have sold the software to him at all had he revealed his true purpose for purchasing the discs. Even on the grounds of social policy reasoning of the kind that Judge Easterbrook deployed, this would mean that the law facilitated Zeidenberg appropriating contractual surplus belonging to ProCD. Constructing the law to countenance outcome would only flow logically from reverence for contract law as existing for its own sake, rather than for enforcement of clear private bargains where private enforcement mechanisms fail.

This Article revisits *ProCD* to reflect on the proper role and function of contract law in transactional settings. It concludes that Judge Easterbrook's detractors may be right about one thing – namely, that his ruling "harmed" a particular vision or version of contract *law*. Critically, though, it did not harm the supplying and consuming of goods and services online, it did not damage the ability of transactors to conduct mutually-beneficial transactions on the terms that make those transactions beneficial, and it did not interfere with productive activity in the name of control.

for much of the non-rational behavior that individuals may exhibit, even if it is systematic rather than random. As a result, the traditional law-and-economics model is far from being supplanted.

3. Assumptions about Firm Behavior

A corollary of the self-interest assumption for individual behavior is the assumption that business firms attempt to maximize economic profit. Economic profit is defined as the total revenues received from selling a product or service *minus* the total costs of producing the product or service, including the opportunity costs

Matsushita Elec. Indus. Co. v. Zenith Radio Corp.

Supreme Court of the United States
475 U.S. 574 (1986)

POWELL, J.

* * *

Petitioners, defendants below, are 21 corporations that manufacture or sell "consumer electronic products" (CEPs) — for the most part, television sets. Petitioners include both Japanese manufacturers of CEPs and American firms, controlled by Japanese parents, that sell the Japanese-manufactured products. Respondents, plaintiffs below, are Zenith Radio Corporation (Zenith) and National Union Electric Corporation (NUE). Zenith is an American firm that manufactures and sells television sets. NUE is the corporate successor to Emerson Radio Company, an American firm that manufactured and sold television sets until 1970, when it withdrew from the market after sustaining substantial losses. Zenith and NUE began this lawsuit in 1974, claiming that petitioners had illegally conspired to drive American firms from the American CEP market. According to respondents, the gist of this conspiracy was a " 'scheme to raise, fix and maintain artificially *high* prices for television receivers sold by [petitioners] in Japan and, at the same time, to fix and maintain *low* prices for television receivers exported to and sold in the United States.' " These "low prices" were allegedly at levels that produced substantial losses for petitioners. . . .

* * *

. . . The thrust of respondents' argument is that petitioners used their monopoly profits from the Japanese market to fund a concerted campaign to price predatorily and thereby drive respondents and other American manufacturers of CEPs out of business. Once successful, according to respondents, petitioners would cartelize the American CEP market, restricting output and raising prices above the level that fair competition would produce. The resulting monopoly profits, respondents contend, would more than compensate petitioners for the losses they incurred through years of pricing below market level.

* * *

. . . According to petitioners, the alleged conspiracy is one that is economically irrational and practically infeasible. Consequently, petitioners contend, they had no motive to engage in the alleged predatory pricing conspiracy; indeed, they had a strong motive not to conspire in the manner respondents allege. Petitioners argue that, in light of the absence of any apparent motive and the ambiguous nature of the evidence of conspiracy, no trier of fact reasonably could find that the conspiracy with which petitioners are charged actually existed. This argument requires us to consider the nature of the alleged conspiracy and the practical obstacles to its implementation.

IV
A

A predatory pricing conspiracy is by nature speculative. Any agreement to price below the competitive level requires the conspirators to forgo profits that free competition would offer them. The forgone profits may be considered an investment in the future. For the investment to be rational, the conspirators must have a reasonable expectation of recovering, in the form of later monopoly profits, more than the losses suffered. As then-Professor Bork, discussing predatory pricing by a single firm, explained:

> "Any realistic theory of predation recognizes that the predator as well as his victims will incur losses during the fighting, but such a theory supposes it may be a rational calculation for the predator to view the losses as an investment in future monopoly profits (where rivals are to be killed) or in future undisturbed profits (where rivals are to be disciplined). The future flow of profits, appropriately discounted, must then exceed the present size of the losses." R. Bork, The Antitrust Paradox 145 (1978).

As this explanation shows, the success of such schemes is inherently uncertain: the short-run loss is definite, but the long-run gain depends on successfully neutralizing the competition. Moreover, it is not enough simply to achieve monopoly power, as monopoly pricing may breed quick entry by new competitors eager to share in the excess profits. The success of any predatory scheme depends on *maintaining* monopoly power for long enough both to recoup the predator's losses and to harvest some additional gain. Absent some assurance that the hoped-for monopoly will materialize, *and* that it can be sustained for a significant period of time, "[the] predator must make a substantial investment with no assurance that it will pay off." For this reason, there is a consensus among commentators that predatory pricing schemes are rarely tried, and even more rarely successful.

These observations apply even to predatory pricing by a *single firm* seeking monopoly power. In this case, respondents allege that a large number of firms have conspired over a period of many years to charge below-market prices in order to stifle competition. Such a conspiracy is incalculably more difficult to execute than an analogous plan undertaken by a single predator. The conspirators must allocate the losses to be sustained during the conspiracy's operation, and must also allocate any gains to be realized from its success. Precisely because success is speculative and depends on a willingness to endure losses for an indefinite period, each conspirator has a strong incentive to cheat, letting its partners suffer the losses necessary to destroy the competition while sharing in any gains if the conspiracy succeeds. The necessary allocation is therefore difficult to accomplish. Yet if conspirators cheat to any substantial extent, the conspiracy must fail, because its success depends on depressing the market price for *all* buyers of CEPs. If there are too few goods at the artificially low price to satisfy demand, the would-be victims of the conspiracy can continue to sell at the "real" market price, and the conspirators suffer losses to little purpose.

Finally, if predatory pricing conspiracies are generally unlikely to occur, they are especially so where, as here, the prospects of attaining monopoly power seem slight. In order to recoup their losses, petitioners must obtain enough market power to set higher than competitive prices, and then must sustain those prices long enough to earn in excess profits what they earlier gave up in below-cost prices. Two decades after their conspiracy is alleged to have commenced, petitioners appear to be far from achieving this goal: the two largest shares of the retail market in television sets are held by RCA and respondent Zenith, not by any of petitioners. Moreover, those shares, which together approximate 40% of sales, did not decline appreciably during the 1970's. Petitioners' collective share

rose rapidly during this period, from one-fifth or less of the relevant markets to close to 50%. Neither the District Court nor the Court of Appeals found, however, that petitioners' share presently allows them to charge monopoly prices; to the contrary, respondents contend that the conspiracy is ongoing—that petitioners are still artificially *depressing* the market price in order to drive Zenith out of the market. The data in the record strongly suggest that goal is yet far distant.[3]

The alleged conspiracy's failure to achieve its ends in the two decades of its asserted operation is strong evidence that the conspiracy does not in fact exist. Since the losses in such a conspiracy accrue before the gains, they must be "repaid" with interest. And because the alleged losses have accrued over the course of two decades, the conspirators could well require a correspondingly long time to recoup. Maintaining supra competitive prices in turn depends on the continued cooperation of the conspirators, on the inability of other would-be competitors to enter the market, and (not incidentally) on the conspirators' ability to escape antitrust liability for their *minimum* price-fixing cartel. Each of these factors weighs more heavily as the time needed to recoup losses grows. If the losses have been substantial—as would likely be necessary in order to drive out the competition—petitioners would most likely have to sustain their cartel for years simply to break even.

Nor does the possibility that petitioners have obtained supra competitive profits in the Japanese market change this calculation. Whether or not petitioners have the means to sustain substantial losses in this country over a long period of time, they have no motive to sustain such losses absent some strong likelihood that the alleged conspiracy in this country will eventually pay off. The courts below found no evidence of any such success, and—as indicated above—the facts actually are to the contrary: RCA and Zenith, not any of the petitioners, continue to hold the largest share of the American retail market in color television sets. More important, there is nothing to suggest any relationship between petitioners' profits in Japan and the amount petitioners could expect to gain from a conspiracy to monopolize the American market. In the absence of any such evidence, the possible existence of supra competitive profits in Japan simply cannot overcome the economic obstacles to the ultimate success of this alleged predatory conspiracy.

* * *

… [P]etitioners had no motive to enter into the alleged conspiracy. To the contrary, as presumably rational businesses, petitioners had every incentive not to engage in the conduct with which they are charged, for its likely effect would be to generate losses for petitioners with no corresponding gains.…

---

3. Respondents offer no reason to suppose that entry into the relevant market is especially difficult, yet without barriers to entry it would presumably be impossible to maintain supra competitive prices for an extended time. Judge Easterbrook, commenting on this case in a law review article, offers the following sensible assessment:

> … There are no barriers to entry into electronics, as the proliferation of computer and audio firms shows. The competition would come from resurgent United States firms, from other foreign firms (Korea and many other nations make TV sets), and from defendants themselves. In order to recoup, the Japanese firms would need to suppress competition among themselves. On plaintiffs' theory, the cartel would need to last at least thirty years, far longer than any in history, even when cartels were not illegal. None should be sanguine about the prospects of such a cartel, given each firm's incentive to shave price and expand its share of sales. The predation recoupment story therefore does not make sense, and we are left with the more plausible inference that the Japanese firms did not sell below cost in the first place. They were just engaged in hard competition.

Easterbrook, *The Limits of Antitrust*, 63 Texas L. Rev. 1, 26–27 (1984) (footnotes omitted).

* * *

… [T]he absence of any plausible motive to engage in the conduct charged is highly relevant to whether a "genuine issue for trial" exists within the meaning of Rule 56(e). Lack of motive bears on the range of permissible conclusions that might be drawn from ambiguous evidence: if petitioners had no rational economic motive to conspire, and if their conduct is consistent with other, equally plausible explanations, the conduct does not give rise to an inference of conspiracy.… Here, the conduct in question consists largely of (i) pricing at levels that succeeded in taking business away from respondents, and (ii) arrangements that may have limited petitioners' ability to compete with each other (and thus kept prices from going even lower). This conduct suggests either that petitioners behaved competitively, or that petitioners conspired to *raise* prices. Neither possibility is consistent with an agreement among 21 companies to price below market levels. Moreover, the predatory pricing scheme that this conduct is said to prove is one that makes no practical sense: it calls for petitioners to destroy companies larger and better established than themselves, a goal that remains far distant more than two decades after the conspiracy's birth. Even had they succeeded in obtaining their monopoly, there is nothing in the record to suggest that they could recover the losses they would need to sustain along the way. In sum, in light of the absence of any rational motive to conspire, neither petitioners' pricing practices, nor their conduct in the Japanese market, nor their agreements respecting prices and distribution in the American market, suffice to create a "genuine issue for trial."

* * *

The decision of the Court of Appeals is reversed, and the case is remanded for further proceedings consistent with this opinion.

It is so ordered.

Notes and Questions

1. Rationality and Profit Maximization: The Court assumes the Japanese firms behaved rationally by selecting profit maximizing strategies. The plaintiffs' theory, in contrast, required irrational behavior by the Japanese manufacturers. Thus, the Court rejected the plaintiffs' theory as being inconsistent with standard assumptions about economic behavior.

2. Managerial Incentives: The Court implicitly assumes that profit maximization is the goal of the firm, but firms are managed by individuals whose incentives may not always be aligned with this goal. Managers may be more interested in increasing their salary or the size of the firm than in maximizing profits. This divergence between managers' incentives and the firm's profit maximization goal is an example of the **principal-agent problem**. Managers, who are agents of the firm's owners (the principals), do not always act in their principals' best interest. For corporations with publicly traded shares and dispersed shareholders, the principal-agent problem is often characterized as the result of a "separation of ownership and control."

3. Profits and Cartel Agreements: The Court recognizes that each Japanese firm, as a profit maximizer, would have incentives to cheat on the cartel agreement both before and after any success in dominating the market. Again, the assumption of profit maximizing behavior strained the credibility of the plaintiffs' theory.

4. Opportunity Cost of Profits Earned in Japan: The plaintiffs argue that the profits earned in Japan and then, allegedly, used to subsidize losses in the United States, are

somehow not as valuable as other profits — herein lies the incentive to accept the present lost profits in hope of greater future profits from exploiting American consumers. But that cross-subsidization argument ignores the fundamental economic concept of opportunity cost. Profits are profits, regardless of their source. In deciding what to do with those profits — whether to use them to finance continued losses in the American market or to put them to some other use — the Japanese manufacturers necessarily will consider the next best alternative use of those profits, i.e., the opportunity cost. Conversely, the losses the manufacturers suffered in the American market are not lessened in any way by the fact that they were making profits in Japan. As a result, that the manufacturers are making large profits in Japan does not make it any more rational for them to incur ongoing losses in a predatory pricing scheme in the United States than if they had to look to other sources of financing to cover their American losses.

5. Costs of Production and Economic Profit: Profit is defined as total revenue minus total costs. Economists use the phrase **costs of production** in a way that reflects the concept of opportunity cost. In order for a firm to undertake a productive activity, it must attract inputs (resources or factors of production) from other alternative uses. The firm attracts inputs by paying the owners of the inputs at least the value of their services in their next best alternative use — opportunity cost. Costs of production (firm outlays) are the payments made to the owners of resources to assure the availability of those resources for use in the firm's production process. In discussing the costs of production, economists and many courts are careful to recognize both the **explicit costs** recorded in the firm's books and the **implicit costs** that reflect the value of resources used in production by the firm for which no explicit payments are made. When a firm does not make an explicit payment for the use of resources it owns, then the implicit cost is the income that those resources could have commanded in some alternative use. A consideration of explicit and implicit costs reveals the total opportunity cost of the production process.

