7

Holdup, Holdout, and Royalty Stacking: A Review of the Literature

Norman V. Siebrasse

7.1 INTRODUCTION

This chapter provides a critical review of the literature relating to remedies for patent infringement in the context of complex products, with a focus on the underlying theoretical issues of holdup, holdout, and royalty stacking.

A royalty can only be considered excessive when measured against some benchmark. Section 2 of this chapter considers the conceptually appropriate benchmark for a fair return to a patentee. Section 3 reviews the theory relating to "holdup," which is used generically to mean any mechanism by which a patentee, bargaining with the expectation of being able to enjoin any unlicensed use, might be able to extract a royalty that exceeds the benchmark. Section 4 reviews mechanisms by which holdup can be mitigated. Section 5 attempts to place this debate relating to holdup into the context of the general literature on property rules versus liability rules. Section 6 considers "holdout" mechanisms, which may allow implementers to force a patentee to accept a royalty that is lower than the fair benchmark. Royalty stacking refers generally to any mechanism by which the total royalty burden is unduly increased by the presence of multiple patentees. It is the focus of Section 7, while Section 8 considers empirical evidence relating to holdup and royalty stacking.

7.2 BENCHMARK RETURN TO PATENTEE

7.2.1 A *Share of the Discounted Incremental* Ex Ante *Value:* $\theta\beta v$

To decide whether a royalty is excessive or inadequate requires comparison with a benchmark. A prominent benchmark is that used by Lemley & Shapiro (2007a), namely a share of the incremental *ex ante* value of the invention as compared with the next best alternative, discounted by the probability of validity and infringement, where the patentee's share is determined by its bargaining power. This can be summarized as $r^* = \theta\beta v$, where r^* is the benchmark royalty, θ is the probability

that the patent is valid and infringed, β is the patentee's bargaining power, and v is the incremental *ex ante* value of the invention.[1]

7.2.2 *Incremental Value Over Best Alternative:* v

1 Overview

On an instrumental view of the patent system, the patentee's incentive to invent should be commensurate with the social value of the invention,[2] and it is widely acknowledged that the social value of a technology is its incremental *ex ante* value over the next best alternative.[3]

The view that the value of an invention depends on its value over the best alternative is premised on the view that the patent system should incentivize the invention of socially beneficial products. If an already known drug treats pain effectively, and there is a new drug that is equally effective but no better in any respect, then it would be wasteful to spend social resources on the new drug that offers no advantages over the old drug. While this basic starting point is widely accepted, there are a number of details that are open to debate.[4]

[1] Lemley & Shapiro 2007a, 1999.

[2] Farrell et al. 2007, 610–11; Shapiro 2007.

[3] *See, e.g.*, Swanson & Baumol 2005, 10–11; Farrell et al. 2007, 610–11; Elhauge 2008, 545; Denicolò et al. 2008, 577–78; Layne-Farrar et al. 2009, 448; Shapiro 2010, 282; Gilbert 2011, 642; Camesasca et al. 2013, 304; Cotter 2013a, 128; Carlton & Shampine 2013, 536, 545; Jarosz & Chapman 2013, 812; Cotter 2014a, 357; Sedona Conference 2014, 23–24; Contreras & Gilbert 2015, 1468–69, 1499–1500; Siebrasse & Cotter 2017a; Lee & Melamed 2016, 411–12.

Golden 2007, 2144 n.119, challenges this, saying "The value of a patented invention is not necessarily merely its worth relative to that of an alternative. This can be appreciated by recognizing, for example, that my ability to purchase a bottle of Soda 2 for $1.00, rather than a bottle of Soda 1 for $1.25, does not mean that Soda 1 is worth only $0.25 – the difference between the values of the two choices." At first glance, this argument apparently fails to recognize the distinction between the value of the invention and the value of an embodiment of that invention. An invention is information, with zero marginal cost, while the embodiment of the invention may well have a substantial marginal cost. Suppose a consumer is indifferent between Soda 2 for $1 and Soda 1 for $1.25, and the cost of the ingredients for the two sodas is exactly the same, but Soda 1 has a flavor-enhancing technology, with a zero marginal cost. The value of the flavor-enhancing technology is $0.25, but the value of a bottle of Soda 1 is $1.25, because of the cost of the tangible ingredients. However, judging from the remainder of the passage, Golden's real concern may have been with cases where the alternative is also patented: *see* the discussion below in Section 7.2.2.a "Patented Alternatives." Golden 2007, 2138 also argues that the marginal value of an invention is difficult to determine. While this is no doubt often true, it is not a conceptual objection to the benchmark, but a practical one, which perhaps more properly goes to the point of whether a reasonable royalty is adequate compensation, which is discussed below in Section 5.1 "Inaccuracy of Damages Awards." Moreover, as Lemley & Shapiro 2007b, 2169 point out, it is not necessary to measure the marginal value to conclude that holdup allows the patentee to extract an excess.

[4] Lemley & Shapiro 2007b, 2169, state that "[c]ertainly, [v] is well defined conceptually." This is true only for the paradigmatic cases.

2 Incremental Value

A) PATENTED ALTERNATIVES. While it is widely acknowledged that the value of the invention is its incremental value over the best alternative, there is not a consensus when the best alternative is patented. Some authors explicitly identify the value of the patented technology as its incremental value over an unpatented alternative, and simply do not consider the case of a patented alternative,[5] but it is quite common to simply remain silent on the issue.[6]

What might be termed a "strict" interpretation of the incremental *ex ante* approach treats patented and unpatented alternatives in exactly the same way. Most prominently, Swanson & Baumol (2005) explicitly take the benchmark to be the strict incremental value of a patented invention over the best patented alternative, so that if two patented technologies are equally effective, the benchmark royalty is its marginal cost (potentially zero).[7] Siebrasse & Cotter (2017a) argue that this strict approach is wrong because if the royalty received by a patentee is equal to the marginal cost of manufacturing and licensing the technology, there will be an insufficient incentive to invest the sunk costs of invention in the first place. As they note: "The fact that two patentees develop equivalent technology at the same time does not mean that neither required the lure of a patent. Viagra and Cialis may be equally effective in treating erectile dysfunction, but that does not imply that they both would have been invented if pharmaceutical patents were not available."[8] Consequently, this strict interpretation of the incremental *ex ante* approach is inconsistent with the basic rationale of the patent system, which is to allow an inventor to recover some part of their sunk costs of invention.

Lemley & Shapiro (2007a) deal with the question of patented alternatives in a footnote, saying that when the best alternative is patented, "[t]he proper comparison is between the cost and value of the patentee's component and the cost and value of the alternative, including patent royalties that would have to be paid on the alternative where appropriate."[9] However, when there is no established royalty for the alternative it is not clear how Lemley & Shapiro would determine the royalty that would have to be paid on the

[5] Farrell et al. 2007, 612–15; Layne-Farrar et al. 2009, 456.

[6] *See, e.g.*, FTC 2011, 191–94; Shapiro 2010, 282.

[7] Swanson & Baumol 2005, 18–19; Layne-Farrar et al. 2007, 686 (expanding on the model of Swanson & Baumol 2005 and making the same assumption); Carlton & Shampine 2013, 541 n.25 (specifying that "[t]he alternatives could be patented or unpatented").

[8] Siebrasse & Cotter 2017a, 1192–93; *see also* Kieff & Layne-Farrar 2013, 1120 (leveling substantially the same criticism of the strict approach in the SEP context, noting it will result in "reduced SSO participation and suboptimal investment in innovation"); Golden 2007, 2144 n.119 (noting that "the fact that a patent has inspired the discovery of a [patented] substitute does not mean that the patented contribution should be considered to be devoid of value").

[9] Lemley & Shapiro 2007a, 2039 n.153; *see also* Contreras & Gilbert 2015, 1468 taking essentially the same position.

alternative.[10] Moreover, even if there were an established royalty for the alternative, competition from the new patented invention would presumably affect the royalty charged by the alternative patented technology. It is not clear whether the royalty to be taken into account for the alternative is the royalty that was actually being charged prior to the introduction of the new technology, or the royalty that would have been charged after the introduction of the new technology.

Consequently, Siebrasse & Cotter (2017a) suggest that there is, as yet, no satisfactory approach to determining the value of an invention in comparison to a patented alternative.[11]

B) INCREMENTAL VALUE TO DIFFERENT USERS. Epstein et al. (2012) argue that "it is a serious mistake to suppose that there is any such unique number that counts as the incremental value of a patent. Generally, different buyers will derive different benefits from implementing any particular technology," and consequently the incremental value "should not be given any prescriptive weight."[12] They appear to view this as both a conceptual and practical criticism of the incremental value approach. It is misplaced as a conceptual criticism, as the incremental social value of the invention is the aggregate of its incremental value to particular individuals.[13] While the incremental value benchmark faces substantial difficulties in implementation, this is true of any conceptual model. No method of assessing damages is perfect, and whether an explicit application of the incremental value approach is so impractical as to not be worth pursuing depends on the evidence available in the particular case and the feasible alternative methods.[14]

[10] *See* Elhauge 2008, 564–65 (pointing out that Bertrand competition between the patentees implies that the royalty will be zero if both patented alternatives are equally valuable as compared with the unpatented alternative).

[11] Siebrasse & Cotter 2017a (stating that "[w]e are not aware of any literature providing a thorough theoretical analysis of this problem, and the solution is not evident").

[12] Epstein et al. 2012, 37.

[13] Epstein et al. 2012, 37, suggest that in the context of SEPs such an interpretation, which would result in a different royalty to different users, would "violate the RAND policies without cause." Presumably they are referring to the nondiscrimination requirement. It is by no means settled that differing royalties to differently situated implementers would violate the nondiscrimination requirement: *see* Carlton & Shampine 2013, 546 (arguing that "non-discriminatory" means that similarly situated firms should pay the same royalty, where firms are similarly situated only "if ex ante they expect to obtain the same incremental value from the patented technology compared with the next best alternative"); *compare* Gilbert 2011, 875 (arguing that all licensees should be able to choose from the same schedule of royalties, even if they do not pay the same rate). In any event, even if there are good policy reasons why in the SEP context the nondiscrimination requirement should be interpreted as meaning that different implementers should pay the same royalty, the point remains that the value of a patented technology is the aggregate of its incremental value to particular individuals.

[14] Epstein et al. 2012, 38 say that "[t]he complex institutional framework makes it apparent that no meaningful 'incremental value' calculation can be done." This is overstated. The courts have regularly attempted to assess the value of the invention at issue over the alternatives, and the results, while no doubt imperfect, have certainly been meaningful: *see*, *e.g.*, *Microsoft Corp.* v. *Motorola, Inc.*

3 Ex Ante

A) WHY "EX ANTE"?. The fair benchmark royalty should relate to the value of the patented technology. As discussed in more detail below, there are situations in which the amount at stake in negotiations between a patentee armed with an injunction and an implementer[15] is more than just the value of the technology. For example, if the implementer has sunk costs in implementing the technology, the patentee might be able to "hold up" the implementer for some part of those costs that would be lost to the implementer if its use of the technology were enjoined. The incremental value is assessed before the potential for holdup arises – "*ex ante*" – on the view that the return due to holdup is not properly attributable to the invention.[16] The intuition is that the true incremental value of the patented technology over the best alternative is the most that a licensee would pay for license to the patented technology in pre-adoption negotiations, on the view that if the patentee demanded a higher royalty it would be more profitable for the user to adopt an unpatented alternative.[17] As Siebrasse & Cotter (2016) emphasize, the construct of an "ex ante negotiation" is only a mechanism for isolating the value of the patented technology from other value that might be appropriated by a patentee armed with an injunction, such as the implementer's sunk costs.

B) WHEN IS "EX ANTE"?. While there is general agreement that the appropriate benchmark is *ex ante* value, there is inconsistency on the specifics: "*ex ante*" is variously used to mean prior to sunk costs being incurred; prior to a standard being adopted (in the context of SEPs); or prior to first infringement.[18] Since the reason for an *ex ante* assessment is to avoid including holdup value, it follows that the precise meaning of "*ex ante*" turns on the type of holdup one is concerned with. If the concern is sunk costs holdup, then *ex ante* means before the implementer incurs sunk costs. If the concern is that a patentee should not be able to capture value

(W.D. Wash. 2013) (U.S.); *In re Innovatio IP Ventures, LLC Patent Litigation* (N.D. Ill. 2013) (U.S.); *Grain Processing Corp.* v. *Am. Maize-Prods. Co.* (N.D. Ind. 1995, p.1390–93) (U.S.).

[15] While the term "implementer" is often associated with the standards context, where it is used to mean a party implementing a standard, in this chapter I will use it more broadly, as a generic term for any party who might use or implement a patented technology. This includes both infringers and parties who may be infringers, though the action is settled because infringement is determined. It also includes noninfringers, such as licensees and potential licensees, parties who choose not to use the technology at all after failed negotiations.

[16] *See* Lemley & Shapiro 2007a, 1999 (describing the benchmark royalty as "the royalty rate that would be reasonable and expected in the ideal patent system without any element of holdup").

[17] Carlton & Shampine 2013, 540; Lee & Melamed 2016, 392.