Consider, for example, a small "Mom & Pop" tavern organized as a sole proprietorship and located on a busy corner of an area undergoing a commercial "boom." "Mom & Pop" own the building in which the tavern is located. Their explicit costs of operation include the payments for inventory, advertising, legal and accounting services, and labor. Each of these payments involves an explicit contractual outlay that is recorded as a cost in the firm's books. On the other hand, "Mom & Pop" may leave out some important implicit costs of production. For example, "Mom & Pop" may not record on the firm books a salary for their own services. Or, they may enter a salary below what they could earn while working for another firm. The opportunity cost of working for themselves is an implicit cost — a real economic cost — that should be relevant in their decision making process. Another implicit cost that may be left out of "Mom & Pop's" accounting records is the opportunity cost of using their building. If "Mom & Pop" did not occupy the building, it could be rented to another business or sold (with the proceeds invested in another income earning asset). The failure to include the implicit cost of using the building means that the firm is ignoring an important opportunity cost: the **opportunity cost of capital** — the value of the payments that they could receive from the next best alternative investment of the capital tied up in the firm. The opportunity cost of capital is typically assumed to be the market rate of return on an asset of similar risk, and is an implicit cost of production.

Including implicit costs as well as explicit costs in the cost calculation lets one determine **economic profit** — the difference between total revenue and total cost of production. In contrast, the **accounting profit** found on a firm's income statement is the amount by which total revenue exceeds total explicit costs. Obviously, if accounting costs do not

account for implicit costs, then accounting profit may overstate actual (economic) profit. In other words, positive accounting profit does not necessarily imply that positive economic profit is being earned. To the contrary, a positive accounting profit can be consistent with an economic loss.

A final, related point about a firm's cost of production concerns the accounting convention of relying on **historical costs**, i.e., payments that were incurred in the past. Historical cost means that the accounting value of an asset—such as inventory or a building—is determined by the price that the firm paid for the asset. This is the exact opposite of the perspective inherent in the concept of opportunity cost. Opportunity cost is forward-looking—it involves a decision today to commit resources to one use as opposed to another use. Thus, in a real sense, opportunity cost is the value of "the road not taken." Decisions by economic actors rely on predictions about the future, not on past events. Past outlays (unless they affect prospective costs) are not costs because they exert no influence on current decisions.

Consider, for example, the market value today of a commercial building purchased in 1995 for $1,000,000. If the market price of the building has fallen to $500,000, should the owner refuse to sell at that price simply because its current **book value** (historical cost minus accumulated depreciation) is $800,000? Alternatively, if the owner of the building tried to use it as collateral for a bank loan, would the banker use historical value or current market value as an indicator of its value as collateral? The prudent banker will look at the market value.

6. Economic Survival of the Fittest: What if businesses did not seek to maximize profits, but instead made business decisions for no apparent reason and changed their strategies entirely at random? Wouldn't the most successful businesses still be those that (however randomly) chose strategies that were the most profitable? And wouldn't businesses that chose unprofitable strategies lose money and eventually go out of business? Thus, even if businesses do not consciously follow a profit maximization strategy, the results of marketplace competition will look much as if they did—the economic fittest will survive. This process is often referred to as **Economic Darwinism**. James Brickley, Clifford Smith, and Jerold Zimmerman apply this perspective to competition among firms: "The collapse of Enron, Charles Darwin might have noted, is an example of how competition tends to weed out the less fit. As described in *The Origin of Species*, natural history illustrates the principle of 'survival of the fittest.' In industry, we see Economic Darwinism in operation as competition weeds out ill-designed organizations that fail to adapt. Competition in the market provides strong pressures for efficient decisions. Competition among firms dictates that only those firms with low costs will survive. If firms adopt inefficient, high-cost policies, competition will place strong pressures on these firms to either adapt or close." James Brickley et al., Managerial Economics and Organizational Architecture 7 (3rd ed. 2004).

B. Property Rights and Exchange in a Free Market Economy

The fundamental problem faced by any economic system is the allocation of society's scarce resources among the unlimited desires of the individuals who make up that society. A capitalistic market economy, which is the basis of the American economic system, attempts to solve this social problem by tapping the individual's self-interest in a manner

Free Market Environmentalism [redacted] (2001)

CHAPTER 1

VISIONS OF THE ENVIRONMENT

For most people, markets are the cause of environmental problems, not the solution. The very notion of free market environmentalism is an oxymoron. Even conservative thinkers who support free enterprise and free trade find themselves uncomfortable with the idea of letting unfettered markets determine how and when natural resources are used. Markets may work fine to produce shoes or software, but the environment is somehow different and is too precious to be thrown to bulls and bears on Wall Street.

The feeling that markets and the environment do not mix is buttressed by the perception that resource exploitation and environmental degradation are inextricably linked to economic growth. This view, which first emerged with industrialization, builds on fears that we are running out of resources. After all, if resources are finite and production to meet material needs uses up some of those finite resources, the world's natural endowment must be getting more scarce. During the Industrial Revolution in England, the Reverend Thomas Malthus articulated this view, hypothesizing that exponential population growth would eventually overwhelm productivity growth and result in famine and pestilence. At the heart of Malthus' logic: population and growing consumption must eventually run into the wall of finite natural resources.

Earth Day 1970 started a plethora of doomsday predictions. Ecologist Kenneth Watt declared: "We have about five more years at the outside to do something," and Harvard biologist George Wald predicted that "civilization will end within 15 or 30 years unless immediate action is taken against problems facing mankind." Barry Commoner wrote: "We are in an environmental crisis which threatens the survival of this nation, and of the world as a suitable place of human habitation."[1]

Modern-day Malthusians have given such dire predictions an aura of credibility by using complex computer models to predict precisely when Malthusian calamities will occur. In the early 1970s, a group of scientists from the prestigious Massachusetts Institute of Technology concluded:

> If the present growth trends in world population, industrialization, pollution, food production, and resource depletion continue unchanged, the limits to growth on this planet will be reached sometime within the next one hundred years. The most probable result will be a rather sudden and uncontrollable decline in both population and industrial capacity.[2]

In a graph generated by their computer model, the scientific team showed that the "uncontrollable decline" would begin shortly after the turn of the century—in 2005, to be exact—with a precipitous decline in industrial output, food supplies, and population.[3]

The *Global 2000* report commissioned by President Jimmy Carter arrived at similar conclusions in 1980. It predicted:

> If present trends continue, . . . the world in 2000 will be more crowded, more polluted, less stable ecologically, and more vulnerable to disruption than the world we live in now. Serious stresses involving population, resources, and environment are clearly visible ahead.[4]

In every resource category, *Global 2000* predicted overuse and declines in quantity and quality.

Each year the Worldwatch Institute, headed by Lester R. Brown, makes headlines with its Malthusian predictions in its annual flagship publication, *State of the World*. In *State of the World 1998*, Brown states:

> Global land and water resources are not sufficient to satisfy the growing grain needs in China if it continues along the current development path. Nor will the oil resources be available, simply because world oil production is not projected to rise much above current levels in the years ahead as some of the older fields are depleted, largely offsetting output from newly discovered fields.[5]

He goes on to predict food crisis due to land and water scarcities: "The world is moving into uncharted territory on the food front, facing a set of problems on a scale that dwarfs those in the past."[6]

And year after year, the predictions from Worldwatch are proven wrong. The world does not run out of food; famine and pestilence are not on the increase; and the environment is generally not more, but rather less, polluted.[7] The one area where Worldwatch predictions tend to be more accurate is in the case of water, but even here they miss the mark. At the Second World Water Forum held in The Hague in March 2000, for example, the Food and Agriculture Organization of the United Nations (FAO) refuted claims that water shortages and food scarcity were imminent dangers. A recent FAO analysis of 93 developing countries revealed that the growing demands on global water are actually slowing. One FAO official commented that the "developing countries should be able to boost food production significantly by increasing irrigated cropland by about one third, yet using only 12 percent more water."[8]

The reason that Malthusian hypotheses are continually refuted is that they fail to take into account how human ingenuity stimulated by market forces finds ways to cope with natural resource constraints. As the late Julian Simon observed, the "ultimate resource" is the human mind, which has allowed us to avoid Malthusian cycles.[9] Human ingenuity is switched on by market prices that signal increasing scarcity and provide rewards for those who mitigate resource constraints by reducing consumption, finding substitutes, and improving productivity. For example, it was rising copper prices that helped stimulate the switch to fiber optics and satellite communication as substitutes for millions of miles of copper wire. Rising prices of aluminum and human ingenuity did more to reduce the thickness of drink cans and thus conserve resources used to produce and transport aluminum than did recycling regulations and campaigns. In fact, in the 1960s and the 1970s, the amount of aluminum in soft drink cans was reduced by 32 percent. These examples of aluminum and copper, two finite resources that Malthusians would doom to extinction, show that the human mind is a powerful tool, particularly when the incentives are right.

Julian Simon was so confident in the "ultimate resource" that he proposed a simple test to another modern-day Malthusian, Paul Ehrlich, author of *The Population Bomb*.[10] To test their competing theories, Simon offered a wager based on resource prices. He said that if Ehrlich's scarcity hypothesis was correct, it should be reflected in rising resource prices. On the other hand, if Simon's confidence in human ingenuity was correct, resource prices should fall. In 1980, Simon proposed that Ehrlich choose five resources and that they hypothetically purchase $200 worth of each. Ten years later they would see whether the $1,000 worth of resources had risen or fallen in value. If they had risen, Ehrlich would be vindicated and Simon would have to pay him the increase, or vice versa. Not surprisingly, Ehrlich accepted the wager and chose tungsten, copper, chrome, nickel, and tin. When they revisited the values in 1990, every single commodity chosen by Ehrlich had fallen in price even without adjusting for inflation. In fact the $1,000 bundle of five resources was worth less than $500 after adjusting for inflation. Ehrlich paid up, saying that he was still confident in his theory and that it would take longer for scarcity to manifest itself, but declined to accept Simon's offer to repeat the wager for $10,000.

Even for cases in which neo-Malthusians agree that human ingenuity can help mitigate the impacts of scarcity, they generally want top-down regulations and incentives from government to stimulate the ingenuity. For example, prices may encourage people to conserve energy or water, but energy and water markets cannot necessarily be depended upon to give the correct signals. In the case of ethanol at current and expected future prices, the corn-based alternative fuel would not be competitive with petroleum-based fuels, so the federal government subsidizes ethanol production. In the case of water, Americans would not be installing low-flow toilets were it not for federal regulations. And where prices are advocated as a stimulus to conservation, they are to be set by government, not by markets. According to Sandra Postel of the World Watch Institute,

Wherever pricing and marketing fail to take into account the full social, environmental, and intergenerational costs of water use, some additional correction is needed. . . . In the case of fossil aquifers, such as the Ogallala in the U.S. High Plains or the deep desert aquifers in Saudi Arabia and Libya, a "depletion tax" might be levied on all groundwater extractions.[11]

From land to water to air, governmental control—which means political control—is seen as a necessary check on the environmental ravages of free markets.

This book will challenge the perception that free market environmentalism is an oxymoron and indeed argue that if we are to continue improving environmental quality in the twenty-first century, we must harness market forces. We offer an alternative way of thinking about natural resource and environmental concerns that contrasts information and incentives generated by markets and politics. In general, free market environmentalism emphasizes the positive incentives associated with prices, profits, and entrepreneurship, as opposed to political environmentalism, which emphasizes negative incentives associated with regulation and taxes.

At the heart of free market environmentalism is a system of well-specified property rights to natural and environmental resources. Whether these rights are held by individuals, corporations, nonprofit environmental groups, or communal groups, a discipline is imposed on resource users because the wealth of the property owner is at stake if bad decisions are made. Moreover, if private owners can sell their rights to use resources, the owners must not only consider their own values, they must also consider what others are willing to pay. In the market setting, it is the potential for gains from trade that encourages cooperation. Both the discipline of private ownership and the potential for gains from trade stand in sharp contrast to the political setting. When resources are controlled politically, the costs of misuse are more diffused and the potential for cooperation is minimized because the rights are essentially up for grabs.

Perhaps the best example of how these characteristics of private ownership can enhance environmental quality comes from the Nature Conservancy, the largest environmental group in the United States that depends on private landownership. When the conservancy obtains title to a parcel of land, it uses a formal system for evaluating whether the property has significant ecological value. Consider the case in which the Wisconsin Nature Conservancy was given title to beachfront property on St. Croix, Virgin Islands. One might think that the Nature Conservancy would go to great lengths to prevent development of oceanside property in the Caribbean. But, indeed, it actually traded the property for a much larger tract in Wisconsin and allowed selective beachfront development to occur under some protective covenants.

Why would an environmental group let this happen? The answer, in a word, is tradeoffs. As owner of the beach, the Wisconsin Nature Conservancy had to ask what would be gained and what would be sacrificed if development was prevented. The gain was obviously beachfront protection. The sacrifice may not be obvious to the casual observer, but it was obvious to the organization. At the time, the Wisconsin Nature Conservancy was actively trying to complete pro-

tection of a watershed in southern Wisconsin. It did not have the money to buy the last parcel of land needed to complete the protection, but it saw an opportunity to trade St. Croix beachfront for that rocky hillside. As owner of the St. Croix property, the Nature Conservancy faced the cost of just saying no or reaping the benefit of an entrepreneurial trade. The discipline and the incentives of private ownership forced the conservancy to make careful decisions and allowed it to accomplish its goal of saving a watershed. As a result, the Nature Conservancy's wealth in the form of environmental amenities was enhanced.

The emphasis of free market environmentalism on private ownership and decentralized decision making should not be taken to mean that there is no role for government. On the contrary, government has an integral role to play in the definition and enforcement of property rights. In the absence of the rule of law, the incentives inherent in private ownership disappear and with them goes the potential for environmental stewardship. With clearly specified titles obtained from land recording systems, strict liability rules, and adjudication of disputed property rights in the courts, market processes can encourage owners to carefully weigh costs and benefits and to look to the future.

Free market environmentalism conflicts with traditional environmentalism in its visions regarding human nature, knowledge, and processes. A consideration of these visions helps explain why some people accept this way of thinking as an alternative to bureaucratic control and why others reject it as a contradiction in terms.

Human nature: Free market environmentalism views man as self-interested. This self-interest may be enlightened to the extent that people are capable of setting aside their own well-being for close relatives and friends or that they may be conditioned by moral principles. But beyond this, good intentions will not suffice to produce good results. Developing an environmental ethic may be desirable, but it is unlikely to change basic human nature. Instead of intentions, good resource stewardship depends on how well social institutions harness self-interest through individual incentives.