[18] The general rule in U.S. law is that a reasonable royalty is assessed on the basis of a hypothetical negotiation taking place at the time of the first infringement; *see* Lee & Melamed 2016, 422–25 (reviewing the cases). However, some courts have said that in the context of SEPs the appropriate time is before the standard is adopted: *see Apple, Inc.* v. *Motorola, Inc.* (N.D. Ill. 2012, p.913) (U.S.); *In re Innovatio IP Ventures, LLC Patent Litigation* (N.D. Ill. 2013) (U.S.); *Microsoft Corp.* v. *Motorola, Inc.* (W.D. Wash. 2013) (U.S.). Some scholarship focusing on lock-in suggests the appropriate time is prior to lock-in occurring: *see, e.g.*, Lee & Melamed 2016.

arising from network effects on standardization, then *ex ante* should be taken to mean before a standard is adopted and network effects arise.

c) ex ante vs. ex post information. As noted above, it is widely accepted that the value of an invention is the amount that would be negotiated by willing parties *ex ante*. It is often assumed that this "*ex ante*" value must only take into account information actually available to the parties *ex ante*, so that if subsequent information ("*ex post* information") reveals that the invention was more or less valuable than would have been anticipated by the parties, that information should be ignored. Siebrasse & Cotter (2016) refer to this as a "pure" *ex ante* approach. They critique this approach, pointing out that the rationale for the *ex ante* nature of the benchmark royalty is to avoid providing the patentee with a return reflecting holdup, and this does not justify excluding *ex post* information. Building on a point made by Mariniello (2011), they argue that the incremental value of the invention should be determined on a "contingent *ex ante*" basis, under which the implementer is assumed not to have invested any sunk costs, but all available information is used to assess that value, including *ex post* information.[19] Siebrasse & Cotter (2016) argue that using all available information allows a more accurate assessment of the true social value of the invention and therefore more accurately aligns the patent incentive with the social value of the invention.[20]

Lee & Melamed (2016) provide the most sustained scholarly argument in favor of a pure *ex ante* approach, in which all *ex post* evidence is excluded except to the extent that it may be used to establish what the parties would have agreed to based purely on *ex ante* information.[21] They have three main objections to the use of *ex post* information: (1) "the rationale [for using *ex post* evidence] assumes that the actual profits would have been unforeseen entirely at the time of the hypothetical negotiation"; (2) "a royalty determined on the basis of ex post evidence will generally include a premium based on ex post economic developments that increase the infringer's reliance on the patent – in particular, lock-in costs – and that are unrelated to the incremental benefit the patent confers"; and (3) "[b]ecause the rationale is meant to avoid undercompensating the patent holder, often the only ex

[19] Carlton & Shampine 2013, 545 n.40, characterize Mariniello 2011 as disagreeing with the *ex ante* approach, when in fact he only disagree with using only *ex ante* information. That is, Carlton & Shampine implicitly assume that if the negotiation is *ex ante* for purposes of sunk costs, it must necessarily also be *ex ante* for the purposes of information.

[20] *See also* Jarosz & Chapman 2013 (arguing that assessment of reasonable royalty damages should consider all available evidence, including information generated after the date of the hypothetical negotiation); Geradin & Layne-Farrar 2007, 98 (criticizing "any ex ante approach" on the basis that for that matter, "it may hinder innovation in those cases in which the value of an invention is unclear at the moment of standardization." This implicitly assumes that all *ex ante* approaches must use only *ex ante* information); Epstein et al. 2012, 34 (arguing that a measure of damages fixed at the time the standard is adopted will fail to recognize changes in the value of technology over time).

[21] *See also* Gooding 2014.

post information considered is that which tends to increase the royalty rate."[22] The analysis provided by Siebrasse & Cotter (2016) does not turn on point (1). With respect to point (2), Siebrasse & Cotter (2016) agree that value arising from lock-in cost should be excluded, but they argue that this does not require excluding *ex post* information generally. And point (3) is not an argument against the use of *ex post* information as such, as opposed to an argument against the one-sided use of *ex post* information. Siebrasse & Cotter emphasize that the rationale for using *ex post* information is that it allows more accurate assessment of the true value of the invention, whether that true value is higher or lower than would have been anticipated by the parties *ex ante*.

In summary, while it is widely accepted that the value of an invention is the amount that would be negotiated by willing parties *ex ante*, there is relatively little scholarship that distinguishes an *ex ante* negotiation from the use of *ex post* information, and there is no consensus as to whether *ex post* information should be used.

D) FULL EX ANTE APPROACH. One criticism of the standard *ex ante* approach is that the hypothetical negotiation is assumed to take place before the implementer has sunk costs into implementing the invention, but after the patentee has invested sunk costs into inventing the patented technology. Epstein et al. (2012) point out that since the purpose of the patent system is to provide an incentive to invent, it is wrong to assume that the invention has already been invented. Instead, they argue that we should consider what bargain would be arrived at in "a truly 'ex ante' setting – that is, at the outset of a new technology, before either inventors or manufacturers have made the investments necessary to the success of that technology."[23] While Epstein et al. (2012) make this point as a criticism of the standard *ex ante* approach, they do not explain how the "truly" *ex ante* approach would differ from the standard *ex ante* approach. As they point out themselves, *ex ante* licensing in the sense used by the standard approach, after the patentee has invented but before the implementer has sunk costs, is common in practice, and voluntary *ex ante* licensing of this type provides the primary return to the patentee, and so the primary incentive to invent, in many, perhaps most areas of technology.[24] Prima facie, the "true" *ex ante* approach reflects this practical bargain; the patentee sinks costs of invention in return for the right to negotiate a license before the implementer has sunk costs of implementation. The incremental *ex ante* benchmark for reasonable royalty damages simply attempts to replicate this bargain. This implies that the standard *ex ante* model corresponds to the "true" *ex ante* model advocated by Epstein et al.

[22] Lee & Melamed 2016, 416.

[23] Epstein et al. 2012, 10 (referring specifically to FTC 2011, the FTC "*ex ante*" model; but the point is equally applicable to Lemley & Shapiro's model); *see also* Layne-Farrar et al. 2014, 29 n.14 (noting that the *ex ante* terminology, though standard, "might be misleading. That period is ex ante for implementers, but it is ex post for patent holders, who have already sunk their R&D investments at that time. A better term would be 'medio amne' or midstream.").

[24] Epstein et al. 2012, 17.

(2012). Of course, it is possible that the voluntary *ex ante* licenses that are in fact the primary source of the incentive to invent do not provide an optimal incentive to invent. If that is the case, it is a problem for the larger patent system to address, perhaps by adjusting the patent term or scope.

7.2.3 *Bargaining Power Discount:* β

1 What is "Bargaining Power"?

"Bargaining power" or "bargaining skill" is a term used in two related but distinct ways.[25] Theoretically, it is used in the context of the solution to the bargaining problem, as initially set out in Nash's famous paper of that name.[26] If two parties with the opportunity to collaborate for mutual benefit are rational, they will engage in an exchange that maximizes the total net benefit to the parties jointly.[27] This net benefit is often referred to as the gains from trade.[28] The bargaining power discount, β, represents the way in which the parties to a negotiation split the gains from trade.[29] So, if the patentee appropriates the entire gains from trade we would say $\beta = 1$, and if the bargaining power is equal the parties will split the gains equally, then $\beta = 0.5$.

Based on this theory, a patent licensing negotiation is often modeled as a bargaining problem in which the gains from trade are the difference between the patentee's minimum willingness to accept and the implementer's maximum willingness to pay (also sometimes referred to as the threat point).[30] In an *ex ante* negotiation over an ironclad patent, the implementer's maximum willingness to pay is normally taken to be the value of the invention, v, as its threat point is to walk away from the negotiation and use the best noninfringing alternative. In assessing a reasonable royalty, the patentee's minimum willingness to accept is normally taken to be its marginal cost.[31]

[25] The term bargaining "skill" is often used to mean what I have been referring to as bargaining "power," as encompassing all residual factors that might affect the split in the gains to trade: Nash 1950 refers only to bargaining "skill," not bargaining "power." Lemley & Shapiro 2007a, use "skill" to refer to the general division of gains from trade, and "negotiating power" or similar terms to refer to specific factors, such as holdup, which affect the negotiated royalty. In my view it is more natural to use "skill" to refer to negotiation skills, as a small cash-constrained patentee might have to settle for a small share of the gains from trade even if its chief negotiator is a very skillful bargainer. However, when factors such as discount rates are modeled explicitly, then it is useful to use "skill" to refer to any residual factors affecting the split.

[26] Nash 1950.

[27] *Id.* at 155, 159.

[28] *See, e.g.*, Lemley & Shapiro 2007a, 1997; Elhauge 2008, 538.

[29] *See, e.g.*, Lemley & Shapiro 2007a, 1995–98 (citing Nash 1950).

[30] *See, e.g., id.* at 1997–98.

[31] Strictly the minimum willingness to accept is equal to the patentee's marginal cost only if it could not exploit the invention itself. If it could exploit the invention, its minimum willingness to accept is its opportunity cost of doing so, but in such a case lost profit damages would normally be appropriate.

In light of this model, bargaining power is also often used to mean that part of the difference between the value of the invention and the patentee's marginal cost that will be captured by the patentee in an actual license negotiation. This practical meaning of bargaining power and the theoretical meaning correspond only if the bargaining model just discussed accurately describes real-world negotiations. In particular, the Nash Bargaining Solution is strictly applicable only to negotiations over pure gains from trade. It is well understood that invention is only the first step toward commercialization, and that an implementer normally must make product specific investments in manufacturing, advertising, and distribution, and so on, in order for a product to be successfully commercialized. A number of authors have suggested that in practice the implementer's share of the profits is, at least in part, a return to the implementer for its contribution to that joint value.[32] If that is correct, then the split in profits negotiated by parties to a real-world agreement does not correspond exactly to the split in gains from trade in the theoretical Nash Bargaining Model.[33]

Returning to the basic theoretical bargaining model, the literature as to what determines bargaining power is thin. Nash's famous paper setting out what is now known as the Nash Bargaining Solution, took equality of bargaining power as a premise, and did not consider any of the factors that might affect it.[34] Formal game theory has added little to the concept to give it more real-world content. The main theoretical refinement is by Rubinstein (1982), who shows that under certain conditions, a party with a higher discount rate (higher time value of money) will have less bargaining power. This supports the informal view that resource constraints

[32] *See* Goldscheider et al. 2002, 130 (noting that "typically 75 per cent of the work needed to develop and commercialize a product must be done by the licensee"); Cotter 2009, 1169 (noting that "[i]n a sense, producers of end products are not merely users of the patented invention, but rather might be thought of as sequential innovators"); Lemley & Shapiro 2007b, 2167 (explaining that the value of the innovation is "jointly created" by various parties "including other patent holders and the downstream firm itself"); Siebrasse & Cotter 2016, 954–55 ("In an actual license agreement, both parties bring something to the table in the process of turning an invention into a commercially valuable revenue-generating product. The patentee's most obvious contribution is the invention, but bringing the final product to market will generally require further development and technical implementation, such as clinical trials, as well as marketing, manufacturing, and distribution, all of which require further investment at risk beyond the investment made by the patentee in the invention itself. These further services may be provided by either of the parties, and the way the parties split the incremental profit in an actual negotiation depends on who provides what services and on the relative importance and cost of those services.").

[33] To the extent that the implementer's share of the profits reflects a return to these kinds of transaction-specific investments, this could in principle be reflected in the bargaining model by adjusting the implementer's maximum willingness to pay accordingly. And to the extent that the bargaining power discount in a particular case is determining by looking to what similarly situated parties actually negotiated, any returns to the implementer that are necessary in the real world will automatically be included. The point remains that the familiar theoretical bargaining problem is probably not a complete model of actual patent license negotiations.

[34] Nash 1950, 159. More precisely, Nash assumes (Proposition 8) that similarly situated parties would split the gains from trade equally, and showed that rational parties would arrive at an agreement that maximizes the gains from trade.

affect bargaining power. Many other factors doubtless also affect real-world bargaining power, such as the ability to drive a hard bargain by psychological negotiating tactics (which might be termed bargaining skill), or repeat play and reputation effects.

The Nash Bargaining Solution is applicable only when there is some degree of bilateral monopoly, at least in the sense of the object of exchange having unique value to one of the parties.[35] This means that the market structure is also relevant to bargaining power. For example, a party negotiating with a counterparty in a competitive market will be able to extract the entire surplus by threatening to license to a different counterparty.

On the whole, the elegance and simplicity of the Nash Bargaining Model has made it a very attractive modeling construct, but a richer description of the factors that affect real-world bargaining power would be useful.

2 Justification for Bargaining Power Discount

The main justification provided by Lemley & Shapiro for using $\theta\beta v$ as the benchmark is that it reflects the royalty that would be negotiated by parties if they negotiated *ex ante*, and the return to voluntary market negotiations is theoretically appropriate in the absence of any known market failure.[36] So, as noted immediately above, the implementer's share of the profits may represent in part a return to investments made by the implementer to commercialize the invention. If so, voluntary market negotiations will provide an appropriate return to that investment, as will a benchmark that mimics the market.

As well as looking to the incentive to invent, the royalty also affects implementer incentives. To the extent that the gains from trade represent pure economic rents, then the particular split does not in principle affect the incentive to implement because any positive share provides a greater return to the implementer than does the best alternative.[37] However, to the extent that the implementer's share of the profits is a return to the implementer's technology-specific investments, holdup is inefficient, even if the royalty is less than θv, because the implementer may have to pay more for the patented technology than the value the implementer derives from the technology. When that is the case, the implementer may avoid implementing

[35] *Id.* at 155 (noting that the article treats "the classical problem of exchange, and, more specifically, of bilateral monopoly").

[36] Lemley & Shapiro 2007a, 1999 (stating that the benchmark is intended "to reflect the royalty rate that would be negotiated, prior to any infringement, if the patent were known to be valid"); *see also* Lemley & Shapiro 2007b, 2165; Cotter 2009, 1182 (preferring Lemley and Shapiro's use of the bargaining power discount over Elhauge's approach for this reason).

[37] Elhauge 2008, 538, says explicitly that in Lemley & Shapiro's model, β reflects only a split in the joint gains from trade, and not any reward for relative contribution of the parties to the creation of that joint value. However, it is not clear that this is an accurate characterization of Lemley & Shapiro's model in particular, or of models of how parties split the value of the invention more generally.

the technology, even though the technology itself would have a net social benefit.[38] It is the avoidance of efficient investments by implementers that is the real downside of holdup.[39]

3 Criticism of Bargaining Power Discount

A) OPTIMAL RETURN TO PATENTEE. Elhauge (2008) argues that the bargaining power discount is inappropriate in principle.[40] His basic argument is that if the return to the patentee is less than the full social value of the invention there will be socially valuable inventions that the patentee will not have an incentive to invent.[41] The bargaining power discount, β, arbitrarily depresses the return to the patentee. Elhauge therefore takes θv to be the benchmark return.[42] Consequently, he argues that the risk of holdup is much less than is suggested by Lemley & Shapiro.