Knowledge: In addition to incentives, good resource stewardship depends on the information available to individuals who make decisions about resource use. Free market environmentalism views this information as being time- and place-specific rather than general and concentrated in the hands of experts. Whether we are considering interactions among humans or interactions between humans and nature, the information necessary for good management varies significantly from time to time and from place to place. Certainly there is some knowledge that is general and concentrated in the minds of experts, such as the laws of physics or principles of ecology, but the complexity of ecosystems makes them impossible to model and therefore to manage from afar. Given that this type of time- and place-specific information cannot be gathered in a single mind or groups of minds that can account for the multiple interconnections of ecosystems, decentralized management guided by the incentives of private ownership becomes an alternative to centralized political control.

Process or solutions: These visions of human nature and knowledge combine to make free market environmentalism a study of process rather than a prescription

for solutions. If man can rise above self-interest and if knowledge can be concentrated, then the possibility for solutions through political control is feasible. But if there are self-interested individuals with diffuse knowledge, then processes must generate a multitude of solutions conditioned by the costs and benefits faced by individual decision makers. By linking wealth to good stewardship through private ownership, the market process generates many entrepreneurial experiments; and those that are successful will be copied, while those that are failures will not. The question is not whether the right solution will always be achieved, but whether good decisions are rewarded and bad ones penalized.

These three elements of free market environmentalism—self-interest, information, and process—also characterize the interaction of organisms in nature. Since Charles Darwin's revolutionary study of evolution, most scientific approaches have implicitly assumed that self-interest generally dominates behavior for higher as well as lower forms of life. Individual members of a species may act in altruistic ways and may cooperate with other species, but species survival depends on adjustments to changing parameters in ways that enhance the probability of individual and species survival. To assume that man is not self-interested or that he can rise above self-interest requires heroic assumptions about *Homo sapiens* vis-à-vis other species.

Ecology also emphasizes the importance of time- and place-specific information in nature. Because the parameters to which species respond vary considerably within ecosystems, each member of a species must respond to time- and place-specific characteristics with the knowledge that each possesses. These parameters can vary widely, so it is imperative for survival that responses utilize the diffuse knowledge. Of course, the higher the level of communication among members of a species, the easier it is to accumulate and concentrate time- and place-specific knowledge. But a giant leap of faith is necessary if humans are to be able to accumulate and assimilate the necessary knowledge to manage the economy or the environment. Evidence from Eastern Europe underscores the problems that can arise with centralized management of either the economy or the environment.[12]

Continuing the analogy between ecology and free market environmentalism is instructive for thinking about the implications for policy. When a niche in an ecosystem is left open, a species will profit from filling that niche and will set in motion a multitude of other adjustments. If an elk herd grows because there is abundant forage, there will be additional food for predator species such as bears and wolves. Their numbers will expand as they take advantage of this profit opportunity. Individual elk will suffer from predation as elk numbers will be controlled. Plant species will survive and other vertebrates such as beavers will be able to survive. This is a process that no central planner could replicate because there is no best solution for filling niches and because each species is reacting to time- and place-specific information.

Comparing free market environmentalism with ecosystems serves to emphasize how market processes can be compatible with good resource stewardship and environmental quality. As survival rewards species that successfully fill a niche, increased wealth rewards owners who efficiently manage their resources.

Profits link self-interest with good resource management by attracting entrepreneurs to open niches. If bad decisions are being made, then a niche will be open. Whether an entrepreneur sees the opportunity and acts on it will depend on his or her ability to assess unique information and act on that assessment. As with an ecosystem, however, the diffuse nature of this information makes it impossible for a central planner to determine which niches are open and how they should be filled. If the information or incentives are distorted because property rights are incomplete or because decision makers receive distorted information through political intervention, then the market process will not necessarily generate good stewardship.

Visions of what makes good environmental policy will change only if we realize that our current visions are not consistent with reality. We must ask ourselves whether well-intentioned individuals armed with sufficient information dominate the political decisions that affect natural resources and the environment. Forest policy analyst and environmentalist Randal O'Toole answered this question in the context of the U.S. Forest Service.

> While the environmental movement has changed more than the Forest Service, I would modestly guess that I have changed more than most environmental leaders. . . . In 1980, I blamed all the deficiencies in the markets on greed and big business and thought that government should correct these deficiencies with new laws, regulatory agencies, rational planning, and trade and production restrictions. When that didn't work, I continued to blame the failure on greed and big business.
>
> About 1980, someone suggested to me that maybe government didn't solve environmental or other social problems any better than markets. That idea seemed absurd. After all, this is a democracy, a government of the people, and what the people want they should be able to get. Any suggestion that government doesn't work was incomprehensible.
>
> But then I was immersed in the planning processes of one government agency for ten years (sort of like taking a Berlitz course in bureaucratic speaking). I learned that the decisions made by government officials often ignored the economic and other analyses done by planners. So much for rational planning. Their decisions also often went counter to important laws and regulations. So much for a democratic government.
>
> Yet I came to realize that the decisions were all predictable, based mainly on their effects on forest budgets. . . .
>
> I gradually developed a new view of the world that recognized the flaws of government as well as the flaws in markets. Reforms should solve problems by creating a system of checks and balances on both processes. . . . The key is to give decision makers the incentives to manage resources properly.[13]

This book provides a "Berlitz course" in free market environmentalism that challenges entrenched visions. Because free market environmentalism depends on clearly specified property rights and accompanying price signals, it works

better for some resources than for others. Markets can allocate land, water, and energy better than they can water quality or the global atmosphere. As we shall see, even on the western frontier, free market environmentalism was at work as cattlemen and farmers developed property rights to land and water and avoided the tragedy of the commons. In contrast, since one-third of the nation was set aside in the late nineteenth century for federal management, special interest politics has resulted in fiscal and environmental mismanagement. But even in the case of public lands, the implications of free market environmentalism are being utilized to improve incentives for federal managers by allowing them to charge user fees and to keep those fees for reinvestment in improving those lands. Similarly, in the case of water allocation, markets are improving efficiency and increasing instream flows for fish, wildlife, and recreation.

If land and water allocation are easy problems for free market environmentalism, pollution concerns challenge the paradigm. Again, however, the focus on property rights, with the accompanying right to be free from trespass by pollutants, provides a way of thinking how polluters can be held accountable. If it is possible to identify who is releasing pollutants into the soil, water, or air and to determine what the impacts of those pollutants are, then broad regulations can be replaced with negotiations between those who are harmed and those who are causing the pollution. In this way, property rights allow those who want cleaner land, water, or air to charge those who want to use it for waste disposal and hence make polluters accountable for costs they create. As we shall see in later chapters, measuring and monitoring pollution is not always easy, but just as barbed wire lowered the cost of fencing the western frontier, technology can lower the costs of holding polluters accountable.

By confronting our entrenched visions about the interface between markets and the environment, we can move beyond the status quo of political regulation and harness market forces to improve environmental quality. Included in these forces are environmental entrepreneurs who can fill market niches that allocate natural resources more efficiently and supply environmental amenities that we demand. These forces also require voluntary exchanges between property owners that promote cooperation and compromise, in contrast to political regulations that tear at the social fabric. In short, free market environmentalism offers an alternative that channels the heightened environmental consciousness into some win-win solutions that can sustain economic growth, enhance environmental quality, and promote harmony. Now join us in considering the details of this alternative vision.

What Do Judges and Justices Maximize?
(The Same Thing Everybody Else Does)

[redacted]

This article presents a positive economic theory of the behavior of appellate judges and Justices. The essay argues that the effort to insulate judges from significant economic incentives, through devices such as life tenure and stringent conflict of interest rules, has not rendered judicial behavior immune to economic analysis. Drawing on analogies to the managers of non-profit enterprises, to those who vote in political elections, and to theatrical spectators, the essay models the judicial utility function in terms that allows judges to be seen as ordinary people responding rationally to ordinary incentives. This model provides a theoretical alternative to the common view of judges as Prometheans or saints, and it suggests new ways of looking at such practical issues as the design of the judicial compensation system.

I. Introduction

This article proposes a new, positive economic theory of judicial behavior. My focus is on the "ordinary" appellate judge with secure tenure (for example, a federal court of appeals judge or U.S. Supreme

* Judge, United States Court of Appeals for the Seventh Circuit; Senior Lecturer, University of Chicago Law School. Earlier versions of this article were delivered at a conference on economic analysis of law sponsored by the George Mason University School of Law, on January 28, 1993, and at the annual meeting of the American Law and Economics Association, on April 30, 1993. I thank the participants in both conferences, plus Gary Becker, Douglas Baird, Harold Demsetz, Frank Easterbrook, David Friedman, Henry Hansmann, Lawrence Lessig, Nelson Lund, Martha Nussbaum, Mark Ramseyer, Eric Rasmusen, and Michael Trebilcock, for many helpful comments, and Benjamin Aller for valuable research assistance.

Court Justice), as distinct from judicial "titans," real or imagined, on whom most of the previous literature explicitly or implicitly focuses. Analogizing judges to the managers of nonprofit enterprises, to voters in political elections, and to spectators at theatrical performances, I present a simple model in which judicial utility is a function mainly of income, leisure, and judicial voting. I show how this model can be used to explain various judicial behaviors (ranging from stare decisis to what I call "go-along voting"), to make falsifiable predictions concerning judicial effort, and to provide a framework both for evaluating changes in judicial compensation and rules of conduct and for comparing judicial with legislative behavior. Although my principal interest is in the economic theory of judicial behavior (with some borrowings from "law and literature" in Part V), the approach may have value for the solution of practical problems of judicial administration, as well as some relevance to the economic analysis of other occupations in which nonpecuniary income is a large part of total compensation.

The motivation for the article should be plain. At the heart of economic analysis of law is a mystery that is also an embarrassment: how to explain judicial behavior in economic terms, when almost the whole thrust of the rules governing compensation and other terms and conditions of judicial employment is to divorce judicial action from incentives—to take away the carrots and sticks, the different benefits and costs associated with different behaviors, that determine human action in an economic model. Since the judges are the central actors in the drama of the common law and play lead roles in statutory and constitutional law as well, the failure thus far to explain their actions in economic terms mocks the claim of economic analysis to explain the salient features, institutional as well as doctrinal, of the law in general and the common law and other judge-made law in particular. The economic analyst has a model of how criminals and contract parties, injurers and accident victims, parents and spouses—even legislators, and executive officials such as prosecutors—act, but falters when asked to produce a model of how judges act.

Instead of trying to explain directly why judges adopt one judicial philosophy or another, or decide a particular case one way or another, or follow or don't follow precedent, I shall concentrate on the anteced-

ent question: Are judges rational? Or have the elaborate efforts made to strip them of incentives placed their behavior beyond the reach of rational-choice models? I shall argue that judges *are* rational and specifically that they can fruitfully be viewed as composites of three types of rational maximizer: the manager of a nonprofit enterprise, the voter, and the theatrical spectator. I bring in a fourth type, the player of games for fun, to explain the ruled quality of most judicial decision-making without having to assume that judges are unusually selfless or self-disciplined. Although none of these types is easy to model in rational-maximizer terms, enough progress has been or can be made in fitting them to the economic model to provide a framework for understanding judicial behavior as rational too. This framework makes it possible to consider in a new light not only the question of judicial rationality but also more practical issues such as limits on judicial moonlighting, the difference between dictum and holding, trends in judicial conflict-of-interest rules, the effect of higher judicial salaries, the tension between judges and law professors, and whether judges are hard-working and conscientious.

My approach downplays the "power trip" aspect of judging, the focus of most of the few previous efforts to model the judicial utility function. In fact, I assume that trying to change the world plays no role in that function. Not that judges are indifferent to power; they enjoy, I shall argue, the power that goes with deciding cases. But only a small minority, whom I shall largely ignore, have a visionary or crusading bent.[1] We know that the framers of the Constitution attempted to design a government that could be operated by moral and intellectual mediocrities, a characterization of officialdom from which not even federal judges are exempt. The inclusion in the design of such rules or institutions as life tenure for judges suggests that the framers believed that judges could not be counted on to behave with consistent courage, although some of course have behaved so.[2] Politics, personal friendships, ideology, and pure serendipity play too large a role in the

---

[1] Hence the common criticism of the Supreme Court in the Burger years as lacking any "vision" or sense of mission. This is true of most courts most of the time, but is not necessarily a good criticism.

[2] Notably the federal judges who supervised public school desegregation in the South after *Brown v Board of Education.*

appointment of federal judges to warrant treating the judiciary as a collection of genius-saints miraculously immune to the tug of self-interest.[3] By treating judges and Justices as ordinary people,[4] my approach makes them fit subjects for economic analysis; for economists have no theories of genius.[5] It is fortunate for economic analysis, therefore, that most law is made not by the tiny handful of great judges but by the great mass of ordinary ones, although I shall consider briefly in Part VI the extension of my analysis to extraordinary judges. Because there are so many ordinary judges, and because anti-intellectualism runs deep in the American soul, there is even a cult of ordinariness in judging. Exceptionally able judges arouse suspicion of having an "agenda," that is, of wanting to be something more than just corks bobbing on the waves of litigation or umpires calling balls and strikes.

My focus is on federal appellate judges: not only because I am one and know this group of judges best,[6] but also because efforts to strip away incentives have progressed furthest with this group. Article III of the Constitution erects such a high hurdle to removing a federal judge from office that pretty much the only thing that will get him removed is criminal activity. A federal judge can be lazy, lack judicial temperament, mistreat his staff, berate without reason the lawyers who appear

---

[3] We are beginning, at last, to get a glimpse, from judges, of some of the unvarnished realities of a judicial career. Besides Richard A. Posner, *The Problems of Jurisprudence* (Harvard, 1990), see Frank H. Easterbrook, *What's So Special about Judges?* 61 U Colo L Rev 773 (1990), and Patricia M. Wald, *Some Real-Life Observations about Judging,* 26 Ind L Rev 173 (1992).

[4] "Most judges, even Supreme Court Justices, have been plucked from well deserved intellectual obscurity." Charles W. Collier, *The Use and Abuse of Humanistic Theory in Law: Reexamining the Assumptions of Interdisciplinary Legal Scholarship,* 41 Duke L J 191, 221 (1991). From political obscurity as well, one might add.

[5] But see Daniel L. Rubenson and Mark A. Runco, *The Psychoeconomic Approach to Creativity,* 10 New Ideas in Psychology 131 (1992), for a stab at one.

[6] In so saying I do not mean to suggest that judges have privileged access to the judicial utility function. A utility function is not a psychological or phenomenological concept but a device for generating hypotheses. I doubt that *any* judge subjectively experiences his work in the way modeled in this article. I know I do not.

before him, be reprimanded for ethical lapses, verge on or even slide into senility, be continually reversed for elementary legal mistakes, hold under advisement for years cases that could be decided perfectly well in days or weeks, leak confidential information to the press, pursue a nakedly political agenda, and misbehave in other ways that might get even a tenured civil servant or university professor fired; he will retain his office. His pay cannot be lowered, either—and neither can the pay of a good judge be raised. All judges of the same rank are paid exactly the same, and so the carrot is withdrawn along with the stick. Another reason there is no carrot is that a judge is, of course, forbidden to accept bribes from litigants, to pocket filing fees or other fees levied on litigants, or to collect royalties from people who cite his opinions. He gets a fixed salary, period.

Well, there is a small carrot. Supreme Court Justices are often appointed from the ranks of federal court of appeals judges,[7] and although the probability of such an appointment is low for any particular judge even if he or she is one of the relatively few who are talked about for such a promotion, it figures in the thinking of some judges. But the impact of a particular decision on the prospects for promotion is normally very slight. Some decisions have no impact at all on those prospects and in the case of almost all the remaining decisions the impact is unpredictable—the decision may offend as many influential people as it pleases.

Promotion to jobs outside the judiciary is discouraged by the structure of judicial compensation, which is heavily backloaded.[8] The pension is extraordinarily generous (one's final salary, for life, including annual cost-of-living increases), but there is no vesting until age 65,

---

[7] Indeed, at present seven of the nine Justices—Blackmun, Stevens, Scalia, Kennedy, Souter, Thomas, and Ginsburg—are former federal court of appeals judges, though Souter was one for only a few months and Thomas for only a year and a half. Historically the percentage has been lower.

[8] The incentive effects of different methods of compensating employees and other agents are the subject of an extensive economic literature, pertinently illustrated by Edward Lazear and Sherwin Rosen, *The Economics of Compensation of Government Officials,* in Robert W. Hartman and Arnold R. Weber, eds, *The Rewards of Public Service: Compensating Top Federal Officials* 101 (Brookings, 1980), and by Ronald N. Johnson and Gary D. Libecap, *Agency Growth, Salaries and the Protected Bureaucrat,* 27 Econ Inquiry 431 (1989). On the effect of the backloading of compensation on the turnover

so anyone who quits sooner gives up a large expected benefit. The attractive pension arrangements are important for inducing judges of advanced age to retire. Article III presumably outlaws mandatory retirement for Article III judges, and makes it very difficult to force judges to retire even for cause. So here a carrot must do all the work.

The compensation and tenure structure for federal district judges is the same as that for appellate judges except that the salary is slightly lower and they have greater prospects for promotion,[9] although for the

---

of federal civil servants, see Richard A. Ippolito, *Why Federal Workers Don't Quit*, 22 J Human Resources 281 (1987) ("Ippolito, *Why Federal Workers Don't Quit*"). I shall take for granted the compensation and tenure regime of federal judges.

[9] Mark A. Cohen, *The Motives of Judges: Empirical Evidence from Antitrust Sentencing*, 12 Intl Rev L & Econ 13 (1992), presents some empirical evidence that desire for promotion affects the behavior of district judges. Another article by Cohen provides additional empirical evidence for a district-judge utility function that contains such arguments as desire to exercise discretion, aversion to heavy workloads, and desire for promotion to the court of appeals. Mark A. Cohen, *Explaining Judicial Behavior or What's "Unconstitutional" about the Sentencing Commission?*, 7 J L Econ & Org 183 (1991) ("Cohen, *Explaining Judicial Behavior*"). Other papers that analyze judicial behavior in utility-maximizing terms are Robert D. Cooter, *The Objectives of Private and Public Judges*, 41 Public Choice 107 (1983) ("Cooter, *Objectives*"); Jeffrey N. Gordon, *Corporations, Markets, and Courts*, 91 Colum L Rev 1931, 1967–71 (1991); Richard S. Higgins and Paul H. Rubin, *Judicial Discretion*, 9 J Legal Stud 129 (1980) ("Higgins and Rubin, *Judicial Discretion*"); Thomas J. Miceli and Metin M. Cosgel, *Reputation and Judicial Decision-Making* (forthcoming in J Econ Behavior & Org); Erin O'Hara, *Implicit Collusion or Social Constraint: Toward a Game Theoretic Analysis of Stare Decisis* (unpublished manuscript, Clemson University Departments of Legal Studies and of Economics) ("O'Hara, *Implicit Collusion*"); Bruce H. Kobayashi and John R. Lott, Jr., *Judicial Reputation and the Efficiency of the Common Law* (unpublished manuscript, 1993) (George Mason University School of Law) ("Kobayashi and Lott, *Judicial Reputation*"). Kobayashi and Lott point out that a judge who wanted to maximize citations to his decisions would have an incentive to upset existing precedents—and to replace them not only with his own decisions but with his own *inefficient* decisions, because, other things being equal, inefficient decisions are likely to generate more litigation, hence more occasions for citing the judge's decisions. Such a judge, however, belongs more to the "Promethean" than to the "ordinary" category, and it is the latter on which I focus. I leave aside the question whether a true Promethean would be likely to be a simple citation maximizer.

most part only to the court of appeals. More important in keeping them in line is the fact that a district judge presides more or less continuously at trials and other proceedings in open court in which he is required to make rulings and talk to lawyers and jurors. If he isn't on the ball this soon becomes known and he gets a bad reputation in the legal community. Appellate judges, in contrast, are largely shielded from direct evaluations of their work. They never have to make rulings in open court, or indeed open their mouth in court. Provided they can pick competent law clerks, they will be able to churn out, regardless of their own efforts or ability, professionally competent opinions—and opinions are virtually their only public product and hence virtually the only basis upon which the legal profession, or the rest of the world, can evaluate them. It is the unique insulation of federal appellate judges from accountability that makes their behavior such a challenge to the economic analysis of law, and more broadly to the universalist claims of the economic theory of human behavior.

[redacted]

[10] See, for just a few examples of this literature, William A. Niskanen, *Bureaucracy and Representative Government* (Aldine, Atherton, 1971); Randall L. Calvert, Matthew D. McCubbins, and Barry R. Weingast, *A Theory of Political Control and Agency Discretion,* 33 Am J Pol Sci 588 (1989); Mark Bagnoli and Michael McKee, *Controlling the Game: Political Sponsors and Bureaus,* 7 J L Econ & Org 229 (1991).

[11] Henry B. Hansmann, *The Role of Nonprofit Enterprise,* 89 Yale L J 835 (1980); see also Henry B. Hansmann, *Ownership of the Firm,* 4 J L Econ & Org 267

V. The Judge as Spectator and as Gamester

A difficult question remains—that of the judge's motivation, when all monetary punishments and rewards have been stripped away and a choice between work and leisure is not in the offing, to vote for one side rather than another, or to vote for one interpretation of a statute or legal doctrine rather than another, or to adopt one judicial philosophy (such as "conservative," "liberal," "activist," or "restrained") rather than another. The traditional objection to the secret ballot—that it promotes, or at least protects, irresponsible voting—has carried the day with respect to voting by judges. Every judicial vote is public,[35] although sometimes a judge will tell his friends that he joined an opinion with which he disagreed because he didn't think the issue important enough to warrant a dissenting opinion. The public character of judicial voting facilitates criticism, which can be expected to have a greater effect on behavior when ordinarily more powerful incentives, such as money, are not in play. Yet most judges in fact are relatively insensitive to criticism other than by other judges, believing conveniently that most of it is motivated by political disagreement, envy, ignorance (willful or otherwise) of the conditions under which judges work, and self-promotion. Moreover, public comment on judicial decisions other than by the Supreme Court is rare. Only a tiny fraction of the tens of thousands of other appellate opinions published each year receive any sort of critical attention that might get back to the judge and alter his future behavior.

Choices of the kind that face a judge who must vote in a case—choices that cannot be made on the basis of wanting to increase one's pecuniary income, leisure, fame, or other forms of utility—are common in other areas of living. They are for example the choices we make when watching dramatic or cinematic performances. Athletic contests are different, mainly because of the built-in "bias" in favor, normally, of the "home" team, a bias that makes the judicial analogy

[35] Not a universal feature of judging: Mark Ramseyer tells me that in Japan, when lower-court judges sitting in three-judge panels decide cases, the votes are not made public and dissents are never written.

strained. The bias is highly relevant to state court adjudication, however, and can help explain not only the federal diversity jurisdiction but also the exclusive federal jurisdiction over many types of case that pit a state resident against federal taxpayers.

The audience for a play or movie is detached, having no tangible stake in the outcome of whatever struggle is being depicted on stage or screen. Yet ordinarily it is induced to "choose" one side or the other. Usually the choice is manipulated by the author—he "tells" us as it were to side with the hero against the villain. But in dramatic works of deep ambiguity, which often are highly popular among intellectuals on that account, such as *Hamlet,* or *Measure for Measure,* or *Pygmalion,* the choice offered to the spectator is a real one, because the author either has not resolved in his own mind the central tension in the situation dramatized or has not been able (or desired) to communicate the resolution clearly. This explains the popularity of revisionist interpretations of literature, such as arguing that the real hero of *Paradise Lost* is Satan. The spectator, or, in the last example, the reader (but a "live" performance provides a closer analogy to the judicial process, though today many cases are submitted for decision without any oral argument or other hearing—much as when a play is read rather than performed), has to weigh the evidence and come to a decision. The position of the judge is similar.[36] If spectators get consumption value out of such choices, it is not surprising that judges do.

Spectators make choices about the meaning of a play or movie by bringing to bear their personal experiences and any specialized cultural competence that they may have by virtue of study of or immersion in the type of drama that they are watching, and often by discussing their reactions with friends who may have a similar competence. The judge brings to bear on his spectatorial function not only a range of

---

[36] The analogy between the judge and the literary reader or spectator is developed, though for a different purpose, in Martha C. Nussbaum, *Equity and Mercy* (forthcoming in Philosophy and Public Affairs). On the dramaturgical character of trials, see Milner S. Ball, *The Play's the Thing: An Unscientific Reflection on Courts under the Rubric of Theater,* 28 Stan L Rev 81 (1975), and on the forensic character of drama, see Kathy Eden, *Poetic and Legal Fiction in the Aristotelian Tradition* 176–83 (Princeton, 1986). There has been some economic analysis of the arts, but I have found none that bears on the issues in this article.

personal and political preferences, but also a specialized cultural competence—his knowledge of and experience in "the law." And if he is an appellate judge he will often discuss with his professional colleagues the proper outcome of the contest before making up his mind.

Of course not every case has the rich ambiguities of *Hamlet*. Many cases involve puzzles soluble with the technical tools of legal analysis—here the judge is like the reader of a detective story. The jury as factfinder performs a similar function. It is a different kind of spectatorship from the one I am stressing here, that of the appellate judge asked to decide not where truth lies but which party has the better case. But in either case the choice, like that of the theater audience, is a disinterested one; the judge's or jury's income is not affected by it. A further point is that the less informed the tribunal is, the more "dramatic" the trial must be to hold the "audience's" attention. It is not surprising that Anglo-American trials, historically dominated by juries, are far more dramatic than Continental trials, historically dominated by professional judges.

The voting and spectatorship analogies to judicial decision-making are similar. This is most easily seen by comparing applause to voting—for in a large audience the clapping of a single spectator contributes little more to the overall decibel level than a single vote in an election contributes to the outcome. The voter is the spectator[37] in a contest between candidates, much as the reader or viewer of *Antigone* is the spectator of a contest between Antigone and Creon. It is no surprise that voter turnout is higher, the more publicized and the closer an election is,[38] just as the audience for a heavily advertised, highly dramatic play is likely to be larger than the audience for a meagerly advertised, undramatic play.

Why has the spectatorship analogy to judging been overlooked?[39] One reason is the piety in which the public discussion of judges is

---

[37] Brennan and Lomasky, *The Impartial Spectator* at 189 (cited in note 26).

[38] John H. Aldrich, *Rational Choice and Turnout*, 37 Am J Pol Sci 246, 266–68 (1993); Gary W. Cox and Michael C. Munger, *Closeness, Expenditures, and Turnout in the 1982 U.S. House Elections*, 83 Am Pol Sci Rev 217 (1989).

[39] Well, not completely. See Yosal Rogat, *The Judge as Spectator*, 31 U Chi L Rev 213 (1964). Rogat's article is a criticism of Justice Holmes for having been *too* detached, too much an observer of rather than a participant in life

usually clothed. The analogy seems to give judging a frivolous air. But serious engagement with the arts as reader or viewer is not a frivolous activity; nor is "play" (contrasted with work) incompatible with adherence to rules. A chess player would reduce rather than enhance the pleasure he received from playing the game if he violated its rules, and so would the theatergoer who refused to enter into the lives of the characters on the stage, on the ground that they were not real people; and likewise the judge who violates the rules of the judicial game. Sports fans, theater fans, movie fans, and opera fans often develop a degree of connoisseurship which enhances their pleasure. In other words, they learn the rules (broadly understood) of the game they are watching, and respond in accordance with those rules. It is the same with judges, but with the important difference that some of the rules of the judicial game are uncertain and contested.

A second reason why the spectatorship analogy to judging has been overlooked is the domination of analyses of judicial behavior by legal academics. The academic is a spectator too, but he is a spectator not of the little drama that the judge witnesses—the trial or other contest that the judge resolves—but of the judge's opinion. He usually does not attend oral argument or even read the briefs in the cases that he discusses. Naturally, therefore, he tends to ascribe more importance to the opinion, to its reasoning, its rhetoric, etc. than to the decision itself. Yet these are secondary factors for most judges. For the judge, as for Hamlet, "the play's the thing."[40] When judges got busy, the first thing to be delegated was opinion writing; yet even today it would be considered a scandal if judges delegated the hearing of testimony or argument (though there is in fact some delegation of these functions to magistrates and masters).