The main objection to Elhauge's analysis is that it is doubtful that full appropriability of the social value of the invention is the appropriate benchmark. As just discussed, the main justification for the bargaining power discount provided by Lemley & Shapiro is that it mimics the market. If parties to a voluntary transaction would include such a discount, then prima facie that is efficient.

Elhauge's implicit response is that there is market failure, because the bargaining discount, even if voluntarily negotiated, provides an inadequate incentive to invent. As noted, the thrust of his argument is that if the patentee cannot capture the full social value of the invention, there will be inventions that would be socially beneficial for which inventors will not have an adequate incentive to invent.[43] However, the dominant view is that full appropriability is probably not optimal and certainly it is not so clearly optimal as to justify a departure from the prima facie benchmark of a voluntary market transaction.[44]

[38] If the implementer knows with certainty that it will have to pay more than the value it can extract from the invention, it will avoid implementing the invention entirely. More generally, the potential for of holdup makes the investment riskier and inefficiently depresses the degree of investment.

[39] Lemley & Shapiro 2007b, 2164 (explaining that "holdup is recognized as a form of market failure that leads to inefficiency, primarily by discouraging what would otherwise be socially desirable investments").

[40] *See also* Denicolò et al. 2008, 577 n.27 (saying that they consider it "more natural to assume that the negotiating parties would agree on a license fee of v, remunerating the patent holder fully for the value its innovation contributes to the product," but they do not explain exactly why they consider this more natural, and in any event their analysis uses the β discount).

[41] Elhauge 2008, 541.

[42] *Id.* at 545.

[43] *Id.* at 543 (arguing that the Lemley-Shapiro model is wrong to ignore this); *see also* Shavell & van Ypersele 2001, 535 (suggesting that full appropriability of the social value of the invention by the patentee is the appropriate baseline).

[44] *See, e.g.*, Frischmann & Lemley 2007, 268–71; Golden 2010, 529–31; Scotchmer 1991, 31; Shapiro 2007, 114–77. *See also* Sichelman 2014 (arguing that traditional remedies may either over- or undercompensate patentees as compared to the socially optimal return, depending on the circumstances).

Elhauge's arguments for full appropriability are not sufficiently persuasive to displace the dominant view. It is true that on the one hand the problem of imperfect appropriability tends to result in too little investment in invention. But on the other hand, the so-called patent race problem tends to lead to excessive research. The patent race arises when multiple parties try to capture the winner-take-all prize of a patent. The marginal social benefit of additional research is the benefit of an earlier invention date, but the marginal private benefit is the increased chance of capturing the entire value of the patented technology, not just the marginal benefit of an earlier patent date. This divergence between social and private benefit tends to lead to wastefully duplicative research by firms competing for the patent prize, or excessively rapid invention, or both.

These two problems tend in different directions, and, in a leading article, Dasgupta & Stiglitz (1980) conclude that "there is no clear presumption whether ... there will be excessive or inadequate research" when the patentee is able to capture the full appropriable surplus.[45] While Elhauge acknowledges this literature, he cites Dasgupta & Stiglitz (1980) for the proposition that an optimal patent term can be set to provide optimal incentives to invent and, "for *small inventions the market always provides inadequate research*."[46] However, that statement was made by Dasgupta & Stiglitz (1980) "within the confines of our simple model," and "for particular parameterizations," which includes in particular patents with an infinite life.[47] They do not suggest that this conclusion is generalizable.

Elhauge (2008) also quotes Dasgupta & Stiglitz (1980) as saying "where, with an infinite-lived patent, there is excessive expenditure on R&D, there is an optimal patent life."[48] He argues "if we assume, as makes sense to isolate the remedial issues at hand, that substantive patent law on issues such as patent length has been optimally set, then this literature supports awarding patent holders the full θv rather than discounting that amount by β."[49] However, the point being made by Dasgupta & Stiglitz is that the patent race problem is driven by the appropriable value of the invention, which increases with the patent term, which means that the patent race problem is at its worst if the term of the patent is infinite. If the patent race problem dominates the problem of an inadequate incentive to invent due to uncaptured social surplus, the patent race problem can be mitigated by reducing the patent term until an optimal balance is achieved. The rest of Dasgupta & Stiglitz's sentence – not

[45] Dasgupta & Stiglitz 1980, 21 (with an infinite-lived patent in markets with free entry into R&D); *see also* Tandon 1983, 156–57 (patent races might result in underinvestment or overinvestment in research).

[46] Dasgupta & Stiglitz 1980, 19 (their emphasis), quoted by Elhauge 2008, 544.

[47] Dasgupta & Stiglitz 1980, 1819. In particular, they assume constant elasticity demand curves, with elasticity less than unity, and an infinite life of a patent. *Id.* at 19. Shapiro 2007 also provides two simple models in which full appropriability is optimal, and again the restrictive requirement of these models illustrate the limits of full appropriability as a benchmark

[48] Dasgupta & Stiglitz 1980, 21, quoted by Elhauge 2008, 544.

[49] Elhauge 2008, 544.

quoted by Elhauge – concludes that "the optimal life of the patent will, however, vary depending on the size of invention and the elasticity of demand in the industry," and they conclude that "there is no simple intervention into the market allocation – no uniform rule applicable for all inventions and industries – which will attain the social optimum."[50] Since we know that the patent term does not, in fact, vary with those parameters, the proper conclusion from Dasgupta & Stiglitz is that we know that the patent term is not optimal, which implies that full appropriability is generally not optimal.

Elhauge (2008) also argues that θv understates the social value of the invention because v does not include social value arising after the patent term expires.[51] However, the patent reward should reflect the value of the inventor's contribution, which is only the earlier date of invention as compared with when the invention would have arisen even without the lure of a patent, as a result of general technological progress. If the patent term is set optimally the consumer surplus after expiry of the patent will not reflect any of the inventor's contribution, because the invention would have arisen anyway. While there is no particular reason to believe that the patent term is optimal, either on average or in any particular industry, neither do we know whether it is generally too long or too short.

In summary, the simple fact that the patentee cannot capture the full social value of the invention does not in itself allow us to conclude that the benchmark return proposed by Lemley & Shapiro, including the bargaining power discount, provides an inadequate incentive to invent.

B) CIRCULARITY. On a related point, Golden (2007) argues that Lemley & Shapiro's argument for the θβv benchmark is "fundamentally circular," because their justification that it represents "the royalty rate that would be reasonable and expected in the ideal patent system without any element of holdup" assumes that a patent holder "should obtain no more than it would receive if an injunction were unavailable."[52] However, this is not really a circularity problem, because Lemley & Shapiro's main point is that the appropriate benchmark reflects the royalty rate that would be negotiated in the absence of market failure, and injunctive relief, in some circumstances, can give rise to holdup, which is a well-known source of market failure.[53] The benchmark is therefore not simply the assumption that injunctions are unavailable, but rather that they do not give rise to holdup. Indeed, their benchmark implicitly assumes that when parties negotiate *ex ante*, they negotiate with the understanding that if they cannot agree, an injunction will be granted to restrain

[50] Dasgupta & Stiglitz 1980, 21.

[51] Elhauge 2008, 543 (arguing that the Lemley-Shapiro model is wrong to ignore this); *see also* Golden 2007, 2138 (noting that the limited patent term means that the patentee cannot capture the full social value of the invention).

[52] Golden 2007, 2139–40, quoting Lemley & Shapiro 2007a, 1999.

[53] *See* Lemley & Shapiro 2007a, 1999 (stating that the benchmark is intended "to reflect the royalty rate that would be negotiated, prior to any infringement, if the patent were known to be valid").

the user from infringing; it is that assumption that sets the user's maximum willingness to pay at the incremental value of the invention over the best alternative.

With that said, while Golden frames the issue as being a problem of circularity, his key point is that Lemley & Shapiro do not adequately recognize the benefits of injunctive relief that might justify its use despite giving rise to holdup.[54] Certainly Golden is right to say that the mere fact that injunctive relief might, or even certainly would, give rise to holdup in a particular case, is not a sufficient justification for denying injunctive relief without consideration of its countervailing benefits.

c) INDEPENDENT CREATION. Lemley & Shapiro (2007b) respond to the argument that patentees are under-rewarded because they cannot capture the social value of the invention after the term expires by saying "this argument is plainly incorrect in the central case where the infringing party independently develops the patented invention, which is common in holdup situations. In those situations, the patent holder's reward typically exceeds its social contribution, the finite patent lifetime notwithstanding."[55] This is a curious response. On its face, the main result from Lemley & Shapiro (2007a) is that, because of holdup, the patent holder's reward typically exceeds its social contribution, even when the invention *was* copied by the infringer. To say that, if the infringing party independently developed the invention, the patentee's reward will exceed its social contribution even when there is no holdup at all is an entirely different argument. Lemley & Shapiro (2007a) do say that "[a]n additional prerequisite for denying an injunction should be that the defendant developed the technology independently rather than copying it from the plaintiff,"[56] but that explicitly turns on what they see as countervailing considerations, and not on the view that there is no holdup if the infringer copied.

If the implementer independently developed the technology covered by the patent, the social value of the patentee's contribution will certainly be less than v, the value of the invention as compared with the best alternative. The discounts for validity and bargaining power are irrelevant to the true value of the patentee's contribution, so that value may well be less than $\theta\beta v$.[57] This may be taken to suggest that in cases of independent creation the benchmark royalty should be discounted by some entirely different factor to reflect the patentee's true contribution. However, this observation really supports the view that an independent invention defense should be introduced into patent law, as the patentee's contribution to the infringer's

[54] Golden 2007, 2140 ("A more satisfactory analysis would at least acknowledge long-recognized benefits of injunctions against infringement and would engage in some substantial analysis of whether their costs nonetheless outweigh their benefits.").

[55] Lemley & Shapiro 2007b, 2169. They attribute this argument to Golden 2007, 2136, though its main point is that the optimal reward to the patentee is indeterminate, not that the patentee is under-rewarded. In any event, Elhauge 2008 does make that argument, and the more significant point here is Lemley & Shapiro's response.

[56] Lemley & Shapiro 2007a, 2036–37.

[57] Shapiro 2007, 115–17; Shapiro 2010, 304.

product is always zero if the infringer develops the technology independently.[58] An independent invention defense would effectively "discount" the royalty – to zero – on a case-by-case basis, and this is more sensible than applying a general discount to all royalties to reflect the general percentage of cases in which the infringer did not copy. While introducing an independent invention defense has considerable theoretical appeal, it is of course not part of patent law. There have been a number of suggestions that an independent invention defense should be introduced into patent law, but the debate is not yet sufficiently developed to decide whether the absence of an independent invention defense is a defect of patent law that should be rectified as a matter of policy, or whether there is some good counterargument against an independent invention defense. Consequently, this is a situation in which we should assume for remedial purposes that substantive patent law is optimal; either it is in fact optimal, or the best solution is to amend substantive patent law. With that said, there is a separate question, as to whether independent invention should be a factor in determining whether injunctive relief is granted. That is different from the question of an independent invention defense, because even if injunctive relief is denied, the patentee would still be entitled to a reasonable royalty in the amount of the benchmark; however, that benchmark will not be adjusted to reflect independent creation.

D) ASYMMETRIC INFORMATION. Golden (2007) suggests that a patent holder will likely approach negotiations at a significant informational disadvantage that may "appear to tilt the likely result of negotiations toward an outcome corresponding to a low value for [β]."[59] However, it is not clear that information asymmetries will systematically favor the infringer.[60] And even if information asymmetries do favor the infringer, it is not clear that this will tilt the result to a lower share of value for the patentee. The patentee's ignorance might cause it to ask for too much, rather than too little, and the main effect of information asymmetry may be only to reduce the chance of settlement and increase litigation rates.[61]

[58] *See* Shapiro 2007, 127–35 (arguing for an independent invention defense for this reason).

[59] Golden 2007, 2132–33; *see also* Elhauge 2008, 549–50.

[60] Golden 2007, 2132 notes that the implementer will have better information about its costs and profit margin, but the patentee will likely have better information about its patent's validity, Lemley & Shapiro 2007b, 2170, and the patentee will likely have better information about previous licenses it has granted that are likely to affect the royalty awarded in litigation. The patentee will also have superior information if it is entitled to lost profit damages. Further, as Lemley & Shapiro 2007b, 2170 point out, and Golden 2007, 2130 acknowledges, any information asymmetry will be reduced by discovery, at least in the U.S. litigation system.

[61] Lemley & Shapiro 2007b, 2170. Elhauge 2008, 550 responds that the implementer "will accept when the patent holder demands too little . . . but won't accept when the patent holder demands too much," and therefore "the actual negotiated royalties will be lower than they predict." This is a variant on the "Option Effect" argument, discussed below, Section 7.5.4.2.b "Option Effect." Even if the option effect does depress the patentee's return (on the assumption that there is a systematic information asymmetry favoring the implementer), this does not affect the point that the problem is difficult to solve by changing legal rules related to remedies.

4 θ: Patent Strength

The benchmark fair royalty rate requires the value of the patented technology to be discounted by the strength of the patent, which is to say, the probability that it is valid and infringed. Otherwise, the implementer will be paying for a technology that it did not use, or for which no patent should have been granted. In U.S. law, a reasonable royalty is in principle awarded on the assumption that the patent was known to be valid and infringed – that is, without any discount for patent strength, since damages are only awarded if the patent has been held to be valid and infringed. This is not inconsistent with the principle that a fair royalty requires a discount for patent strength; on the contrary, it is necessary to avoid double discounting.[62] While these principles are not controversial, the extent to which the courts appropriately apply or ignore the patent strength discount is another question. The most complete analysis of that issue is Masur (2015), who characterizes existing U.S. law on this point as "both incoherent and backwards."[63]

7.3 HOLDUP

7.3.1 *Varieties of Holdup*

Despite the centrality of the concept of "holdup," it does not have any precise definition – or rather, it has a variety of precise definitions. In the broadest sense, holdup is used to mean any mechanism by which a patentee can extract a royalty that is higher than a fair benchmark royalty. In a slightly narrower sense holdup is used to mean any mechanism by which the royalty that might be demanded by a patentee *ex post* is higher than that which might be demanded *ex ante*, where *ex ante* is defined variously as the time at which infringement began, or sunk costs were

[62] Suppose that the parties would agree to a $1 million royalty *ex ante* if they knew the patent to be valid and infringed, but they each believe there is only a 70 percent probability of validity. The license they would actually negotiate would be appropriately discounted, to $700,000. If there is infringement and the patentee files suit, the patentee knows that it only has a 70 percent chance of obtaining a favorable judgment. If the amount of a favorable judgment is the actual $700,000 the parties would have negotiated, the patentee's expected pay-off from going to trial is only $490,000 (70 percent of $700,000), which means that the patentee will be worse off as a result of the infringement than if the infringer had licensed. The assumption of validity and infringement corrects for this problem by awarding the patentee $1 million if she prevails, so that her expectation pretrial is $700,000, exactly the amount she would have agreed to *ex ante*: for further discussion, *see* Cotter 2009, 1183; Taylor 2014, 115–16.