The analogy to spectatorship can help us see how judicial outcomes reflect both the judges' preferences going in *and* the quality of the briefing and argument in particular cases. It can also help us understand the function of confirmation hearings in enabling legislators to

---

(like his friends Henry Adams and Henry James)—in short an insufficiently empathetic spectator.

[40] Compare Joseph Bensman and Robert Lilienfeld, *Craft and Consciousness: Occupational Technique and the Development of World Images* 19–22 (Wiley, 1973).

ascertain a judicial candidate's policy preferences, since those preferences can be expected to guide or at least influence a judge's decisions. We might also expect that "ideologues" would be appointed to judgeships at an earlier age on average than other candidates. Not only may it be difficult to determine the trajectory of the nonidealogue's views save by a process of inference from behavior over a long career, but to the extent that an ideologue is inherently more predictable there is less worry that if appointed young he will have a long time to change his views.[41]

Contrary to appearances, this analysis does not justify complacency about judicial performance. To eliminate by means of rules governing conflicts of interest all personal stakes from the judge's decision-making increases the weight (by reducing the cost) of ethical considerations, including the ethical duty to follow legal rules, but at the same time, as with the nonprofit enterprise, it reduces the penalty for careless, erratic, inattentive, or willful decision-making. The problem is deeper. We can see it by returning to the analogy between political and judicial voting, and by examining more critically than before the analogy between judicial and theatrical spectatorship. Most political campaigners appeal primarily, though not exclusively, to the voter's self-interest.[42] The judge, in contrast, like the theatrical spectator, is asked to cast a disinterested vote. It is easy to see why the spectator's vote is likely to be disinterested; what has he—being powerless—to gain by refusing to play the spectator's game? But a judge has some power. Supposing that the conflict of interest rules are effective in insulating his decision from any consequences for his personal or family wealth, one can still imagine a host of inappropriate considerations that might enter into his utility function: personal dislike of a lawyer or litigant in the case, gratitude to the appointing authorities, desire for advancement, irritation with or even a desire to undermine a judicial

---

[41] Compare John R. Lott, Jr. and W. Robert Reed, *Shirking and Sorting in a Political Market with Finite-Lived Politicians*, 61 Public Choice 75, 87–88, 91 n 26 (1989).

[42] Brennan and Lomasky, *The Impartial Spectator* at 189 (cited in note 26), emphasize the element of disinterest in political voting. They follow Adam Smith's *Theory of Moral Sentiments* in arguing that disinterested actions are likely where the cost of such actions to the actor is low.

colleague or subordinate, willingness to trade votes, desire to be on good terms with colleagues, not wanting to disagree with people one likes or respects, fear for personal safety, fear of ridicule, reluctance to offend one's spouse or close friends, and racial or class solidarity. These are common factors in the decisions of everyday life—why not in the decisions of judges, unless we ascribe to them a utility function different from that of the ordinary person, which would be inconsistent with treating them as ordinary persons?

Such factors do influence judicial decisions, but less often than the suspicious layman might suppose. The reason is not that judges have different utility functions from other people but that the utility they derive from judging would be reduced by more than they would gain from giving way to the temptations that I have listed. It is the same reason that many people do not cheat at games even when they are sure they can get away with cheating. The pleasure of judging is bound up with compliance with certain self-limiting rules that define the "game" of judging. It is a source of satisfaction to a judge to vote *for* the litigant who irritates him, the lawyer who fails to exhibit proper deference to him, the side that represents a different social class from his own; for it is by doing such things that you know that you are playing the judge role, not some other role, and judges for the most part are people who want to be—judges. This is consistent with most judges' not wanting to work too hard, for working as hard as a lawyer in private practice is not one of the rules of the judicial game. It is also consistent with judges' often voting their policy preferences and strong personal convictions. For in our system the line between law and policy, the judging game and the legislating game, is blurred. Many cases cannot be decided by reasoning from conventional legal materials. Such cases require the judge to exercise a legislative judgment, although a more confined one than a "real" legislator would be authorized to exercise.[43]

The analogy to playing games is worth pursuing a bit. Rules are not always irksome restraints. They are often constitutive. It is difficult to write a sonnet, because the sonnet is a genre with rigid rules; but without the rules, there would be no sonnets, and this would be a loss

[43] The legislative function of the judge is one reason that a judge is not permitted to use random methods, such as coin flipping, for deciding even cases that are truly indeterminate from a conventionally "legal" standpoint.

not only for the reader but for the sonneteer. And similarly with a game,[44] for example chess. If you decided that your bishops should be allowed to make the same moves as your queen, or that some of your pieces should be allowed to make moves off the chessboard, you would no longer be playing the game of chess. It is true that people sometimes cheat at games when they think they can get away with it. But that is because the pleasure of the game is not the only argument in their utility function. A person might cheat at tennis (this is very common in fact) because he saw an advantage from winning, but if at all reflective he would realize that his pleasure from playing the game itself was diminished, that he was trading off that pleasure against another source of utility. The judicial "game" has rules that lawyers learn in law school and then in practice or teaching, and both self-selection and the careful screening of federal judicial candidates help to assure that most lawyers who become federal judges will be lawyers who enjoy this particular game. They are therefore likely to adhere, more or less, to the rules limiting the materials and considerations that enter into their decisions.[45] The rules, it should be noted, are not the rules of substantive law, to which the community is subject but to which judges in their judicial capacity relate differently, as law givers and law appliers; they are the institutional rules of judging, to which only judges are subject. These rules, as I have said, are not altogether clear or uniform, especially in application; they are probably less clear and uniform, at least

---

[44] For pertinent discussions of games, see Bernard Suits, *What Is a Game?,* 34 Philosophy of Science 148 (1967); Arthur Allen Leff, *Law and,* 87 Yale L J 989, 998–1003 (1978); Johan Huizinga, *Homo Ludens: A Study of the Play-Element in Culture* chapters 1, 4 (Roy, 1950); and especially Thomas Morawetz, *The Concept of a Practice,* 24 Philosophical Studies 209 (1973). I emphasize that the notion of "game" that I am using here is unrelated to that used in game theory. Game theory is concerned with strategic behavior; I am not. For my purposes, solitaire is a good example of a game, but it is not a game in game-theory terms because it does not involve an interaction between persons.

[45] This is related to Wittgenstein's important point that rules bind because they are accepted—rather than being accepted because they bind. Nothing *in* a rule imposes an obligation to follow it. The decision to obey comes from outside, from force or socialization or, I am arguing, from the fact that the rule is constitutive of a pleasurable activity. Compare Sabina Lovibond, *Realism and Imagination in Ethics* 55–57 (U Minnesota, 1983).

in the United States, than the substantive law. Some judges play by "activist" rules, more by rules of "restraint" because those rules are more congenial to the legal profession's self-image; and judges, like game players, sometimes bend or break the rules for the sake of other values, such violations being in fact rather common because detection and sanctioning are difficult. Nevertheless most judicial decisions do have a "ruled" quality, and the analogy to games helps to show how this property of decisions is consistent with utility maximization and how therefore it does not presuppose—what would be contrary to the assumptions of my analysis—heroic self-abnegation on the part of the judges.

Another aspect of rules makes it reasonable to expect most judges to abide by the rules of the judicial game. In creating games, as in creating art, people create a reality in which they can find temporary refuge from, by imaginative transformation of, the sinister realities of ordinary life, the realities of hatred, disease, crime, betrayal, war and so forth. The judicial game has aspects of this refuge and transformation. Its raw materials are the ugly realities of life, but they are transformed in the judicial game to intellectual disputes over rights and duties, claims and proofs, presumptions and rebuttals, jurisdiction and competencies. And that is a comfort; it spares the judge who inflicts or upholds the death penalty from having to think of himself as a killer. But to get this comfort the judge must play by the rules of the judicial game, because the rules constitute the game.

Elected judges play the judge game too, and legislators play a related game called statesmanship or public service. But unlike life-tenured federal judges, these players face higher costs (and obtain no greater benefits) from abiding by the rules of the game and therefore break them more often. Not always, which is why many and indeed most decisions by nonelected judges have a ruled quality and why much legislation has a genuine public-spirited character—not necessarily because the voters are public spirited but because legislators derive satisfactions from acting in the public interest that may outweigh the costs when those costs are small.

[redacted]

VII. Conclusion

Judges are rational, and they pursue instrumental and consumption goals of the same general kind and in the same general way that private persons do. That has been the main thesis of this article. The article does not explain the content of judicial doctrines, and it may seem to undermine the theory that the common law and other areas of

judge-made law are on the whole efficiency-enhancing[59] by disparaging the ability and the will of judges to conceive and impose so singular and ambitious a vision. Actually it supports that theory by showing that the conditions of judicial employment enable and induce judges to vote their personal convictions and policy preferences—or in a word their values.

Reference to the legislative function of the judiciary and to the rules of the judicial game further contributes to dissolving the tension between the analysis of the judicial utility function in this article and the efficiency theory of the common law. When a really new case arises, the rules of the judicial game require the judge to act the part of a legislator and therefore vote his values, although the rules do not require and may even forbid him to acknowledge that this is what he is doing. Efficiency – not necessarily by that name – is an important social value that judges, given their limited remedial powers and the value pluralism of our society, can promote in a reasonably effective, consistent, objective, and hassle-free fashion. One therefore expects it to be influential in judicial decisionmaking when judges are called upon to exercise a legislative function. The decision of a really new case establishes a precedent for future guidance. At this point the rules of the judicial game require the judges to follow the precedent rather than to decide each future case from the ground up. Hence the subsequent cases may not bear any visible imprint of economic thinking, but they will be efficient decisions if the precedent, reflecting the judges' legislative activity, was based implicitly or explicitly on a desire to enhance efficiency. Law can thus be efficient even if only a small minority of cases concern themselves with efficiency. Since the judges have a larger legislative role in common law than statutory cases, we expect common law rules to be on average more efficient than statutory ones.

But this is a side issue. The value of the approach taken in this article is that it demystifies judges at the same time that it domesticates them for economic analysis, and that it both explains the principal features of judicial behavior and generates a number of empirically testable predictions. I have focused on the federal judiciary but the approach can and should be extended to elected judges, to Continental

[59] See, for example, Posner, *Economic Analysis of Law* part 2 (cited in note 13). But see Koyayashi and Lott, *Judicial Reputation* (cited in note 9).

European judges, to jurors, and to legislators. For example, my analysis predicts that judges elected for a term work harder than appointed judges and also that they are less inclined to follow precedent because they are more responsive to the current balance of political power, hence less "independent." Through the testing of these predictions and those suggested earlier in the article, we can move toward the goal of a theoretically sophisticated and empirically validated understanding of the basic institution of our legal system.

THE SELECTION OF DISPUTES FOR LITIGATION

[redacted]

I. The Problem and Some Earlier Attempted Solutions

This paper addresses the relationship between litigated disputes and disputes settled before or during litigation. The specification of this relationship is important for the analysis both of the legal system and of the influence of the legal system on society. Virtually all systematic knowledge of the legal system derives from studies of appellate cases. Appellate cases, of course, provide the most direct view of doctrinal developments in the law. Few scholars today, however, are content to study doctrinal developments alone without regard to the broader influence of legal rules on social affairs. Appellate cases may tell us which disputes courts find troublesome and which they find easy to decide. But this doctrinal information discloses very little about how legal rules affect the behavior of those subject to them or affect the generation of legal disputes themselves.

If all legal disputes, or even a random sample of these disputes, were tried to judgment and then appealed, the inference from legal rules to social behavior would be straightforward. The facts of appellate cases

*[redacted]

[*Journal of Legal Studies*, vol. XIII (January 1984)]
 0047-2530/84/1301-0008$01.50

would indicate all the types of social problems requiring legal resolution. Differences over time in the types of problems appealed would show how changes in legal rules affect these problems and thus how rules influence behavior. It is well known, however, that only a very small fraction of disputes comes to trial and an even smaller fraction is appealed. In a study of insurance company claims files, H. Laurence Ross reports that, of his sample, only 4.2 percent of claims ultimately reached trial and 0.2 percent of claims were appealed.[1] In a more comprehensive survey of police automobile accident reports, Alfred Conard *et al.* find that 0.7 percent of accident victims press their claims to trial and only 0.09 percent of victims appeal trial verdicts.[2] It is very difficult to infer specific characteristics from observations of 0.2 percent or less of a population, especially where there is no evidence that the observations (the disputes selected for appeal) were selected randomly.

Many legal scholars have expressed concern about the peculiar sample of cases that reach trial and appeal, but none has developed an accepted means to adjust analysis of appellate data in response to the problem. Karl Llewellyn, for example, regarded litigated cases as "pathological": bearing the same relation to the broader set of disputes "as does homicidal mania or sleeping sickness, to our normal life."[3] Llewellyn attempted to obtain a broader view of the legal system from his studies of appellate decisions by disregarding leading decisions and, instead, sampling in a manner that approaches randomness. Llewellyn studied "the cases in sequence as they stand in the reports" (that is, for example, the first 94 pages of New York reports from 1842)[4] or decisions rendered by a court on a "single opinion-day" (that is, for example, all decisions rendered by the Pennsylvania Supreme Court on March 20, 1944).[5] These decisions represented to Llewellyn the "mine-run stuff as it comes unselected from the mine."[6] Llewellyn's methodological innovations, however, have not appeared attractive to modern legal scholars. His approach ignores completely doctrinal innovations of any individual court (thus, Llewellyn could address only the process and not the substance of judicial decision making). And few accept his implicit assumption that legal doc-

[1] H. Laurence Ross, Settled out of Court: The Social Process of Insurance Claims Adjustments 216 (1970).

[2] Derived from Alfred F. Conard, James N. Morgan, Robert W. Pratt, Jr., Charles E. Voltz, & Robert L. Bombaugh, Automobile Accident Costs and Payments: Studies in the Economics of Injury Reparation 155, 241 (1964) (hereinafter cited as Conard).