[63] Masur 2015, 127 (arguing that it is generally very difficult to apply an appropriate adjustment for patent strength because estimates of patent strength are private information not normally available to the courts, and further that licenses that are negotiated as litigation settlements in circumstances where the infringer was losing at trial are the best gauge of patent value, and yet such licenses are systematically excluded).

incurred by the implementer, or, in the SEP context, as the time at which the standard was adopted.[64]

Focusing on the *ex ante/ex post* version of holdup, Siebrasse & Cotter (2017a) note that there are three different mechanisms by which *ex post* royalties may be higher than *ex ante* royalties. They refer to these as (1) sunk costs holdup, (2) network value appropriation, and (3) the apportionment problem. These are discussed subsequently in this section. Lemley & Shapiro (2007a) provide a very influential model of holdup, which extends the holdup analysis to probabilistic patents. All of these mechanisms are said to be a potential source of excess returns to the patentee. High litigation costs are also said to be another potential source of holdup. However, the effect of the distortion due to litigation costs is ambiguous, as is discussed in Section 6 "Holdout/Reverse Holdup."

1 Sunk Costs Holdup

Farrell et al. (2007) describe "opportunism" or "holdup" as follows: Holdup can arise, in particular, when one party makes investments specific to a relationship before all the terms and conditions of the relationship are agreed upon.[65]

They provide the following example of holdup in a case where the patented technology costs $40 to implement, exclusive of any royalty, and the best alternative technology costs $50, so that the inherent advantage of the patented technology is $10, and a benchmark reasonable royalty is any amount less than this:

> [S]uppose that, of the $40 cost of using the patented technology, $25 was spent before the royalty was negotiated and that this $25 is specific to the patented technology, i.e., would be wasted if the user later decided against adopting that technology. Then, at the time of negotiations, the forward-looking cost of using the patented technology (exclusive of royalty) is $40 – $25 = $15, while the cost of using the unpatented technology remains $50 (the $25 already spent has no value if the user adopts the alternative technology) … [T]he maximum royalty that the user is willing to pay remains the added value of the patented technology, but with the key difference that this amount is now $50 – $15 = $35, or $25 more than in our first calculation. Ex post negotiation increases the user's willingness to pay for the patented technology because the user finds the alternative relatively less attractive after spending $25 on the patented technology. The patented technology's ex post advantage … exceeds its inherent advantage … by an amount equal to the user's $25 investment … The patent holder thus captures a share (proportional to its bargaining skill) of sunk investments by the user.[66]

[64] These concepts may be related, because one definition of a benchmark fair royalty is the royalty that would have been negotiated *ex ante*.

[65] Farrell et al. 2007, 604.

[66] *Id.* at 612–13.

That is, the fact that the user has made transaction-specific investments prior to negotiating for the right to use the technology means that the patentee can capture part of the user's sunk costs, in addition to the inherent advantage of the patented technology. (This analysis implicitly assumes that the successful patentee will be granted an injunction.) It is convenient to refer to holdup arising from such transaction-specific investments as "sunk costs holdup" where the transaction specificity is left implicit.[67]

Sunk costs holdup relies centrally on the transaction-specific investment being sunk before any negotiations take place. It does not turn on the product being complex, as it can arise when only one patent covers the product. Nor does it turn on the probabilistic nature of the patent, or on the cost of litigation – in the above example, litigation costs are assumed to be zero.

If sunk costs holdup does occur, it has adverse effects on both patentee and implementer behavior. It allows the patentee to capture more than the value of the invention, thus creating an excessive incentive to invest in patented technologies; and the prospect of being held up increases the *ex ante* risk to the implementer, thus reducing the attractiveness of investments in products that are potentially subject to sunk costs holdup.

2 Network Effect Appropriation

Another type of holdup, applicable primarily in the context of standards, is referred to by Siebrasse & Cotter (2017a) as network effect appropriation, which they define as follows:

> [N]etwork value appropriation, arises whenever the value of a particular technology increases upon standardization due to the presence of network effects. As with sunk costs holdup, an injunction would enable the patentee to extract a higher royalty ex post than it could have negotiated ex ante, and thus again might be described as resulting in the capture of some of the value of the standard – though in this context, the increase in value is due to network effects and does not depend on the presence of transaction-specific sunk costs.[68]

[67] The general analysis of this type of opportunism, which arises whenever a transaction is subject to "durable investments in transaction specific human or physical assets" is associated with Williamson 1985, 61. It is not specific to patent law, or even intellectual property; Williamson originally discussed it in the context of contracts. Williamson famously defined "opportunism" as "self-interest seeking with guile," *id.* at 47, and he emphasized the investment of sunk costs (which he referred to as "the fundamental transformation," *id.* at 61) as giving rise to the possibility of opportunism. However, as Farrell et al. 2007, 604, point out, "[t]he pure economics are largely unaffected by whether or not guile is involved . . . ." While guile is involved in some cases of sunk costs holdup, for example in case of so-called patent ambush, many holdup scenarios of central concern to authors such as Lemley & Shapiro 2007a, Farrell et al. 2007, and Lee & Melamed 2016, do not turn on any deceitful behavior by the patentee.

[68] Siebrasse & Cotter 2017a, 1166.

U.S. courts have consistently held that a reasonable royalty should not reflect "any value added by the standardization of that technology."[69] On its face this appears to say that a patentee should not be able to capture any value arising from network effects, though as Siebrasse & Cotter (2017a) point out, courts routinely award damages, including ongoing royalties in lieu of an injunction, in the form of running royalties, which do allow the patentee to capture value arising from network effects.[70]

From a policy perspective, Siebrasse & Cotter (2017a) argue that allowing the patentee to capture some part of the value of a patented technology that arises due to network effects is desirable from a dynamic efficiency perspective, because it provides the correct incentive to invent, and it is not undesirable from a static efficiency perspective, as it has no adverse effects on implementer incentives.[71] While there are many articles arguing that a patentee should not be able to extract a higher royalty *ex post* than it could have obtained *ex ante*, which suggests that the patentee should not be able to capture any value arising from network effects, such articles typically do not distinguish between sunk costs and value arising from network effects. The two often go hand in hand, because adopting a standard and the consequent network effects, is often accompanied by substantial sunk costs. Two exceptions are Swanson & Baumol (2005) and Lee & Melamed (2016), which both specifically assert that the patentee should not be able to capture value arising from network effects. However, both treat network effect appropriation equivalently to sunk costs holdup, and they do not offer independent policy justification for not allowing the patentee to capture any of the value arising from network effects.[72] Chao (2016) also takes issue with Siebrasse & Cotter (2017a) on this point, but his discussion turns on what Siebrasse & Cotter (2017a) characterize as the distinct problem of apportionment, which is discussed in the next section.[73]

[69] *See Ericsson, Inc.* v. *D-Link Sys.* (Fed. Cir. 2014, p.1232) (U.S.); *see also CSIRO* v. *Cisco Sys., Inc.* (Fed. Cir. 2015, p.1304) (U.S.).

[70] Siebrasse & Cotter 2017a, 1220. They also note that there is some ambiguity in these statements, as the courts do not clearly distinguish value arising from network effects from sunk costs or problems of appropriation.

[71] *See also* Geradin & Layne-Farrar 2007, 93 (suggesting that it is not clear why the essential patent holder should not capture part of the value arising on standardization).

[72] Swanson & Baumol 2005, 8–10; Lee & Melamed 2016, 429–30.

[73] Chao 2016, 304 (stating that a patentee should not be able to capture any part of what he calls "*ex ante* compatibility value"). While Chao states that he disagrees with Siebrasse & Cotter 2017a on this point, the example he gives to illustrate this point is of a patented technology that does not make the standard any better as compared with existing alternatives. Chao does not specify whether the alternatives were unpatented. In a case in which the patented technology included in the standard was no better than an unpatented alternative, this would be an example of what Siebrasse & Cotter 2017a describe as the problem of apportionment, and in their analysis such a patent would receive a royalty of zero.

3 The Apportionment Problem

Another type of holdup may arise when a patent claims a relatively minor feature of a complex product. If the patentee can get an injunction that prevents sale of the entire product, it can extract part of the value of the entire product, even though the patented technology contributes little to that value. This point is explored at length by Lemley & Shapiro's model, which is addressed next.

The apportionment problem, when it exists, has the same adverse incentive effects on patentee and implementer incentives as does sunk costs holdup. It allows the patentee to capture more than the value of the invention, creating an excessive incentive to invest in minor patented technologies. The prospect of being held up in this manner increases the *ex ante* risk to the implementer, thus reducing the attractiveness of investments in products that are potentially subject to the apportionment problem.

The apportionment problem is exacerbated in the context of a standard. As discussed in more detail below, the excessive royalty that can be extracted by a patentee armed with an injunction is generally capped by the losses that would be suffered while the technology is designed around. This implies that if the technology can quickly and easily be removed or designed around, then the royalty overcharge will be small. However, this is not true in the SEP context because licensing terms of SEPs almost always specify that the SEPs are only licensed for use in the products that comply with the standard.[74] That means that if the technology covered by a SEP could easily be designed around or removed as a technical matter, the product would no longer be compliant with the standard and the other licenses to the truly important SEPs would lapse. This would allow the owner of the unimportant SEP to capture the value of the standard as a whole.[75] It is not enough to design around a SEP technically; instead, the implementer would have to be able to lobby the standards organization to remove the technology in question from the standard. While not necessarily impossible, this will certainly be a very lengthy process.[76] In such a case, the "redesign period" referred to in the discussion below

[74] *See* American Bar Association (ABA) 2007, 60–61 .

[75] Alternatively, the design-around costs would include the cost of lobbying the relevant standard development organization to adopt a new version of the standard that excluded the controversial technology, and the lost profits during that period. If the licenses for the other SEPs do not contain such a term, the problem might still arise if it was necessary for marketing purposes to advertise that the product was compliant with the standard. In other cases, the holdup effect would relate primarily to the lost profits during the period of redesign, as identified by Lemley & Shapiro 2007a.

[76] Consider, for example, the interlaced video SEPs at issue in *Microsoft Corp.* v. *Motorola, Inc.* (W.D. Wash. 2013) (U.S.). On the evidence, these added little to the value of the standard, and presumably it would have been relatively simple to remove support from interlaced video from Microsoft's products, since it involved disabling a feature rather than adding one, but doing so would have rendered Microsoft's products noncompliant with the standard. Support for interlaced video was eventually removed from the standard.

should be interpreted to mean the time needed to get the SEP removed from the standard, rather than the time needed for a technical redesign.

4 Probabilistic Holdup: Lemley & Shapiro Model

Lemley & Shapiro (2007a) and Shapiro (2010) provide a widely discussed model of holdup.[77] Their model incorporates both sunk costs holdup and the apportionment problem, and additionally addresses the effect of the probabilistic nature of patents; that is, the fact that the validity and scope of granted patents is uncertain until they are litigated.[78] Their model extends the simple sunk costs holdup model in two other respects. First, in the simple sunk costs model the implementer's option is to license or redesign its product to avoid using the patented technology; Lemley & Shapiro develop a more explicit model of the litigation process in which the implementer may choose to redesign either during the litigation period, or after the end of litigation. Second, in the simple model, the implementer is at risk of being held up for transaction-specific sunk costs that are generally conceptualized as being machinery or other tangible goods. Lemley & Shapiro point out that the implementer is also at risk of being held up for lost profits during the period that its product is being redesigned to avoid infringement. Lemley & Shapiro also focus on redesign costs (sometimes referred to as switching costs), rather than sunk costs. (The relationship between switching costs and sunk costs is discussed below.)

They consider two scenarios: a "surprise" scenario in which the implementer is already selling its product when it learns of the patent, and an "early negotiation" scenario in which negotiations take place before the product is designed.[79] For "ironclad" patents – those that are certainly valid – their model is a variant on the standard sunk costs model of holdup; there is no overcharge when *ex ante* negotiations are possible, and in the *ex post* scenario the patentee extracts part of the costs of switching to a noninfringing alternative.

When considering probabilistic patents, Lemley & Shapiro further distinguish between two scenarios. If the patent is relatively weak, it will make more sense for the implementer to refrain from redesigning until after it has lost in litigation, in which case its threat point is determined by the sunk costs plus the lost profits on the entire product during the period of redesign. This is the "Litigate" scenario. On the other hand, if the patent is relatively strong, the implementer's best negotiating strategy is to threaten to redesign its product during litigation ("Redesign and Litigate"), in which case its threat point is determined by the redesign costs. This means that a weak patent will have a higher relative overcharge because it can extract not just redesign costs, but also lost profits on the entire product during the period of

[77] Lemley & Shapiro 2007a, 1995 n.7 (noting that their technical economic analysis is based on a working draft of Shapiro 2010).

[78] Regarding probabilistic patents, *see generally* Lemley & Shapiro 2005.

[79] These terms are taken from Shapiro 2010.

redesign. The overcharge is discounted by the probability of validity, so the absolute overcharge for a weak patent will normally be smaller than for a strong patent covering the same technology.

Lemley & Shapiro show that the probabilistic nature of patents can result in an overcharge even when *ex ante* negotiations are possible. This is because the implementer's threat is to avoid using the patented technology entirely, and adopt the best noninfringing alternative instead. This is appropriate for an ironclad patent because it allows the patentee to obtain part of the true value of the invention. The problem is that the implementer's threat is exactly the same, and so the outcome of the negotiation is exactly the same, even if the patent is potentially invalid. This implies an overcharge, because the royalty should be discounted by the probability of invalidity. The problem, as they put it, is that "the accused infringer has chosen to give up without a fight, effectively agreeing to treat a possibly invalid patent as certainly valid, and so the chance that it would have invalidated the patent will not be reflected in the negotiated royalty."[80] (Because the overcharge can be extracted even when an *ex ante* negotiation takes place, it is perhaps not strictly correct to refer to it as "holdup," which normally implies that a higher royalty can be extracted *ex post*, than could have been negotiated *ex ante*.)

To summarize their results:

Scenario 1 – Surprise – "Litigate" strategy

- Applicable when patent is weak, redesign costs are high.
- Overcharge because patentee can extract lost profits on the entire product during redesign, plus redesign costs, both discounted by probability of validity. Because of the discount the absolute value of the overcharge will be small, but because of the lost profits on the entire product, the percentage overcharge will be large.
- Overcharge increases with (a) redesign costs; (b) lost profits during redesign period; and (c) value of the product relative to the value of the invention.

Scenario 2 – Surprise – "Litigate & Redesign" strategy

- Applicable when patent is strong, redesign costs are low.
- Overcharge because P can extract redesign costs, not discounted.
- Percentage overcharge (a) increases with redesign costs, and (b) decreases with probability of validity (i.e., is greater for weak patents).

Scenario 3 – Early Negotiation

- Either just like surprise case,
  or
- Overcharge because implementer's threat is not to use the invention with certainty, in which case percentage overcharge decreases with the probability of validity.

[80] Lemley & Shapiro 2007a, 2005.

Their results do not turn directly on the complex nature of the product, but complex products are likely to be subject to a greater overcharge because they are likely to face Scenario 1, in which a weak patent with relatively little value to the product can nonetheless extract a portion of the value of the entire product.

The adverse economic effects of holdup in Lemley & Shapiro's model in the surprise scenario are the same as for sunk costs holdup (though the mechanism is somewhat different). The economic effects of probabilistic holdup in the early negotiation scenario are slightly different. Again, the patentee is capturing more than the value of its contribution, which creates an excessive incentive to invest in patenting. But in principle the overcharge will not increase the risk to the implementer, because it knows how much it has to pay *ex ante*. Nor will it cause the implementer to avoid using the patented technology, because the patentee will not charge so much that the implementer would prefer to use the alternative. It will in principle reduce the implementer's expected profit, thus creating a distortion in the direction of investments. The degree of the distortion will presumably depend on the market structure.

While this model is well-known and influential for its implications respecting injunctive relief, within the context of remedies, and particularly the withholding of injunctive relief, another implication is that additional effort should be devoted to weeding out weak patents before they are licensed or litigated.[81]

5 Sunk Costs, Switching Costs, and Lock-in

Holdup is sometimes described as involving switching cost, on the view that once one technology is selected, it may be that the cost of switching to the alternative technology is prohibitively expensive.[82] This characterization is used most often in the standards context, where the implementer is said to be "locked in" to the standard once it is chosen, but similar reliance on switching costs as giving rise to holdup is also found in other contexts.[83] This contrasts with the traditional focus of the general economic holdup literature on sunk costs, in which holdup occurs when a party tries to charge a higher price than it would have been able to before those sunk costs were incurred. The puzzle is that sunk costs were necessarily incurred in the past – a party cannot be held up for costs that it has not yet incurred – while "switching costs" on the other hand, imply costs that would take place in the future, after failed negotiations, to switch to an alternative, nonstandard technology.

Cotter et al. (2018) provide a general framework for reconciling concepts of switching costs and sunk costs. They explain that the threat of adopting the next best alternative always disciplines the royalty that can be extracted by the

[81] Shapiro 2010, 307.
[82] *See, e.g.*, Gilbert 2011, 862; DOJ & FTC 2007, 35; FTC 2011, 5.
[83] *See, e.g.*, Lemley & Shapiro 2007a, 2037.

patentee, but the value of both the patented technology and the alternative may change. After costs are sunk, the selected technology is more profitable going forward, because the costs of implementation have already been incurred. So sunk costs holdup can be thought of as representing holdup due to the differential profitability of the selected technology *ex ante* versus *ex post*. The differential profitability of the alternative technology represents a separate source of holdup. If the profitability of the alternative technology changes, either because its costs change, or because its revenues change – as when it is not selected to be the standard – the disciplining value of the user's threat to switch also changes. Switching costs as such, in the sense of the forward-looking cost of implementing the alternative technology, are irrelevant to holdup. If the cost of implementing the alternative technology is the same *ex ante* or *ex post*, any amount that could be extracted by the patentee *ex post*, because the implementer wants to avoid incurring those costs, could also have been extracted *ex ante*. Implementers become "locked in" to a standard, not because of the costs of switching, but because the expected revenue from the alternative technology will have been reduced once the original technology was adopted as part of the standard.

This has practical implications. Lemley & Shapiro (2007a) recommend that "the court should evaluate the cost that the infringing firm would have to incur to redesign its product to avoid infringing the patent. If this cost is high relative to the value that the patented technology has added to the infringing firm's product, no permanent injunction should be issued."[84] But as Denicolò et al. (2008) point out, the relevant comparison is not the cost of redesign, but the additional cost of adopting the alternative technology *ex post* as compared with *ex ante*. Looking only to the cost of redesign risks penalizing "the most valuable patents – precisely, those that are most difficult to circumvent even with full knowledge of the patent."[85] They note that instead "the policy should indicate that to avoid injunctive relief an infringer must show not only that it is costly to redesign the product in a non-infringing way ex post, but also that it could easily have designed the product in a non-infringing way ex ante if only it had been aware of [the patent holder's] patent (which again emphasizes the importance of the inadvertent infringement assumption)."[86] The point made by Denicolò et al. (2008) is consistent with the analysis provided by Cotter et al. (2018); it is not the cost of switching to the alternative that is important, but whether the cost of switching has changed.[87]

[84] Lemley & Shapiro 2007a, 2037.
[85] Denicolò et al. 2008, 596.
[86] *Id.*
[87] It should be emphasized that this analysis of the source of the differential between *ex ante* and *ex post* royalties does not imply that all of that difference constitutes undesirable "holdup." Their analysis helps identify the specific source of the differential; whether allowing the patentee to capture part of that differential is undesirable is a separate question.

6 Caveats and Critiques

A) OVERVIEW. A number of theoretical critiques of the holdup model are discussed in the remainder of this section. While most of these points were directed at Lemley & Shapiro's model in particular, several are applicable to sunk costs holdup and the apportionment problem generally, as their model is in some respects simply the best known elaboration of these general problems. There is another general critique of the holdup argument, to the effect that even though holdup might be a problem in theory, there are a number of countervailing mechanisms, such as the potential for *ex ante* bargaining, that mean it is not a substantial problem in practice. These arguments are discussed subsequently in Section 4 "Mitigating Mechanisms." The empirical evidence is reviewed in Section 8 "Empirical Evidence."

B) LITIGATION COSTS AND WEAK PATENTS. Golden (2007) argues that "for a weak infringement case for which θ is sufficiently near 0, litigation costs can again be expected to dominate the potential infringer's concerns."[88] The intuition is that the implementer's exposure due to holdup is discounted by probability of validity, while litigation costs, under the U.S. rule (each party bears its own costs), are not. Therefore, for weak patents litigation costs will dominate (so long as litigation costs are roughly independent of the strength of the patent).[89] Recall that in Lemley & Shapiro's analysis, the overcharge factor – the overcharge as a percentage of the benchmark royalty – is very high for weak patents, but the absolute amount of the overcharge may be relatively small, because the overcharge due to holdup is discounted by the probability of validity, and so is small for a weak patent. One response to this might be that litigation costs drop out of Lemley & Shapiro's formal model, as they are assumed to be symmetric.[90] But costs are not necessarily symmetric in fact, and in practice negotiations might be driven by litigation costs. In that case, the transaction cost analysis discussed below in Section 7.5.2 would be more pertinent to the potential for holdup (or holdout).[91]

C) PATENTS CENTRAL TO THE PRODUCT. Denicolò et al. (2008) say that "[w]hen the infringed patent is essential to the innovative product . . . the logic of the holdup problem changes significantly."[92] They note that "for holdup to be a significant threat not only must the patent cover a single component of a larger complex product, but that one component must be minor (ν small) and a stand-alone product

[88] Golden 2007, 2131.

[89] *See* Golden 2007, 2130–31 (discussing separately the cases in which the implementer's best strategy is to design around only if found liable, and in which the implementer would design around in any event).

[90] Lemley & Shapiro 2007b do not specifically respond to this point in their reply to Golden.

[91] *See* Section 6 "Holdout/Reverse Holdup."

[92] Denicolò et al. 2008, 593.

excluding v must have been commercially and technically feasible ex ante."[93] This is not really a challenge to Lemley & Shapiro's central point, which is precisely that holdup is especially severe for a complex product with a minor patented feature.[94] It is true that when the patent is more central to the product, the holdup in Lemley & Shapiro's *ex post* scenario is driven by sunk costs (as opposed to the loss of profits from the entire product being held off the market), and in the "early negotiation" case it is driven by the probabilistic nature of the patent. The question then is whether Denicolò et al. (2008) show that these factors do not result in holdup for essential patents. The answer is no.

To support their argument they give the example of the case in which the patentee and the implementer both have technology that is strictly complementary in the sense that both technologies are necessary to the success of the product. The proper benchmark in such a case is the royalty the parties would have negotiated prior to either sinking costs into their respective technologies.[95] If the parties negotiate *ex post*, and the patentee can obtain an injunction in the case of breakdown, their positions will not have changed much, since either will be able to block the project. The difference is that both will have sunk R&D costs into their technologies, but if those costs are similar, and the bargaining power does not change, then the *ex post* bargain will be the same as the *ex ante* bargain.[96]

This argument is evidently directed primarily at the "early negotiation" scenario in which sunk costs are the driver of holdup. While their example is correct so far as it goes, it is not strong support for their proposition. First, there is no particular reason to believe that the R&D costs will generally be similar. An example that approximates the situation they describe is *NTP* v. *Research in Motion*.[97] NTP had a patent on a technology essential to RIM's principal products, but RIM had spent substantial amounts implementing the technology, and there is no reason to believe that NTP's patent, which was a paper invention never commercialized by the inventor or NTP,[98] had been particularly costly to develop. No doubt there are cases where the patentee's R&D costs are roughly on the order of the implementer's technology-specific sunk costs, but that does not justify granting an injunction in cases like *NTP* v. *RIM*, simply on the basis that NTP's technology was essential to the product. The centrality of the patented technology to the product is not a good proxy for symmetry of investment between the patentee and implementer.

[93] *Id.* at 596.

[94] Lemley & Shapiro 2007a, 2001 (noting high holdup for lost profits during redesign "[f]or a complex product and a minor patented feature"); *Id.* at 2002–03 (noting the holdup potential when "the patented feature is nothing special").

[95] Denicolò et al. 2008, 594. The benchmark they give is equivalent to the Shapley pricing solution advocated by Siebrasse & Cotter 2017a.

[96] Denicolò et al. 2008, 593–94.

[97] *NTP, Inc.* v. *Research in Motion, Ltd.* (E.D. Va. 2003) (U.S.).

[98] Lohr 2010.

Secondly, the example of an implementer that has a strictly complementary technology is largely unrelated to the scenario in which the infringed patent is essential to the innovative product. Denicolò et al. (2008) argue that Lemley & Shapiro are wrong to focus on the implementer's sunk costs without considering the costs that the patentee had sunk into R&D.[99] This reflects the "true *ex ante*" argument discussed above. But how does this generalize to a case in which the patentee has a patent that is essential to the product? They say that "since both firms must sink a specific investment before they can contract, both may actually be subject to a hold up problem."[100] But that is true only if the patentee has no option other than to negotiate with that particular implementer. This emerges from their model because they assume that the patentee and the implementer have strictly complementary technologies. But that is a special case. As Denicolò et al. (2008) themselves point out, if the implementer market is perfectly competitive the patentee will be able to extract the full value of the invention. At the other extreme, if there is a monopsony in the implementer market, then the implementer does indeed have additional leverage, on standard monopsony pricing theory. But that arises from the structure of the implementer market, not because the patented technology is essential or otherwise to the product. In effect, Denicolò et al. (2008) are arguing that when the patentee has a patent that is essential to the product and the implementer is a monopsonist, the patentee should be entitled to an injunction in order to counterbalance that monopsony power. But recall that they are arguing that holdup is only significant when the patent covers a single component of a larger complex product, and one component is minor, and a stand-alone product was commercially and technically feasible *ex ante*, or, more generally, the infringed patent is essential to the innovative product. It is not clear how any of these are related to a case in which the implementer has monopsony power, whether because it has complementary technology, or for some other reason.

A model of parties with proprietary rights to complementary technologies is entirely appropriate when discussing multiple patentees with patents reading on a product sold by an implementer, as is notoriously the case with SEPs. This does indeed raise a difficult question of how to allocate royalties, and whether any party should be entitled to an injunction. It is not uncommon that one of those patentees with complementary technology might also be an implementer, but it does not follow that all patentees should be entitled to injunctions against all implementers in order to give them appropriate leverage against a particular implementer that happens to also be a patentee.[101]

[99] Denicolò et al. 2008, 594.