[3] K. N. Llewellyn, The Bramble Bush: On Our Law and Its Study 58 (1960).

[4] Karl N. Llewellyn, The Common Law Tradition: Deciding Appeals 6, 64–68 (1960).

[5] *Id.* at 92–96.

[6] *Id.* at 6.

trine is meaningless. Yet Llewellyn cannot transcend the prospect of the legal system provided by the appellate bench.

In a more recent effort, William Whitford studied the impact that the adoption of the strict liability standard has had on automobile manufacturers and consumers. Reaching beyond leading cases, Whitford studied every reported decision in the period 1960–67 in which auto defects were involved (including cases involving only procedural issues).[7] Whitford, however, despaired of the representativeness of even a census of decisions,[8] which he supplemented with interviews with consumers, employees of dealers and manufacturers, and attorneys involved in auto litigation.[9]

Most legal scholars, however, either ignore the problem of the representativeness of appellate decisions or presume representativeness. The most common presumption is that the facts of disputes that reach trial or, more commonly, appeal resemble the facts of disputes that are settled. Richard Posner, for example, inferred the efficiency of nineteenth-century negligence law from the observation that there were no cases within his large appellate sample in which parties to a contract agreed to a standard of liability different from the legal standard.[10] Posner's conclusion requires the presumption that there are no cases involving alternative liability standards that were settled prior to appeal.[11]

Similarly, it is very common to infer the influence of a legal standard or the attitudes of judges or juries toward plaintiffs or defendants by observing the proportion of cases in which plaintiffs recover verdicts.[12] In the classic *The American Jury,* for example, Kalven and Zeisel purport to

[7] William C. Whitford, Strict Products Liability and the Automobile Industry: Much Ado about Nothing, 1968 Wis. L. Rev. 83, 102–3. For other surveys of a universe of appellate decisions, see Richard A. Posner, A Theory of Negligence, 1 J. Legal Stud. 29 (1972); and George L. Priest, Breach and Remedy for the Tender of Nonconforming Goods under the Uniform Commercial Code: An Economic Approach, 91 Harv. L. Rev. 960 (1978).

[8] Whitford, *supra* note 7, at 103–4.

[9] *Id.*

[10] Posner, *supra* note 7, at 74.

[11] There may have been no litigated cases because it was clear to the parties that courts would give effect to any voluntary agreement, because the alternative standards drafted by the parties were unambiguous, or because no party found appeal of a trial court decision involving such an agreement worthwhile. See, for evidence of voluntary agreements, Richard A. Epstein, The Historical Origins and Economic Structure of Workers' Compensation Law, 16 Ga. L. Rev. 775, 787–97 (1982).

[12] See for example, Jerome Frank, Law and the Modern Mind 112 (1930). Indeed, in an earlier paper, one of us concluded that case-law interpreting the Uniform Commercial code is consistent with an efficiency standard from the observation that the rate of plaintiffs' victories on appeal seemed responsive to differences in the costs of detecting breach of contract. Priest, *supra* note 7, at 997–98.

measure the extent of differences between jury and bench trials by comparing respective rates of plaintiff victories in civil cases and respective rates of conviction in criminal cases.[13] Kalven concludes that the finding that plaintiffs recover verdicts in 55 percent of personal injury cases tried to juries proves that juries are not "monolithically pro-plaintiff," but rather tend to follow the equities of the individual case.[14] The Supreme Court has inferred the procedural fairness of a disability benefits system from a near 50 percent rate of reversals of benefit denials.[15] The presumption underlying these inferences is identical. For the rate of plaintiff verdicts to be an accurate measure of the influence of a legal standard, of judicial or jury attitudes, or of the substantive fairness of any adjudicatory process, litigated disputes must be representative of the entire class of underlying disputes.

This paper presents a model of the litigation process that clarifies the relationship between the set of disputes settled and the set litigated. According to our model, the determinants of settlement and litigation are solely economic, including the expected costs to parties of favorable or adverse decisions, the information that parties possess about the likelihood of success at trial, and the direct costs of litigation and settlement. The most important assumption of the model is that potential litigants form rational estimates of the likely decision, whether it is based on applicable legal precedent or judicial or jury bias. From this proposition, the model shows that the disputes selected for litigation (as opposed to settlement) will constitute neither a random nor a representative sample of the set of all disputes.

In particular, the standard of decision will influence the observed rate of success of plaintiffs and defendants in litigation only partially. If disputes were selected for litigation randomly, then the rate of plaintiff verdicts would differ as the standard of decision becomes more or less favorable to plaintiffs.[16] Our model, however, demonstrates that, where the

[13] Harry Kalven, Jr., & Hans Zeisel, The American Jury, passim (1966), discussed *infra*.

[14] Harry Kalven, Jr., The Dignity of the Civil Jury, 50 Va. L. Rev. 1055, 1072 (1964).

[15] Richardson v. Perales, 402 U.S. 389, 410 (1971). But see Mathews v. Eldridge, 424 U.S. 319, 346 (1976) ("Bare statistics rarely provide a satisfactory measure of the fairness of a decision-making process," although accepting general inference). The Supreme Court has also concluded that the standards applied by six- and twelve-person juries are substantively indistinguishable from observing comparable rates of plaintiff success. Colgrove v. Battin 413 U.S. 149, 159–60 & n. 15 (1973), discussed in text at notes 100–115 *infra*.

[16] The selection question has not been considered explicitly in earlier treatments of the litigation-settlement decision. However, the approach of Landes and Posner, who developed the dominant model of litigation (from which our model derives), suggests that disputes will be selected randomly. See William M. Landes, An Economic Analysis of the Courts, 14 J. Law & Econ. 61 (1971), and Richard A. Posner, An Economic Approach to Legal Proce-

gains or losses from litigation are equal to the parties, the individual maximizing decisions of the parties will create a strong bias toward a rate of success for plaintiffs at trial or appellants at appeal of 50 percent regardless of the substantive standard of law.[17] Thus, plaintiff victories will tend toward 50 percent whether the legal standard is negligence or strict liability, whether judges or juries are hostile or sympathetic. The model also specifies the economic conditions that lead plaintiffs' success rates to differ from 50 percent.

Section II of the paper first presents the basic model and derives the 50 percent implication. Then, an important assumption of the basic model is relaxed to consider the set of cases in which the stakes of the dispute are different for the parties, as, for example, where a single consumer sues a manufacturer whose future practices will be affected by the outcome. Section II shows that the relative stakes to the parties will greatly influence the rate of success in litigation and are likely to be the principal reason why success rates differ from the 50 percent baseline. Finally, Section III presents data bearing on the model's implications, data deriving from our own empirical investigations and from the major empirical studies of the legal system of the past fifty years.

This article builds on an insight of earlier papers by Priest that defined the selection hypothesis and discussed the influence of private decisions

---

dure and Judicial Administration, 2 J. Legal Stud. 399 (1973). According to the Landes-Posner model, the principal determinant of litigation is the absolute difference between the parties' estimates of the likelihood of success. Thus, the parties will litigate if their estimates differ by, say, 5 percent, where 5 percent times the stakes in the case exceeds the difference between joint litigation and settlement costs. Obviously, a 5 percent difference in estimates is just as likely to generate litigation, whether the plaintiff and defendant estimate the plaintiff's chances, respectively, as .95 and .90 or .10 and .05. Posner and Landes assume—reasonably—that differences in the parties' estimates will be distributed randomly. *Id.* at 423–26. It follows that cases litigated will represent a random sample of the set of underlying disputes. Thus, if the underlying set of disputes is such that if every dispute were litigated, plaintiffs would win 30 percent of the time, the Landes-Posner approach predicts that a 30 percent success rate will be observed at trial. This conclusion is made explicit in Posner, *supra* note 7, at 92. For the contrasting predictions of our model under these conditions, see Table 2 *infra*.

[17] In a recent paper, Baxter independently intuits that plaintiffs will win approximately 50 percent of the time. The Political Economy of Antitrust: Principal Paper by William Baxter 16 (Robert M. Tollison ed. 1981) (hereinafter cited as Baxter). Baxter's hypothesis derives from an assumption similar to that of Posner and Landes, *supra* note 16, that disputes, in general, will be litigated randomly, coupled with Baxter's personal assumption that the standard of decision is defined such that plaintiffs would win 50 percent of the time if all disputes were litigated. See Baxter, *supra* at 16 (explaining 50 percent result) and *id.* at 18 (rejecting hypothesis that the true decision standard differs from 50 percent). Since currently there is no basis for believing that the decision standard bisects the distribution of disputes (see Figure 1, *infra*), Baxter's 50 percent conclusion is insupportable. Baxter, however, describes the influence of different stakes to the parties in a manner very similar to ours. See text at notes 120–29, *infra*. Some of Baxter's statistics are reinterpreted *infra*, at note 129.

to settle or litigate on the formal content of the law.[18] One of these papers demonstrated that where litigation costs are substantially greater than settlement costs, the bias toward 50 percent plaintiff victories makes the formal structure of the law appear indeterminate to any scientific, empirical method of observing judicial decisions.[19] This article refines that implication, but it goes further. First, it defines with substantially greater precision and completeness the determinants of the selection effect. Second, it derives empirical implications of the selection hypothesis that are less qualitative and hence more easily observable, such as success rates in litigated cases, addressed in Section III. Finally, it defines empirical relationships between litigated and settled cases; litigation costs, settlement costs, and expected judgments; sets of disputes and legal standards; and changes in the parties' information about the law over time and across legal subject areas—relationships which have been only dimly perceived in the past. We hope, therefore, that the article will provide the basis for a wide and diverse range of new empirical investigations into the operation of the legal system.

II. The Selection Model and Its Implications

A. *Preliminary Assumptions*

Disputes and Verdicts. We define a "dispute" as any occasion in which a plaintiff asserts a claim for some injury against a defendant.[20] A dispute may be resolved either by a verdict after trial or by a settlement at any time prior to a verdict. For our purposes, a dispute will be regarded as "litigated" only if a verdict is rendered; all terminations of the dispute short of a verdict are regarded as "settlements."[21] All trial verdicts constitute the relevant set of disputes for appeal. We presume that the deci-

[18] George L. Priest, The Common Law Process and the Selection of Efficient Rules, 6 J. Legal Stud. 65 (1977), and Selective Characteristics of Litigation, 9 J. Legal Stud. 399 (1980) (hereinafter cited as Priest, Selective Characteristics).

[19] Priest, Selective Characteristics, *supra* note 18, at 410.

[20] The model can also incorporate claims for equitable relief (potential injuries).

[21] Our definition is chosen to correspond with available data. Though it is more common to regard a dispute as "litigated" if trial has begun—even if settled prior to a verdict—we know of no data with respect to cases settled during trial. Our model could be extended to consider settlement during trial—or at other points between the injury and verdict—if assumptions are made concerning the sequential acquisition during these periods of information relating to the outcome. We have no firm basis for such assumptions (however, compare the suggestive discussion in Baxter, *supra* note 17, at 14–15) and thus consider as "litigated" disputes never settled.

Dependent No More: It's been a hundred years since the United States granted citizenship to American Indians. Will it take a hundred more before it frees tribes to make their own decisions?

Date: Spring 2024
From: Hoover Digest(Issue 2)
Publisher: Hoover Digest
Document Type: Article
Length: 2,978 words
Content Level: (Level 5)
Lexile Measure: 1420L

Full Text:
In 1924, Congress passed the Indian Citizenship Act (ICA), declaring that "all non-citizen Indians born within the territorial limits of the United States be, and they are hereby, declared to be citizens of the United States." President Coolidge signed it into law on June 2, 1924. The act, partially inspired by robust Native American enlistment during the First World War, is celebrated as a step toward honoring American Indians--regardless of whether they abandoned their tribal affiliations--by welcoming them to US citizenship, giving them the right to vote. There is, however, dissonance in "declaring" people as citizens who were the continent's first inhabitants and who were already citizens of their own nations--Iroquois, Comanche, Osage, Sioux, Crow, and so on. Those nations long had rules of law, boundaries, and rituals for choosing leaders and adopting others into their tribes. Because US citizenship brought with it subjugation to federal laws, accepting it was controversial among some tribal leaders, who foresaw that it would create a morass of legal questions about where and to whom tribal laws would still apply.

The ICA came a century after the US Supreme Court declared in Cherokee Nation v. Georgia (1831) that tribes were "domestic dependent nations," making "their relation to the United States" resemble "that of a ward to his guardian." Before 1924, the "wards" were usually not US citizens until a Bureau of Indian Affairs (BIA) agent of the federal government deemed them to be "competent and capable." In practice, this generally meant becoming farmers and assimilating into the white population.

This was part of a more general effort to "detribalize" Native Americans. As trustee for Indian wards, the federal government, to this day, oversees land use and title transfer of millions of Indian reservation acres. The US citizenship announced in 1924 did not terminate this ward-to-guardian relationship. On the contrary, these terms remain part of modern federal Indian law under the Burke Act of 1906, which requires that the government assess the competence of individual Indians before giving them fee-simple patent to their allotted land.

These laws help explain the theme of the book and recent movie Killers of the Flower Moon. Because Indian "headrights" to subsurface oil in Oklahoma were held in trust during this 1920s episode, revenues from leases were held by the Department of the Interior and could be released to tribal members only at the discretion of the department. Hence, Golden Globe winner Lily Gladstone, playing the part of an Osage Tribe citizen, must grovel before her Indian agent, stating her allotment number and declaring, "I'm Mollie Kyle, incompetent."

Can a United States citizen simultaneously be a ward and a free American?

POWER IN AUTONOMY

On the 100th anniversary of the Indian Citizenship Act, it is appropriate to reflect that Native Americans held a status before the conquest by Europeans that allowed them to be, in the words of Crow tribal citizen Bill Yel-lowtail, "strong, self-sufficient, self-initiating, entrepreneurial, independent, healthful, and therefore powerful individual persons. Human beings, Indians." Yet the guardian role assumed by the federal government too often has limited the benefits of both tribal citizenship and US citizenship, leaving Native Americans to operate in a no-man's land of uncertain and limited privileges and rights.