[100] *Id.*

[101] Denicolò et al. 2008, 595, also dispute the assertion by Lemley & Shapiro that the magnitude of the holdup problem increases approximately linearly with the number of infringed patents; they conclude instead that the increase in holdup is less than linear. That point is discussed in more detail below in Section 7.7.2 "Cumulative Effect of Holdup." In the present context, their point is

D) MARKET STRUCTURE. The Lemley & Shapiro model assumes a patentee negotiating with a single downstream firm, and while they make some observations respecting markets with multiple downstream firms, they acknowledge that a thorough discussion is beyond the scope of their article.[102] Elhauge (2008) argues that "there is every reason to think the results are totally different if the downstream market is competitive."[103] The gist of his argument seems to be that in a competitive market the patentee will extract the entire expected value of the invention,[104] and so there cannot be any overcharge because an implementer would prefer to exit the market entirely.[105] He then argues that the royalty the patentee can charge is constrained to no more than $v\theta$:

> Assuming damages are properly set at v times X_i [number of units sold] for any infringing seller, the expected damages for infringement will be $v\theta X_i$. Thus, if the patent owner tried to charge a royalty of more than $v\theta$, all the downstream firms would decline the license because they would incur expected losses from agreeing.[106]

That is not correct, or at least it is overly simplistic. This statement is addressed at the early negotiation scenario, and in that case the implementer's threat point is to use the best noninfringing alternative. Suppose the value of the patent is reflected in a cost saving, and the patentee negotiates with one implementer. If the other implementers do not take a license, their costs will be higher than that of the licensee by v (by the definition of v). The licensee can afford to pay more than $v\theta$ and still undercut the other implementers. On the other hand, if the patentee makes the same offer to all implementers simultaneously, we are essentially back in the scenario of a single downstream firm. If the patentee demands more than v, they will all prefer to use the alternative, but otherwise they will be willing to pay more than $v\theta$, because they all have to pay the same amount, and so all will earn the same zero economic return that is standard in a competitive market.

Elhauge then says, "[e]ven if the downstream firms had already used the technology inadvertently, the patent owner could not charge more than $v\theta$ by trying to holdup the downstream firm for some of the costs of redesign, because if it did so the downstream firm would expect to lose money and prefer to exit the market."[107] If the

apparently that stacking will not result in significant holdup if the patents involved are not essential to the product. Given my discussion in the text, this point need not be addressed here.

[102] Lemley & Shapiro 2007a, 2005–08.

[103] Elhauge 2008, 561.

[104] *Id.* at 562 (noting that the patentee can play one implementer off against another, so that in effect "$\beta = 1$ if the downstream market is competitive"). This is subject to the point that splitting the value of the invention may amount to paying the implementer for product-specific services, in which case the value would be split even in a competitive market; but the implementer's expected profit would still be zero.

[105] *Id.*

[106] *Id.*

[107] *Id.*

implementers were not aware of the patent *ex ante*, their expected profit in the competitive market would be zero, and all the cost saving of the technology would be passed on to consumers. If the patentee then emerges, *any* positive royalty, even a royalty of less than $v\theta$, will result in the implementers losing money, unless they raise their prices. If the patentee approaches only one implementer, it will lose money if it takes a license and none of the others do, so it will exit the market and the patentee will get no revenue. If the patentee then approaches another implementer, this process will continue until there are so few implementers left that the market is no longer competitive and the remaining implementer can take a license and raise its prices. In effect, by selectively licensing, the patentee will have transformed a perfectly competitive downstream market into an imperfectly competitive market. There may be circumstances in which that strategy would be rational, but on this route we are no longer dealing with a competitive market, so Elhauge's point would not apply. Alternatively, the patentee might license all implementers at the same royalty, in which case each implementer could raise its price by the same amount without losing its market. Each implementer would be willing to pay the royalty and stay in the market (strictly, it is indifferent between leaving and staying in the market, but that was also true under *ex ante* negotiations) until the royalty is so high it would be preferable to redesign the product – which is the standard point that the implementer can be held up for the redesign costs. The implementers would lose money, but they would lose less money than if they left the market; that is the standard sunk costs holdup result.[108] The only real difference is that the implementers cannot be held up for lost economic profits during the redesign phase, because they are not making any economic profits. But if they are making accounting profits because they have fixed costs, they could be held up for those profits.

This is not to say that the market structure does not affect Lemley & Shapiro's result at all; a thorough discussion is beyond the scope of this chapter, just as it was beyond the scope of Lemley & Shapiro's original article. But Elhauge's critique does not give any reason to think that the basic result does not extend to different market structures.

E) ELASTIC DEMAND. Elhauge (2008) asserts "the Lemley–Shapiro model would overstate royalties because it assumes inelastic output."[109] It is true that inelastic demand is a dubious assumption. It is also true that the overcharge will be less when demand increases in the presence of the patented technology; the intuition is that when the patented technology adds value to the product, the implementer will

[108] *Id.* at 563 (noting that "[t]he same is true if the market downstream is marked by recurring fixed costs or product differentiation, making models of "monopolistic competition" more appropriate," and the same counterargument is applicable).

[109] *Id.* at 547 ("Third, even with the above problems, their assumption of inelastic output is unrealistic and inflates predicted royalties"); *id.* at 551 ("[T]he Lemley–Shapiro model would overstate royalties because it assumes the downstream output X is constant and totally unaffected by whether D incorporates a patented feature that increases product value.").

normally get value from the patent in the form of increased sales, as well as in the form of a higher price, and the increased profit from increased sales partially offsets the overcharge. But Lemley & Shapiro's model does not turn on the assumption of inelastic output. That is merely an example they provide by way of illustration.[110] Elasticity of demand will mitigate the overcharge problem to some degree, but it seems unlikely to provide significant relief in the context of complex products, where thousands of patents may read on a single product.

7 Competing Patentees

Lemley & Shapiro state that their analysis is limited to situations in which the patentee's predominant commercial interest in bringing a patent infringement case is to obtain licensing revenues and it does not apply to settings in which the patent holder practices the invention and seeks to use the patent to exclude a competitor from the market in order to preserve its profit margins.[111] Golden (2007) and Elhauge (2008) argue that Lemley & Shapiro's distinction is not compelling, and they indicate that even patentees seeking only royalties should be entitled to injunctive relief.[112]

The holdup problem faced by the implementer is just as severe whether the patentee competes in the market or not.[113] The key question is therefore whether there are countervailing considerations that imply that a patentee who competes in the market should be granted injunctive relief notwithstanding these holdup concerns.

Lemley & Shapiro's model considers only reasonable royalty damages, and they equivocate when considering a patent holder who would be entitled to lost profit damages, saying in cases involving "significant" lost profits, they favor a presumption that the patent holder will be granted a permanent injunction,

> perhaps with a stay to allow the infringing firm to redesign its product. The presumptive right to a permanent injunction in these cases is justified in part for reasons of equity and in part because of the grave difficulties associated with calculating and awarding lost profits on an ongoing basis.[114]

[110] Lemley & Shapiro 2007a, 2046 Appendix – A.

[111] Lemley & Shapiro 2007a.

[112] After making the point that there is no evident basis for distinguishing between a patentee who seeks lost profit and one who seeks reasonable royalties, Golden 2007, 2155 asks "[w]hy not simply curtail injunctive relief for all patent holders?," but he appears to be asking the question rhetorically.

[113] *See* Denicolò et al. 2008, 588–89.

[114] Lemley & Shapiro 2007a, 2036. Lemley & Shapiro 2007b, 2171–73 also address this point, but they do not develop the substantive rationale for the distinction beyond saying that what matters is "the nature of the patent holder's contribution and how it seeks compensation in the marketplace." However, it is not clear exactly what Lemley & Shapiro mean by "the patent holder's contribution" or why it should vary systematically between practicing and nonpracticing entities, and they do not elaborate on why the availability of injunctive relief should depend on how the patentee seeks compensation in the

This suggests that the costs of denying injunctive relief, in the form of increased error costs of damages calculation, are greater in the context of lost profits.[115] There are two problems with this argument.

First, while it is no doubt difficult to assess lost profits on an ongoing basis, it is not easy to accurately quantify a reasonable royalty either. It is not evident that lost profit calculations are generally so much more difficult than reasonable royalty calculations, particularly in the case of complex products, as to justify a sharp distinction between cases in which the patentee is seeking lost profits and those in which it is not.[116]

Moreover, Golden (2007) points out that the difficulty in assessing reasonable royalty damages has traditionally been one of the principal rationales for granting permanent injunctions.[117] Lemley & Shapiro respond by noting that "all that is required for reasonable royalties to play their role in guiding parties to a negotiated settlement in the shadow of litigation is that they be unbiased."[118] But this same point undermines their distinction between reasonable royalties and lost profits; even if lost profits are more difficult to assess, that makes no difference so long as the errors are unbiased. The important question is not whether lost profit damages are more difficult to assess than reasonable royalties, but whether they are more likely to be biased against the patentee. There is no obvious reason why errors in lost profit damages are less likely to be unbiased than reasonable royalty damages.

Apart from the relative accuracy of the two types of damages, Elhauge (2008), and Denicolò et al. (2008) suggest that the holdup problem might be worse when the patentee is able to seek lost profit damages because it competes in the downstream market. In that case the patentee may hold up the implementers even more because higher royalties provide it with increased market share in the downstream market, as well as directly benefiting from high royalties itself.[119] In effect, the patentee has increased bargaining power when it competes in the downstream market; when it

marketplace. Shapiro 2010, 304, similarly adverts to the difficulty of determining lost profits on a forward-looking basis.

[115] Lemley & Shapiro's reference to "equity," is obscure. They do not refer to any particular equitable principles, which suggests they mean equity in the sense of fairness rather than equity as a legal term of art, but neither do they elaborate on any relevant fairness intuitions.

[116] Golden 2007, 2155. Reasonable royalties are often based on comparable licenses, but as Lemley & Shapiro 2007a themselves point out, at 2022, information about comparable licenses is limited and biased. *See also* Masur 2015 (explaining the difficulties associated with assessing reasonable royalties based on comparable licenses).

[117] Golden 2007, 2152.

[118] *See* Lemley & Shapiro 2007b, 2172. By the same token, the incentive to innovate is maintained if lost profit damages are accurate and unbiased. Lemley & Shapiro give no reason to think that lost profit damages are less likely to be unbiased than reasonable royalty damages. *Id.* (acknowledging that sometimes nonpracticing entities should be able to get injunctive relief and vice versa, but their examples are tied to whether the entity suffers lost sales, which begs the question of why that should be a determinative factor).

[119] Elhauge 2008, 560–61; *see also* Denicolò et al. 2008, 588–89. Elhauge views this point as a criticism of Lemley & Shapiro's model, but it is more properly viewed as an extension.

only licenses, the royalty is constrained because it will make nothing if the royalties are so high as to unduly restrict sales, but if the patentee competes in the downstream market that constraint is lifted, as the patentee might anticipate capturing those sales itself.

A distinct reason for preferring injunctive relief in the case of a patentee that practices the invention is that in such a case we should expect the patentee to be more efficient, because if the infringer were more efficient than the patentee, the patentee would have been willing to license. Allowing the patentee to exclude the infringer in such circumstances gives the market to the more efficient producer.[120] However, granting an injunction to a nonpracticing patentee should have the same effect, and the patentee would license to the more efficient producer.

As a final point on this issue, Geradin (2010a) argues that Lemley & Shapiro's distinction between patentees seeking lost profits and those seeking reasonable royalties "would unduly affect innovators which have opted for a licensing business model for perfectly legitimate reasons, such as for instance the fact that they do not have the skills or the resources to develop and manufacture products embedding their technologies," and "effectively tip the market in favor of vertically-integrated incumbents . . . [, which] would impede efficiency-enhancing specialization allowing firms to focus on what they do best and harm innovation."[121] However, it is not clear that the lost profit damages per unit, properly assessed, will be greater than the royalties per unit. That will only be true, in an economic sense, if the vertically integrated firm has the capacity to satisfy the market and is a more efficient producer than the implementer, in which case it is not inefficient to give the patentee extra leverage against the implementer. If the vertically integrated firm is actually worse at commercializing the invention, this implies that its lost profits will be less than the reasonable royalty it could have obtained from licensing to a more efficient implementer; as Geradin (2010a) points out, the innovator is more likely to opt for a licensing model when it does not have the skills or resources to manufacture the product, and the lower return from a royalty reflects these shortcomings. While this follows as a matter of economic logic, it requires that the lost profits calculation properly accounts for the patentee's costs of production, including fixed costs. If lost profits calculations are excessively generous to the patentee, then the vertically integrated patentee will have greater leverage because of the excessive damages for past infringement whether or not it is granted an injunction. As discussed above, the implementer's share of the surplus may be best understood as compensation for its investment in the success of the product, through marketing and distribution, etc., which would represent costs to the patentee.

In summary, despite their protestations, Lemley & Shapiro's holdup model does prima facie apply to patentees who compete with the infringer. This does not imply

[120] Blair & Cotter 1998, 1626–28.

[121] Geradin 2010a, 126–27. To the same effect, *see also* the second point made by Elhauge 2008, 561.

that their model should be rejected, but it does suggest that their model is incomplete,[122] and/or that the holdup problem needs to be taken seriously in that context as well.

7.4 MITIGATING MECHANISMS

7.4.1 *Introduction*

There are a variety of mechanisms that have been suggested as being effective in mitigating the effects of holdup in a variety of contexts. This section discusses the theoretical plausibility of those mechanisms. Whether they are effective to mitigate the effects of holdup (if any) is an empirical question that is discussed below in Section 8 "Empirical Evidence."

It is sometimes suggested that holdup is not a serious problem in practice because legal constraints, such as the FRAND commitment or oversight by competition authorities, are effective in preventing abuse of patent power.[123] While this may be true, it is not helpful to consider such legal constraints to be relevant mitigating mechanisms. The ultimate question is how to interpret the FRAND commitment when faced with a decision as to whether to grant injunctive relief, and to say the FRAND commitment helps prevent abuse tells us nothing about how to interpret that commitment. If anything, the implicit suggestion that the FRAND commitment and competition law oversight are necessary implies that holdup would be a problem in their absence.

7.4.2 Ex Ante *Licensing*

If *ex ante* licensing is possible, the holdup problem is substantially mitigated.[124] In the SEP context the dominant view appears to be that licensing prior to the standard being adopted is rare and generally impractical,[125] though it does appear that *ex ante*

[122] *See* Section 7.5 "Property Rules and Liability Rules" (suggesting that the theory presented by Smith 2004 might provide the basis for a distinction between the two scenarios).

[123] *See, e.g.*, Nokia Corp. 2011 (stating that "[e]specially for complex standards as in telecoms, Nokia believes that (F)RAND is the only workable solution to prevent patent hold up"); Denicolo et al. 2008, 597 n.80 (referring to Rambus's attempt to hold up its licensees, which was struck down by the FTC).

[124] *See* Lee & Melamed 2016, 460–61. However, it is not necessarily eliminated entirely. As Lemley & Shapiro 2007a show, *ex ante* licensing only avoids holdup entirely for ironclad patents. For probabilistic patents, holdup may occur even with *ex ante* bargaining: *see* above Section 3.1.4 "Probabilistic Holdup: Lemley & Shapiro Model." Further, royalty stacking is not addressed by *ex ante* licensing as such.

[125] *See* Contreras 2013, 59 (stating that "*very few* [FRAND] licenses are negotiated prior to market adoption") (emphasis in original); Intel Corp. 2011, 9 (stating that "ex ante licensing is unlikely to occur in the most common licensing scenarios: those involving new technologies, new product markets, and/or early versions of standards"); Nokia Corp. 2011, 6 (stating that in the telecoms

licensing is at least occasionally possible, and the view is also sometimes expressed that it is common and generally feasible.[126] The extent to which *ex ante* licensing is feasible is an empirical question that this chapter cannot resolve.[127]

Outside of the SEP context whether *ex ante* negotiations are possible will depend on the ability of an implementer to undertake an effective preclearance search. For some types of products, preclearance searches may be generally feasible and costs-effective. However, for complex products, effective *ex ante* negotiation may be even more difficult than in the SEP context. In the SEP the development of the standard will be well-known to those in the industry, and the identity of the patentees will be known, and the hurdle to *ex ante* negotiations is primarily the cost and delay associated with actually negotiating the agreements.[128] For equivalently complex products outside the SEP context, implementers will face the same difficulty, plus the additional burden of actually identifying all the relevant patents.[129]

7.4.3 Ex Ante *Validity Challenge*

Denicolò et al. (2008) note that in Lemley & Shapiro's model, holdup can occur even if the implementer had the opportunity to negotiate *ex ante*, because of the probabilistic nature of the patent. They argue that this result is based on the

environment "[i]t is simply not possible to determine a meaningful value/price long before it is known what kind of products will eventually implement the standard").

[126] Qualcomm Inc. 2011, 11 (stating that Qualcomm entered into *ex ante* WCDMA licenses with firms representing more than 60 percent of royalty-bearing unit sales in 2005); Epstein et al. 2012, 17–18 ("Manufacturers can, and do, engage in bilateral patent licensing before seriously investing in patented technology, both in settings in which SSOs are deployed and those in which they are not" (citing Qualcomm Inc. 2011, 8)); Geradin & Layne-Farrar 2007, 91 ("[V]oluntary ex ante disclosure of licensing terms by IPR owners and ex ante negotiations of license agreements with IPR owners are already regular occurrences" (citing Holleman 2002, 2)); Geradin 2010a, 111 (stating that "the majority of key patent owners and standard implementers commonly engage in ex ante licensing negotiations – that is, they routinely negotiate patent portfolio licenses or cross-licenses pertaining to an anticipated standard, or to a standard under development, well before the standard is finalised," though without citing supporting sources); Microsoft 2011, 14 (noting that potential implementers can sometimes negotiate with SEP holders before the standard is finalized).

Ganglmair et al. 2012, 251 n.5–6 assert that "[o]ption contracts have been shown to be a robust solution to hold-up problems," and that "[c]ontracts with an option feature were used by Qualcomm with its innovative CDMA technology for mobile telephony"). However, this is effectively a type of *ex ante* licensing, as it requires entering into an option-to-license contract before the implementer incurs sunk costs, *id.* at 252, so the feasibility of this solution turns on the feasibility of *ex ante* negotiations.

[127] *Ex ante* negotiation is clearly not possible if the implementer only enters the industry after the standard has issued. *See* Gilbert 2011, 860. In principle the "non-discrimination" branch of the FRAND requirement would protect against holdup in such circumstances, though in practice it might not be possible for the late entrant to find out the terms that were offered to others.

[128] *See* Contreras 2013, 59–62 (explaining the practical factors making *ex ante* negotiations difficult in the standards context).

[129] *See* Lee & Melamed 2016, 405–08; Kieff & Layne-Farrar 2013, 1105–08 (suggesting that *ex ante* licensing is often possible if implementers exercise due diligence, though not always).

assumption that the implementer cannot contest the validity of the patent before designing its product.[130] However, this critique turns on the validity challenge being costless and immediate. Denicolò et al. (2008) assert that "similar conclusions also hold with costly litigation,"[131] but it is not clear that this is true. Apart from the cost, litigation takes time, and the point made by Lemley & Shapiro is that the implementer can be held up for redesign costs and lost profits on its product during the period of redesign, because the implementer's threat point in the negotiation is not to use the invention at all. All of these sources of holdup will arise unless the validity can be determined before the implementer begins to produce the product. If the implementer holds off on selling until validity is decided, it can be held up for the opportunity cost of its foregone profits during that period. Lemley & Shapiro's model gives the same results whether the litigation is assumed to be an infringement action by the patentee, or a declaratory judgment action by the implementer.

7.4.4 *Norms*

Elhauge (2008) suggests that even in a single-shot game, fairness-based norms may help prevent or mitigate excessive royalties by making the implementer's threat to reject such royalties credible.[132] However if fairness norms anchor negotiations even between sophisticated parties, that can only lead to a fair royalty if the norm itself is fair. He says that "[i]f parties believed that $\theta\beta v$ was the fair benchmark, as Lemley and Shapiro argue, then they are likely to refuse royalties above that, making royalties even more undercompensatory."[133] Given that Elhauge is of the view that $\theta\beta v$ is unfair, it is not clear why he believes that it would be adopted as the fairness norm.

More generally, as discussed at the outset of this section, the theory of how parties to a negotiation split the gains to trade is incomplete, and it is certainly possible that fairness norms play a role. But if fairness norms are thought only to influence β, that would affect the degree of holdup – one way or the other – but it would not affect the fact of holdup, unless the fairness norm is so powerful as to displace the standard assumption that the parties negotiate in the shadow of the litigation outcome. A much more substantial argument than is provided by Elhauge would be required to make either point.

7.4.5 *Repeat Play*

Lemley & Shapiro's model of holdup considers a one-shot game. Elhauge (2008) argues that if negotiations over patent royalties are repeated between an

[130] Denicolò et al. 2008, 590, suggest that the implementer might contest the validity when it is aware the patent is weak, but the general point applies regardless of the strength of the patent.

[131] *Id.*

[132] Elhauge 2008, 549–51.

[133] *Id.* at 550–51.

implementer and multiple sequential patent holders, the equilibrium royalty will be lower than the rates predicted by Lemley & Shapiro, essentially because the implementer can improve its bargaining position by developing a credible reputation as a hard negotiator.[134] Elhauge argues that the bargaining is more appropriately modeled as being between an implementer and multiple patentees because the implementer of a complex product necessarily faces multiple patentees, and it will therefore be in the implementer's interest to develop a reputation as a tough negotiator. However, the conclusion that a repeated game will lead to a lower royalty does not appear to be robust to the details of the way in which the game is modeled. For example, similar reasoning suggests that if the negotiations took place between a single patentee and multiple implementers, the royalty might be higher than the rates predicted by Lemley & Shapiro, because the patentee can improve its bargaining position by developing a credible reputation as a hard negotiator. And in many cases, it will be realistic to model both parties as repeat players, as when negotiations are between NPEs with a large portfolio and large implementers who are often targeted by NPEs.[135] On the other hand, patents may be asserted by a special-purpose entity formed solely to assert a single patent portfolio, which is, by definition, not a repeat player, and does not have a market reputation to defend.[136] Also, Elhauge's formal model considers an implementer and multiple sequential patent holders, not simultaneous patent holders, and it is not obvious that there will be a reputational effect when the negotiations are simultaneous. On the whole, there is little doubt that repeat play and reputation effects can have a significant effect on bargaining outcomes, but it is difficult to generalize about exactly what that effect might be.

1 Modified Injunction

A) STAY OF INJUNCTION. A modified injunction may mitigate holdup problems. Lemley & Shapiro (2007a) recommend that if the cost of designing around the patent is moderate or low, the permanent injunction be granted with a stay that is long enough to permit the infringing firm to complete the redesign.[137] The option of a stay is attractive because it reduces the risk of holdup in cases where redesign is not too costly, while also minimizing the risk of undercompensation, because even if damages are undercompensatory, the marginal effect of that undercompensation is felt only during the period of the stay.[138] This option has at least occasionally been

[134] *Id.* at 547–49.

[135] *See* Qualcomm Inc. 2011, 25–26 (suggesting informally that demands will be moderated if both parties are repeat players).

[136] Chien 2014, 31.

[137] Lemley & Shapiro 2007a, 2038.

[138] Of course, damages for pretrial infringement may be undercompensatory, but this is not affected by staying the permanent injunction. Denicolò et al. 2008, 602–03 accuse Lemley & Shapiro of ignoring litigation delays in making this suggestion for a stay, saying patent infringement cases "can take years to wend their way through the courts." This criticism conflates the effect of the stay with the

employed by U.S. courts,[139] but apart from the recommendation by Lemley & Shapiro, it has not featured prominently in the scholarly literature.[140] A stay will not be effective in preventing holdup where the cost of redesign is high relative to the value of the invention.[141]

B) PATENTEE PAYS SWITCHING COSTS. Lee & Melamed (2016) propose a novel form of modified injunction. They distinguish between a willing licensor, who would have been willing to license to the infringer, and unwilling licensors, which includes both patent holders who wanted to practice the patents themselves, as well as those who wanted to license a limited number of others, and so would not have been willing to license the infringer.[142] They propose that an unwilling licensor should generally be able to obtain an injunction against a "guilty" infringer, who could in practice have entered into *ex ante* licensing negotiations, thereby avoiding any holdup problem. In a case involving an unwilling licensor and an "innocent" infringer, who could not as a practical matter have negotiated *ex ante*, they propose as a prospective remedy that the licensor be provided with a choice between an ongoing royalty[143] and an injunction, but the injunction would be available only on the condition that the patentee would pay the infringer's cost of switching to a noninfringing alternative.[144]

This type of injunction protects the implementer from holdup based on switching costs even more effectively than a stay because the patent holder rather than the implementer would bear the costs of switching. As a result, their proposal would protect the implementer even when switching costs are high relative to the value of the invention.

One caveat is that Lee & Melamed do not specify whether a stay of the injunction would also be granted to the implementer.[145] If not, the implementer might be subject to holdup based on lost profits on the product during the redesign period, as argued by Lemley & Shapiro (2007a). It may be that Lee & Melamed would consider such lost profits to be part of the cost of switching, in which case it would

independent effect of litigation delay; it is more properly directed at the U.S. practice of rarely granting preliminary injunctions.

[139] *See* FTC 2011, 238 (citing *i4i Ltd. Partnership* v. *Microsoft Corp.* (Fed. Cir. 2010) (U.S.)).

[140] *But see* Shapiro 2016, 27 (reiterating the stay recommendation).

[141] Lee & Melamed 2016, 458 n.332. In that case, Lemley & Shapiro 2007a, 2036, recommend denying the permanent injunction entirely.

[142] Lee & Melamed 2016, 445.

[143] The royalty would be at the same rate as past compensatory damages. If the patentee would have been entitled to lost profits for past infringement, the ongoing royalty would be at the same rate; otherwise, it would be equal to a reasonable royalty: *id.* at 445.

[144] *See id.* at 390, table 1 (summarizing their proposal); *id.* at 457–60 (discussing the proposal in more detail).

[145] Lee & Melamed 2016, 458 n.332 discuss the possibility of a stay without mentioning it as part of their proposal. This indicates that under their proposal injunctive relief would not be conditioned on a stay.

either be borne by the patentee, or the patentee would agree to a stay voluntarily to avoid having to bear those costs.

Their proposal also captures the intuition that a patentee who practices the invention, and so would normally be entitled to lost profits, should have a stronger entitlement to injunctive relief; but their distinction between willing and unwilling licensors avoids the difficulties associated with distinguishing competing patentees as such.[146]

7.5 PROPERTY RULES AND LIABILITY RULES

7.5.1 *Inaccuracy of Damages Awards*

In their landmark article, Calabresi & Melamed (1972) introduced the now standard distinction between property rules and liability rules. A property rule, in which an entitlement is protected by injunctive relief, gives an individual the right to keep an entitlement unless he chooses to part with it voluntarily. In contrast, if the entitlement is protected by a liability rule, the owner of the entitlement must give it up to another who is willing to pay its fair value, as objectively determined by the court. According to Calabresi & Melamed, the disadvantage of a property rule is that it allows the owner of the entitlement to hold out for an excessive price when there is market failure; the corresponding advantage of a liability rule is that it avoids such holdup.[147] Conversely, the advantage of a property rule is that the owner of the entitlement determines its value, and having the court assess the value of the right, as under the liability rule, is inherently less accurate. This implies that the decision as to whether the patent holder should be granted an injunction turns on whether the holdup problem is worse than the valuation problem.[148]

This analysis transfers directly to the context of patents for complex products. Golden (2007) points out that "[t]he difficulty of assessing a reasonable royalty has in fact been one of the principal rationales for granting permanent injunctions."[149] He notes that "[t]he difficulty of assessing even a retrospective reasonable royalty is notorious," and expert evidence may differ by an order of magnitude.[150]

[146] *See* Section 3.1.7 "Competing Patentees."

[147] Calabresi & Melamed 1972, 1107–08, refer to the "holdout" problem, and their examples turn on collective action problems rather than sunk costs, but their insight applies whenever voluntary bargaining does not result in an exchange based on the true value of the right, and so encompasses what is referred to as "holdup" in the patent context.

[148] *See* Epstein 1997, 2094 ("Stated formally, the task of a legal system is to minimize the sum of errors that arise from expropriation and undercompensation, where the two are inversely related.").

[149] Golden 2007, 2152.

[150] *Id.*, 2150–51 (also noting that the difficulty is compounded in assessing a reasonable royalty going forward, where the market for the invention may be permanently distorted by the infringement); *see also* Cotter 2013a, 54–56 (arguing that the difficulty of accurately valuing patent rights is an important justification for granting injunctive relief).

However, there are a variety of other possible justifications for the use of property rules apart from inaccuracy of damages, and the argument based on inaccuracy of damages is itself problematic. These points are discussed in turn below.