Citizenship requires two things: territorial control by a government, and a legal (written or unwritten) relationship between an individual and the state that specifies rights and duties for both individuals and the state. By these criteria, most Native Americans were citizens of nations before they were declared to be citizens of the United States. Whether it is a matter of enforcing individual ownership claims, protecting territorial borders, or producing public goods such as trails, meeting halls, or irrigation systems, tribes were organized to make and enforce laws and to produce collective goods. In this sense, tribes were nations and members were citizens.

Consider the League of Five Nations, or Iroquois League, as the Europeans called it. The league was a loose alliance among the Oneida, Mohawk, Seneca, Cayuga, and Onondaga tribes, governed by the Great Law of Peace. The Iroquois League was governed by a constitution that influenced language in the US Constitution. Its constitution limited the powers of the collective and enforced the rights of the subgroups and individual citizens. By any definition, the organization of the Iroquois League under the Great Law of Peace constituted a nation, and its individual members were citizens of that nation.

In Lakota America: A New History of Indigenous Power, Pekka Hamalainen describes how the Lakota Nation evolved from an upper Great Lakes hunter-gatherer tribe to one of the most powerful Indian nations west of the Mississippi. The Lakota took control of trade in the Missouri Valley and charged tolls to the trappers and traders, as well as to Lewis and Clark's Corps of Discovery, moving furs and trade goods up and down the river. Where buffalo were plentiful, they excluded other tribes from hunting, and they had individual and clan property rights to cultivated land from which crops such as corn added to their diet.

All of this required complicated governance structures, at levels ranging from the nation to the band to the citizen. The nation was not "a formal state or confederacy" but instead "a manifestation of deep voluntary attachments that bound the seven fires [tribes] together...from the bottom up, with language and kinship as the main cohesive," Hamalainen writes. At the lower end were smaller groups in which "individuals, families, and bands moved around constantly, arranging themselves into different constellations as circumstances demanded... creating a thick lattice of kinship ties that transcended local and regional identities." The bottom-up structure limited the power of the chiefs and councils, but they did have the power to organize armies, with the consent of the subgroups, to defend or expand their territories. The Lakota enforced rules on the inside and protected citizens from invaders from outside. The rules within the tribe provided incentives for individuals, families, and clans to make investments in their personal wealth that allowed them to thrive in new territories and to trade with other tribes when gains from trade were available. Furthermore, those investments, especially in equestrian skills, provided the human and physical (e.g., horses) capital that could be called on by war leaders to protect tribal territories and acquire new lands.

These are but two examples of Native American nations and the citizenship of those nations. European conquest overtook tribal territories and subjected American Indians to new rules over which they had little control.

"DEAD CAPITAL"

Once Indians were declared to be wards of the federal government, they became more like colonial citizens. The nature of their tribal governments, the laws to which they must abide, the structure of their property rights, and even their racial identity were mostly determined by a bureaucracy unaccountable to them. The legacy consists of today's policies regulating everything from health care to education to reservation land use that are still manifest in federal agencies such as the BIA, the Bureau of Indian Education, and the Indian Health Service.

Bureaucracy went so far as to tie Native American citizenship to ethnicity by inventing a pseudoscientific "blood quantum" system of enumeration that persists to this day. Blood quantum was determined by federal Indian agents who tracked the fraction of ancestors documented as full-blooded Indian in the nineteenth and twentieth centuries, sometimes just by looking at the person. These assessments determined eligibility for federal payments under treaties, land cessions, and litigation settlements. Before the Indian Citizenship Act of 1924, low blood quantum was the implicit criterion for US citizenship and high blood quantum for federal payments and wardship. In a nation of immigrants where the Fourteenth Amendment acknowledges citizenship as automatic by place of birth, for Native Americans that status depended on dubious assessments of ancestry.

This federal bureaucracy wields jurisdiction over what is known as Indian Country. It is defined by the US federal criminal code (18 US Code [section] 1151): land within the boundaries of an Indian reservation, including rights of way through a reservation; dependent Indian communities within the borders of the United States; and all individual Indian allotments, including rights of way, and excluding allotments for which Indian title has been extinguished. Real property in Indian Country may be owned in fee simple by either Indians or non-Indians, but sixty-six million acres in Indian Country are held in trust by the Bureau of Indian Affairs, meaning the bureau must approve how the land is used and must monitor the distribution of revenues generated from the land and minerals. In essence, the land is owned by the federal government. American Indians are the only US citizens with such a preponderance of their assets held in trust.

Despite the large acreage held in trust, the lands have yielded a pittance of monetary return when compared to land and natural resources owned by non-Indian American citizens. For example, Indian reservations have abundant energy resources, but have profited little from these supplies. Reservations contain almost 30 percent of the nation's coal reserves west of the Mississippi, 50 percent of potential uranium, and 20 percent of known oil and gas reserves. They also contain almost 10 percent of the nation's wind and solar energy potential and a large stock of critical minerals. Ye t Senate committee hearings have concluded that only two million of fifteen million acres of energy resources were developed on reservations, and that $1.5 trillion worth of subsurface reserves remain untapped. Commercial renewable-energy production is also lacking, with only a handful of tribes capitalizing on the national momentum for sources of alternative energy.

Why do reservation resources held in trust often fail to generate wealth for the reservation? Some tribes choose not to develop their

natural resources. Others wish to develop, but stifling federal bureaucracy stands in the way. As Ernest Sickey, the late chairman of the Coushatta Tribe, put it, federal bureaucracy has strangled American Indian enterprise with "white tape."

The white tape traces back to Cherokee Nation v. Georgia, the 1831 Supreme Court decision, which was not undone by the Indian Citizenship Act. Instead, that court's assertion that Indians "are in a state of pupilage" remains in force as trusteeship continues to govern natural-resources use.

The result is an unparalleled regulatory burden on Native American citizens. As the Environmental Protection Agency notes, "Activities in Indian country ...often require a greater level of NEPA [National Environmental Policy Act] involvement than the same activities in nontribal areas." This means, for instance, that forty-nine regulatory steps were required to get an oil lease in Indian Country, compared to four steps elsewhere. A similar regulatory morass awaits tribes wanting to develop wind and solar in Indian Country.

Trusteeship also helps explain why tribal members in the water-thirsty West do not benefit from paper water rights valued in the billions annually, and why members of North Dakota's Three Affiliated Tribes missed out on millions in royalties during the 2010s petroleum-fracking boom. Bureaucracy and regulatory rules also help explain why good farmland is often left unused despite price booms in agriculture. These valuable resources are effectively "dead capital," to use the phrase of Peruvian economist Hernando de Soto.

And when land and natural resources are put to use, trustee management of revenue has allowed federal bureaucrats to lose billions of dollars belonging to individual Indians through poor recordkeeping and by brokering leases of Indian land for pittances. This was documented in great detail in the so-called Cobell litigation--named for its lead plaintiff, Blackfeet elder Elouise Cobell--a class-action suit accusing the Departments of the Interior and Treasury of mismanaging Indian trust funds. Despite a settlement of $3.4 billion reached in 2009, Cobell said at the time, "There is little doubt this is significantly less than the full accounting to which individual Indians are entitled." The Indian Citizenship Act did not prevent these losses because it did not undo the guardian's power.

LONG ROADS TO SUCCESS

The Indian Citizenship Act was a step forward in tribal and individual Indian relations with the federal government, but as long as the historic legal relationship remains, it is difficult for tribes to be sovereign governments and Native Americans to be sovereign citizens. Tribes refer to themselves as "sovereign nations," but trusteeship and federal bureaucracy weaken both the sovereignty of Indian nations and the benefits of US citizenship.

Furthermore, unlike other governmental units (states, counties, cities) beneath the federal government, tribal governments have little power to tax and therefore few sources of revenue to produce public services such as police protection, education, health care, and roads. Without revenue sources, they depend on grants to produce those services or on federal agencies to provide them directly.

The Southern Ute Tribe's struggle for economic independence illustrates the legacy of nineteenth-century Indian policy that made tribes "domestic dependent nations," and it also shows a path out of dependency. The Southern Ute Reservation in southwest Colorado sits on massive energy resources, yet for most of the past century the small tribe--now numbering fewer than fifteen hundred members--was impoverished. The prosperous hunters and traders who once roamed the Great Basin were forced onto a reservation spanning about one-third of what would become Colorado, and their strength and wealth quickly dissipated. By 1895, the tribe had been squeezed into a fifteen-by-seventy-five-mile strip consisting of a mosaic of land tenure, including both private and trust property, the latter under the control of the federal government.

It took more than a century, but the Southern Utes persevered, slowly winning court judgments to reclaim their water, land, and mineral rights and the revenue from those sources. Revenue from five tribal energy companies is invested in the Southern Ute Growth Fund, estimated to be worth $4 billion, and dividends from the fund are distributed annually to tribal members.

With profits from oil, gas, and other enterprises, the tribe was able to take control of and manage the reservation's infrastructure. It runs a medical clinic, formerly operated by the federal Indian Health Service. It built a state-of-the-art recreation center and introduced a Ute-language program in its schools. The tribe's Southern Ute Community Action Programs include substance-abuse treatment centers, a senior citizen center, and job-training programs. Oil and gas profits provide scholarships for every tribal member who wants to attend college, dividends for members between twenty-six and fifty-nine, and retirement benefits to those over sixty.

In western North Dakota, the Three Affiliated Tribes--Mandan, Hidatsa, and Arikara--on the Fort Berthold Reservation also managed to gain more control of their oil and gas resources. As a result, the former tribal chairman, Tex Hall, said the tribes were gaining "sovereignty by the barrel"--making clear the links between sovereignty, citizenship, and wealth.

Examples extend beyond oil and gas. The Salish-Kootenai Confederated Tribes on the Flathead Reservation in Northwest Montana took over their forest management under a special agreement with the Bureau of Indian Affairs. Not only has their management made profits for the tribe while neighboring federal forests lost money, but the tribes' forests have outperformed federal forests by virtually all environmental standards.

Sovereign control gives tribes an incentive to maximize economic returns from their land. This is in sharp contrast to federal agencies, for which the incentives "might be not just weak but actually perverse," according to a 1994 article in the Journal of Economic Behavior and Organization, because the BIA's budget tends to grow when it fails to fulfill its federal trust responsibility.

There are risks to achieving economic independence, and not every tribal project succeeds. But the data show that Indigenous self-

governance succeeds over the long run. Research showed that tribes opting out of federal oversight through the Indian Reorganization Act of 1934, for instance, have had 12 to 16 percent greater economic growth when compared to those without self-governance.

A PROFOUND SHIFT

The recent landmark case of McGirt v. Oklahoma (2020) provides an opportunity to redefine the meaning of citizenship for American Indians. The case began over the question of whether the federal or state government had jurisdiction over a case involving Jimcy McGirt, a citizen of the Seminole Tribe, who was accused of sexually assaulting a non-Indian minor. Ultimately, the case was heard by the US Supreme Court, which ruled that criminal cases belong in federal or tribal courts if the crime was committed in "Indian Territory," and it concluded that this land makes up nearly half of the state of Oklahoma.

Some citizens of Oklahoma who are not citizens of the "Five Civilized Tribes"--the Cherokee, Choctaw, Chickasaw, Creek, and Seminole--feared that the new jurisdictional arrangement would stunt the state's economy. Governor Kevin Stitt worried that "no investment is going to come" because the "uncertainty is huge." The fears, however, have not come to pass, because the "competent and capable" Indian nations understand the importance of stable governance rules for tribal, state, and US citizens alike.

The history of tribal governance all the way up to McGirt illustrates that citizenship is only as useful as the rules that govern citizens. Before colonization, Native Americans lived under rules that limited tribal authority, enhanced individual freedom, and allowed members to thrive, not just survive. The Indian Citizenship Act of 1924 should have lessened the federal government's guardian role because laws that consider a class of citizens to be incompetent and incapable clash with citizenship in a free country. As we celebrate the ICA, we should also work to recognize the continent's first inhabitants as free, competent, and capable citizens.

by Terry L. Andersonand Dominic P. Parker

Terry L. Anderson is the John and Jean De Nault Senior Fellow at the Hoover Institution. He is past president of the Property and Environment Research Center (PERC) in Bozeman, Montana, and a professor emeritus at Montana State University. Dominic P. Parker is a senior fellow (adjunct) at the Hoover Institution and a professor of applied economics at the University of Wisconsin-Madison. He and Terry Anderson direct the Hoover Project on Renewing Indigenous Economies.

Source Citation (MLA 9th Edition)
Anderson, Terry L., and Dominic P. Parker. "Dependent No More: It's been a hundred years since the United States granted citizenship to American Indians. Will it take a hundred more before it frees tribes to make their own decisions?" *Hoover Digest*, no. 2, spring 2024, pp. 177+. *Gale In Context: Opposing Viewpoints*, link.gale.com/apps/doc/A793021874/OVIC?u=viva_gmu&sid=bookmark-OVIC&xid=4fe427b3. Accessed 26 Aug. 2024.
Gale Document Number: GALE|A793021874

The Institutions of Governance

[redacted]

The new institutional economics (NIE) is an idea whose time has come. That was evident to R. C. O. Matthews in 1986, who in his presidential address to the Royal Economic Society declared that the NIE was "one of the liveliest areas in our discipline" (1986 p. 903) and thereafter described it as a body of thinking based in two propositions: institutions matter, and institutions are *susceptible to analysis* (1986 p. 903). The proposition that institutions matter is embraced by institutional economists of all kinds, old and new. What distinguishes the NIE from earlier (and some contemporary) work on institutions is that institutions are susceptible to analysis. Older-style institutional economics was content to critique orthodoxy and collapsed for failure to advance a positive research agenda (Ronald Coase, 1984; Matthews, 1986). The NIE has responded to the challenge by (i) developing a comparative institutional logic of organization to which (ii) many applications and refutable implications accrue and in relation to which (iii) many empirical tests have been conducted and are broadly corroborative (Paul Joskow, 1991; Howard Shelanski and Peter Klein, 1995; Bruce Lyons, 1996; Keith Crocker and Scott Masten, 1996).

The institutions of principal interest to the NIE are the institutional environment (or rules of the game—the polity, judiciary, laws of contract and property [Douglas North, 1991]) and the institutions of governance (or play of the game—the use of markets, hybrids, firms, bureaus). Although it has been customary to work at one level at a time, each level informs the other, and recent work of a combined kind has appeared as applications to economic development and reform are attempted. Also, the study of economic institutions needs to make provision for two background conditions: the condition of societal embeddedness (to which sociologists refer; [Mark Granovetter, 1985; Victor Nee, 1997]) and the attributes of human actors (Herbert Simon, 1985; Leda Cosmides and John Tooby, 1994, 1996). I return to these two background conditions in Section II. Suffice it to observe here that the NIE is, by its very nature, an interdisciplinary undertaking.

* [redacted]

I. The Governance of Contractual Relations

Transaction-cost economics is located on the branch of the NIE that is predominantly concerned with governance, the branch that has its origins in Ronald Coase's treatment of firms and markets in his classic 1937 paper on "The Nature of the Firm." Rather than take the organization of economic activity in firms and markets as preexisting, defined largely by technology, Coase described firms and markets as alternative means for doing the very same thing. The allocation of activity as between markets and hierarchies was no longer taken as given, but needed to be derived. Should a firm make or buy? Which transactions go where and why? The firm was reconceptualized for these purposes as a governance structure (which is an organizational construction).

Much of the predictive content of transaction-cost economics works through the discriminating-alignment hypothesis: transactions, which differ in their attributes, are aligned with governance structures, which differ in their cost and competence, so as to effect a (mainly) transaction-cost economizing result. Implementing this requires that transactions, governance structures, and transaction-cost economizing all be described. What are the defining attributes of transactions? What are the attributes with respect to which governance structures differ? What main purposes are served by economic organization? How is transaction-cost economizing accomplished?

According to John R. Commons (1932 p. 4), "the ultimate unit of activity ... must contain in itself the three principles of conflict, mutuality, and order. This unit is a transaction." Transaction-cost economics concurs that the transaction is the basic unit of analysis and regards governance as the means by which order is accomplished in a relation in which potential conflict threatens to undo or upset opportunities to realize mutual gains.

The problem of conflict upon which transaction-cost economics originally focused is that of bilateral dependency (Williamson, 1975; Benjamin Klein et al., 1978). Whereas the organization of transactions that are supported by generic investments is easy (classical market contracting works well because each party can go its own way with minimal cost to the other), potential problems arise when nonredeployable investments are made. Parties that are joined in a condition of bilateral dependency (by reason of asset specificity) and are confronted with contractual incompleteness (by reason of bounds on rationality) must confront strains (by reason of opportunism) when faced with the need to adapt cooperatively. How should such contracts be managed?

A comparative assessment of markets and hierarchies is needed. Taking adaptation to be the central problem of economic organization, of which autonomous and cooperative kinds are distinguished, markets enjoy the advantage in autonomous-adaptation respects, and the advantage shifts to hierarchies as the needs for cooperative adaptation build up. Discrete structural differences between markets and hierarchies in incentive intensity, administrative control, and contract-law regimes are responsible for these adaptive differences (Williamson, 1991).

Implicit in the foregoing and important to the transaction-cost economics enterprise is the assumption that contracts, albeit incomplete, are interpreted in a farsighted manner, according to which economic actors look ahead, perceive potential hazards, and embed transactions in governance structures that have hazard-mitigating purpose and effect. Also, most of the governance action works through private ordering, with courts being reserved for purposes of ultimate appeal. Furthermore, although vertical integration is the archetype transaction out of which transaction-cost economics works, labor, capital, corporate governance, regulation/deregulation, vertical market restrictions, multinational and public-sector transactions are variations on a theme to which numerous public-policy ramifications accrue. Finally, although bilateral dependency (asset specificity) is a widespread condition, the basic transaction-cost setup, with its emphasis on *ex post* governance, applies to contractual hazards more generally.

Additional hazards that transaction-cost economics has begun to address include (i) the hazards of weak property rights (David J. Teece, 1986); (ii) measurement hazards of multiple-task (Bengt Holmstrom and Paul Milgrom, 1991), oversearching (Yoram Barzel, 1982; Roy Kenney and Klein, 1983), and multiple-principal (Avinash Dixit, 1996) kinds; (iii) intertemporal hazards, which can take the form of disequilibrium contracting, real-time responsiveness, long latency, and strategic abuse; (iv) hazards that accrue to weaknesses in the institutional environment (North and Barry Weingast, 1989; Brian Levy and Pablo Spiller, 1994; Weingast, 1995), which are pertinent to economic development and reform, and (v) weaknesses of probity, which are of special concern for what James Q. Wilson (1989) refers to as "sovereign transactions" (see also Williamson [1998]). To be sure, lenses other than transaction-cost economics can and have been brought to bear. It is nonetheless noteworthy that transaction-cost economics relates to all of these hazards in the following four respects: (i) all of these hazards would vanish but for bounds on rationality and opportunism, (ii) the magnitude of the hazards varies systematically with the attributes of transactions, (iii) *ex post* governance (as well as *ex ante* incentive alignment) is an important instrument in effecting hazard mitigation, and (iv) the discriminating-alignment hypothesis applies. Hazard mitigation through the *ex post* governance of incomplete contracts is the general rubric.

Whereas organizational variety was once ascribed principally to monopoly purpose and/or efficient risk-bearing, much of this variety is usefully interpreted in transaction-cost-economizing terms. It matters a lot, as it turns

out, whether firms are described as governance structures rather than production functions.

What came to be known as the "inhospitality tradition" in antitrust was the public-policy consequence of treating the firm as a production function (which is a technical construction). Because a technological justification for nonstandard and unfamiliar contractual and organizational practices was lacking under the neoclassical setup, such practices were presumed to be anticompetitive.

However, that is because the study of economics was needlessly truncated. Technological economies of scale or scope and related "technical or physical aspects" do not exhaust the possibilities. Organizational economies that are attributable to the alignment of governance structures with the attributes of transactions had been ignored or suppressed. Upon working through the logic of discriminating alignment, an altogether different rationale for vertical integration, vertical market restrictions, and other forms of nonstandard contracting took shape.

Many of the hitherto puzzling economic institutions of capitalism were interpreted more constructively as this broader approach to economizing progressed. Thus, although anticompetitive purpose remains a concern if the relevant market-power preconditions are satisfied, those are stringent preconditions. As a consequence, anticompetitive effect is better regarded as the exception rather than the rule.

The concepts of credible commitment and discriminating alignment that inform private-sector governance have also turned out to be instructive for understanding economic development and reform. Viewing the institutional arrangements (rules of the game) through the lens of contract and governance has helped, among other things, to disclose when and why privatization efforts will succeed or fail.

II. Background Conditions

The two background conditions to which I referred earlier are the conditions of embeddedness and the attributes of human actors. The first of these is antecedent to the polity and refers to societal features (norms, customs, mores, religion) which differ among groups and nation states and operate as societal supports, or the lack thereof, for credible contracting. A recurring concern is when and why do reputation-effect mechanisms work well and poorly?

Turning to the attributes of human actors, Simon (1985 p. 303) takes the position that "Nothing is more fundamental in setting our research agenda and informing our research methods than our view of the nature of the human beings whose behavior we are studying." The huge literature on "psychology and economics," as recently summarized by Matthew Rabin (1996), speaks to many of the issues. Of special interest is research stemming from Daniel Kahneman and Amos Tversky (1982), in which the limits and biases of individual decision-makers in making probabilistic choices are revealed. Also pertinent are recent studies by cognitive anthropologists (Edwin Hutchins, 1995) and evolutionary psychologists (Cosmides and Tooby, 1996).

Hutchins's book on *Cognition in the Wild* makes the distinction between cognition "in the laboratory, where cognition is studied in captivity, and the everyday world, where human cognition adapts to its natural surroundings" (1995 p. xiv). Issues of organization are brought in by viewing individuals as "part of a larger computational system" (Hutchins, 1995 p. xv). Because, moreover, each generic form of organization has both strengths and weaknesses, organization needs to be studied as both problem and solution. Cognitive anthropology can help to inform both.

The evolutionary-psychology literature notes that cognitive methods "drawn from logic, mathematics, and probability theory" (Cosmides and Tooby, 1994 p. 319) and cognitive processes that had good survival properties for human agents operating in the Pleistocene sometimes differ. For example, framing problems of choice under uncertainty in frequentist terms, which is the manner in which they arise in the wild, rather than in point estimates, which is the way they are posed by statistical decision theory, is consequential. Cognitive biases and disabilities are often reduced when problems are presented in frequentist terms.

Evidence, moreover, that individuals on average are poor probabilists speaks only to the

mean. If there is variance in the aptitude to make probabilistic choices among members of the population, and if it is cost-effective to screen for differential competence, then an efficient alignment of differential competence to tasks can be effected through specialization (organization). Not only does organization matter in this respect, but it also relieves problems of limited cognitive ability by permitting composite tasks to be broken down into nearly separable parts, each of which is more tractable, thereafter to be recombined (Simon, 1962). The upshot is that the study of individual decision-making needs to be extended to make provision for the efficient deployment of individuals and groups within an organization. Cognitive anthropology, organization theory, and evolutionary psychology are all implicated.

III. Concluding Remarks

The papers and discussions at the inaugural conference of the International Society for New Institutional Economics in September 1997 record that the NIE continues to extend its analytical reach and inform new issues. New conceptual challenges arise as new applications are attempted. A deeper understanding of complex economic organization is in progress.

REFERENCES

Barzel, Yoram. "Measurement Cost and the Organization of Markets." *Journal of Law and Economics*, April 1982, *25*(1), pp. 27–48.

Coase, Ronald H. "The Nature of the Firm." *Economica*, November 1937, *4*(4), pp. 386–405.

__→__. "The New Institutional Economics." *Journal of Institutional and Theoretical Economics*, March 1984, *140*(1), pp. 229–31.

Commons, John R. "The Problem of Correlating Law, Economics, and Ethics." *Wisconsin Law Review*, December 1932, *8*(1), pp. 3–26.

Cosmides, Leda and Tooby, John. "Better than Rational: Evolutionary Psychology and the Invisible Hand." *American Economic Review*, May 1994 (*Papers and Proceedings*), *84*(2), pp. 327–32.

→ _____. "Are Humans Good Intuitive Statisticians After All?" *Cognition*, January 1996, *58*(1), pp. 1–73.

→ **Crocker, Keith and Masten, Scott.** "Regulation and Administered Contracts Revisited: Lessons from Transaction-Cost Economics for Public Utility Regulation." *Journal of Regulatory Economics*, January 1996, *9*(1), pp. 5–39.

Dixit, Avinash. *The making of economic policy: A transaction cost politics perspective*. Cambridge, MA: MIT Press, 1996.

Granovetter, Mark. "Economic Action and Social Structure: The Problem of Embeddedness." *American Journal of Sociology*, November 1985, *91*(3), pp. 481–501.

→ **Holmstrom, Bengt and Milgrom, Paul.** "Multitask Principal–Agent Analysis." *Journal of Law, Economics, and Organization*, Special Issue 1991, *7*, pp. 24–52.

Hutchins, Edwin. *Cognition in the wild*. Cambridge, MA: MIT Press, 1995.

→ **Joskow, Paul.** "The Role of Transaction Cost Economics in Antitrust and Public Utility Regulatory Policies." *Journal of Law, Economics, and Organization*, Special Issue 1991, *7*, pp. 53–83.

→ **Kahneman, Daniel and Tversky, Amos.** "On the Study of Statistical Intuitions." *Cognition*, March 1982, *11*(2), pp. 123–41.

Kenney, Roy and Klein, Benjamin. "The Economics of Block Booking." *Journal of Law and Economics*, October 1983, *26*(3), pp. 497–540.

Klein, Benjamin; Crawford, T. A. and Alchian, A. A. "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process." *Journal of Law and Economics*, October 1978, *21*(2), pp. 297–326.

→ **Levy, Brian and Spiller, Pablo.** "The Institutional Foundations of Regulatory Commitment: A Comparative Analysis of Telecommunications Regulation." *Journal of Law, Economics, and Organization*, October 1994, *10*(2), pp. 201–46.

→ **Lyons, Bruce.** "Empirical Relevance of Efficient Contract Theory: Inter-firm Contracts." *Oxford Review of Economic Policy*, Winter 1996, *12*(4), pp. 27–52.

Matthews, R. C. O. "The Economics of Institutions and the Sources of Growth." *Economic Journal*, December 1986, *96*(384), pp. 903–18.

Nee, Victor. "Sources of the New Institutionalism in Sociology." Unpublished manuscript, Cornell University, 1997.

North, Douglass. "Institutions." *Journal of Economic Perspectives*, Winter 1991, *5*(1), pp. 97–112.

North, Douglass and Weingast, Barry. "Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in 17th Century England." *Journal of Economic History*, December 1989, *49*(4), pp. 803–32.

Rabin, Matthew. "Psychology and Economics." Unpublished manuscript, University of California-Berkeley, 1996.

→ **Shelanski, Howard and Klein, Peter.** "Empirical Research in Transaction Cost Economics: A Review and Assessment." *Journal of Law, Economics, and Organization*, October 1995, *11*(2), pp. 335–61.

Simon Herbert. "The Architecture of Complexity." *Proceedings of the American Philosophical Society*, December 1962, *106*(6), pp. 467–82.

———. "Human Nature in Politics: The Dialogue of Psychology with Political Science." *American Political Science Review*, January 1985, *79*(2), pp. 293–303.

→ **Teece, David J.** "Profiting from Technological Innovation." *Research Policy*, December 1986, *15*(6), pp. 285–305.

→ **Weingast, Barry R.** "The Economic Role of Political Institutions: Market-Preserving Federalism and Economic Development." *Journal of Law, Economics, and Organization,* April 1995, *11*(1), pp. 1–31.

Williamson, Oliver E. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press, 1975.

———. "Comparative Economic Organization: The Analysis of Discrete Structural Alternatives." *Administrative Science Quarterly*, June 1991, *36*, pp. 269–96.

———. "Public and Private Bureaucracies." Unpublished manuscript, University of California-Berkeley, 1998.

Wilson, James Q. *Bureaucracy*. New York: Basic Books, 1989.