7.5.2 *Transaction Cost Arguments*

One solution to the puzzle is that injunctive relief might be justified on a variety of other grounds broadly related to transaction costs. Injunctive relief might save the litigation costs associated with quantifying damages; reduce administrative costs associated with judicial supervision of the ongoing royalties; encourage development of transaction cost-reducing institutions; provide an incentive to avoid litigation in the first place; and/or avoid the risk of the implementer otherwise holding out through manipulation of delays in the litigation system.[151]

The problem with this solution is that these second-order arguments require a difficult empirical assessment of the relative severity of these various factors if they are to serve either as a normative basis for recommendations regarding injunctive relief, or as a descriptive theory of current trends and practices. For example, courts are far more likely to grant injunctive relief to a patentee that competes with the infringing firm.[152] As discussed above, Lemley & Shapiro argue that such a preference is justified in order to avoid the costs of damages calculations,[153] but it is far from clear that this justifies a distinction between patentees who would be entitled to lost profits and those entitled only to a reasonable royalty, as it is not evident that there is a substantial difference in the difficulty of the two calculations.[154]

7.5.3 *Generating Information Regarding Potential Use*

Smith (2004) argues that the basic flaw in the pro-liability rules literature is the assumption that the underlying risk distribution is known.[155] He argues that the problem with liability rules is not so much undercompensation for loss of known uses, but failure to compensate the owner of the entitlement for uses that are themselves speculative.[156] While Smith presents this as an argument in favor of

[151] *See* Cotter 2009, 1175–76 (reviewing a variety of justifications for injunctive relief, while nonetheless stating that the valuation advantage is the "first" reason for preferring injunctive relief); *see also* Cotter 2013a, 54–56.

[152] *See* Seaman 2016, 1990, figure 4.

[153] Lemley & Shapiro 2007b, 2172; *see also* Kaplow & Shavell 1996, 741–42 (making a similar point in the context of the general debate about the proper use of property and liability rules).

[154] *See* Section 3.1.7 "Competing Patentees."

[155] Smith 2004, 1721–22.

[156] That is, the value of a property right depends on the range of its potential future uses, as well as the expected value of each of those uses. If some potential future uses are not known to the court, then the assessment of the value of the right will be inaccurate, even if the assessment of the value of the known uses is correct in expectation. Since any positive value potential future use (as opposed to

property rights generally, it does not particularly support injunctive relief for patent infringement in contexts in which the patentee would have been willing to license. In such cases the use itself is known, and any uncertainty relates only to the value of the use. Smith's argument might be relevant to the question of whether there is a sound distinction between a patentee who exploits the patent by practicing the technology and one who seeks only to license it.[157]

7.5.4 *Inaccuracy of Damages Assessment*

1 Inaccuracy v. Biased Damages

It is largely uncontroversial that the assessment of damages for patent infringement is likely to be inaccurate, as Lemley & Shapiro acknowledge in their reply to Golden.[158] They say, however, that

> all that is required for reasonable royalties to play their role in guiding parties to a negotiated settlement in the shadow of litigation is that they be unbiased, so that deviations from the benchmark royalty are not systematic one way or the other.[159]

This exchange reflects a similar debate in the general literature on property rules versus liability rules. There has been a substantial literature responding to Calabresi & Melamed, arguing that liability rules are superior to injunctive relief in a range of circumstances, to the point that "[p]roperty rules find relatively few defenders among legal economists."[160] Smith (2004) points out that the pro-liability rule literature turns on two basic assumptions: that the risk distribution is known; and that errors in judicial determination of damages are unbiased.[161] Lemley & Shapiro's response to Golden reflects the second assumption in particular.

a potential liability), will necessarily increase the expected value of the property, failure to take into account a potential use will result in an assessment of the expected value of the right that is biased downward. Property rules, according to Smith, solve this problem by giving the owner of the right a generalized entitlement to all future uses. This allows the owner to assess the potential future uses herself, without having to convince a court. A second advantage is that a potential use that is not known even to the owner of the right cannot affect the value of the right even under a property rule, but a property right gives the owner an incentive to investigate and discover potential uses, whether or not the value of those potential uses can be proven to a court.

157 *See* Section 3.1.7 "Competing Patentees." Smith 2007, applies his theory to various issues in intellectual property law, and at 1781–82, discusses the standard for injunctions in patent law, but the discussion is so brief as to add little on this issue to his general theory. Smith does not argue that damages assessments are indeed unbiased; his theory is an alternative justification for property rules, which is not necessarily inconsistent with the view that damages are generally undercompensatory.

158 See, for example, Judge Learned Hand's observation in *Cincinnati Car Co.* v. *New York Rapid Transit Corp.* (2d Cir. 1933, p.595) (U.S.) (quoted by Golden 2007, 2123, 2152) that assessment of the patentee's loss "is really incalculable" and a damages assessment can be no more than an approximation.

159 Lemley & Shapiro 2007b, 2172.

160 Smith 2004, 1721–22. *See also* Kaplow & Shavell 1996; Smith 2004, 1741–48 (reviewing the literature).

161 Smith 2004, 1725–26, 1746.

If the courts can reliably award damages that are equal to the loss suffered by the patentee, at least in expectation, the basic argument for injunctive relief would be much weaker, as the assessment of damages would perfectly compensate the patentee while avoiding the holdup problem. The puzzle is that this proves too much: If damages are accurate in expectation, then even a slight possibility of holdup would be enough to warrant denying injunctive relief, given that there will be no impact on the incentive to invent. One response to this puzzle is to say that injunctive relief is justified on the basis of the transaction cost arguments or Smith's theory, discussed in the preceding sections.

Another response is to posit that damages are systematically undercompensatory, apart from any feedback effects from holdup and the availability of injunctive relief, that would tend to make damages overcompensatory.[162] If so, this would not in itself imply that injunctive relief should routinely be granted. Indeed, that would provide a ready explanation, at least in principle, for the observed pattern of injunctive relief in patent cases. Ever since Calabresi & Melamed, the pro-property rights literature has acknowledged that a liability rule is justified when there is a serious risk of holdup. The shift in patent law can be reconciled with the traditional dominance of property rules, and traditional property rights theory, on the basis that shifting realities, such as the rise in patent NPEs and SEPs, and perhaps also a general increase in patents for complex products, have substantially increased the circumstances in which there is a serious risk of holdup.

If damages are systematically undercompensatory, the difficulty is not conceptual, but practical. As Lemley & Shapiro point out, "all advantages are comparative." They argue that "since, as we have demonstrated, injunctive relief will systematically overcompensate patent owners in component industries, there is a strong reason to prefer damages rules in those cases."[163] But this observation cuts both ways. Even if it is true that injunctive relief will systematically overcompensate patent owners, that in itself only gives a strong reason to prefer damages rules if there is no counterbalancing reason to prefer property rights. It is not enough to simply point to a risk of undercompensation to justify a property rule, but neither is it enough to simply point to the risk of holdup to justify a liability rule. Instead, the question would turn on whether the problem of undercompensation is outweighed by the holdup problem. If damages are undercompensatory, this kind of balancing inquiry is inherently difficult, as it turns not just on the existence of undercompensation or holdup, but also on an estimate of the relative severity of each.

[162] *See, e.g.*, Lemley & Shapiro 2007a; Lee & Melamed 2016.
[163] Lemley & Shapiro 2007b, 2172.

2 Are Damages Biased?

a) Direct Evidence

I am not aware of any direct evidence assessing whether damages awards are biased, in the form of a comparison between damages awards and the plaintiff's true loss. It is difficult to imagine how such a comparison could be carried out, given that a legal damages assessment, at least when carried out by a judge, is the most rigorous method we have for assessing the plaintiff's true loss.

b) Option Effect

Denicolò et al. (2008) and Elhauge (2008) argue that even unbiased errors in determining the reasonable royalty rate could favor the infringer, as the downstream firm could pay the court-determined royalties when they are too low and redesign the product when they are too high.[164] For convenience, I will refer to this as an "option effect," as the argument is that the court-determined royalties effectively provide the infringer with an option that can be exercised when it is in the money. Shapiro (2010) agrees with this basic point, but he states that sufficiently small errors will not affect the basic model and its implications so long as the court-determined royalties are unbiased, and further, the option effect "might not arise, even for fairly large errors, for patents covering a small feature of a high-margin product: the downstream firm would pay greatly excessive royalties rather than withdraw its product from the market while engaging in the redesign."[165] However, Shapiro (2016) provides a model in which the option effect is the only source of under-compensation to the patentee when the patentee would be willing to license, and recommends that injunctive relief should sometimes be granted for this reason.[166]

The option effect will be larger if the damages error is large and the intrinsic holdup is small. It also seems that the option effect will be relatively larger for a stronger patent, because it is important only if the patentee wins and is awarded damages. On the other hand, the option effect will have no impact in the early negotiation scenario, where the implementer's threat point is to avoid using the invention entirely, though presumably it would affect the exact probability of

[164] Denicolò et al. 2008, 578–80; Elhauge 2008, 557–58. *See also* Kaplow & Shavell 1996, 761–62 (making essentially the same point to argue that property rights are appropriate for protecting entitlements to things).

[165] Shapiro 2010, 305–06.

[166] Shapiro 2016, 11–12 (describing the option effect), 22 (noting that "the value of the downstream firm's option to negotiate rather than pay the court-awarded royalties declines as the switching costs grow," and discussing when injunctive relief should consequently be awarded). Shapiro 2016, 13–14, also describes a variant of the option effect that arises when the implementer would not have found it worthwhile to use the invention *ex ante*, but the royalty awarded by the court is sufficiently low to make it worthwhile *ex post*. In the absence of reverse payments from the patentee to the implementer, the implementer may pay the unduly low royalty, and the patentee will be undercompensated.

validity at which that becomes the relevant threat point. It is at least clear that the size of the option effect depends on the magnitude of the variance in the error in damages awards, and without empirical evidence on this point, it is difficult to know how important this effect might be in practice.

Further, Cotter (2014a) adds that this strategy "seems to require a good deal of foresight on the part of infringers, as well as a willingness to ignore the high cost of attorney fees and (in some countries) the risk of enhanced damages if the defendant knowingly infringes."[167]

On the whole, it is plausible in principle that the option effect might result in undercompensation if damages are awarded in lieu of an injunction, but it is not clear how significant the effect will be in practice. It would be helpful to be able to estimate the variance in the error of damages awards, but that will be very difficult given that it is not even clear how we could estimate the error term itself.

c) Burden of Proof

Kieff & Layne-Farrar (2013) point out that putting the burden on the patentee to prove its loss may be problematic in the context of reasonable royalty damages because the royalty is often assessed as a portion of the value to the infringer, which requires the patentee "to adduce evidence about a decision made long ago inside the secret business workings of the infringer's enterprise to select the infringing technology over any alternatives that may or may not have existed at that time."[168] More generally, the general principle that the plaintiff must prove its loss may in principle result in undercompensation. The plaintiff's actual losses will be supported by a range of evidence, with some losses supported by more evidence than others. This directly implies that in at least some cases the plaintiff will suffer actual losses that it cannot recover, and that in turn implies that damages are normally undercompensatory.[169] This is not to say that it is wrong to put the burden on the plaintiff to prove its loss, as the opposite rule would result in systematic overcompensation, but the point remains that the rule implies that the plaintiff will be systematically undercompensated.

[167] Cotter 2014a, 345.

[168] Kieff & Layne-Farrar 2013, 1117.

[169] This result follows because proof on the balance of probabilities is a threshold that cuts off some losses entirely. In contrast, under an alternative rule in which damages would be awarded for all losses for which there is evidence, but discounting for the strength of the evidence, difficulty of proof would not in principle be a source of undercompensation. That is, if the plaintiff identified a $1 million loss, but could only establish a 10 percent probability that the loss was caused by the tort, it would be awarded $100,000. To be clear, I am not advocating such a rule, but merely using it to illustrate why the rule that the plaintiff must prove its loss results in undercompensation, at least in principle. No doubt the plaintiff will often attempt to prove losses that did not occur, but it is reasonable to suppose that the evidence supporting losses that did not occur will systematically be weaker than that supporting losses that did occur. This implies that unwarranted compensation for loss that did not occur will not be sufficient to offset denial of compensation for actual losses.

d) Hindsight Bias

Elhauge (2008) suggests that damages might be systematically undercompensatory due to hindsight bias, on the view that juries may underestimate the value of the invention because inventions often seem more obvious after they have been created. However, as Cotter (2013a) points out, hindsight bias might just as plausibly lead to overcompensation.[170] More generally, behavioral economics has identified a variety of psychological mechanisms that give rise to systematic biases in decision-making, so it is plausible that such mechanisms might lead to systematically biased damages, but these mechanisms turn on the details of the decision-making context, and it is not clear how these effects will play out in the context of patent damages.

e) Jury Bias

Jury trials are often used in the U.S. system. Juries are more sympathetic toward patentees than judges, and are more likely to award greater damages.[171] It is not uncommon for jury awards to be overturned on appeal as not being adequately supported by the evidence, and this suggests that jury awards are systematically overcompensatory. Even if jury awards are systematically overcompensatory, this does not imply that awards made by a judge alone are unbiased. Since damages awarded by judges and juries appear to be systematically different, both cannot be unbiased, but it is possible that both are biased.

f) Interest

Damages will be undercompensatory if interest is not awarded, as is the case in some jurisdictions.[172] This may lead to holdout, but it is a problem that impacts patent litigation, and indeed all litigation, well beyond patents for complex products, and the obvious solution is to award interest at compensatory rates.

g) Presumption of Unbiased Damages

Lemley & Shapiro give no reason for believing that errors in reasonable royalty damages (or any other damages) are unbiased. They are not alone in this; I am unaware of any scholarship in the general property and liability rules literature that makes a positive case for the view that damages assessments are unbiased, as opposed to simply assuming it. The implication is that we should presume that damages are unbiased in the absence of any evidence to the contrary, so that the burden of proof

[170] Cotter 2013a, 345.

[171] *See* Chien 2014, 22 and sources cited therein.

[172] *See* Cotter 2013a, 276 (noting that interest is routinely awarded in some jurisdictions, but not in others); Denicolò et al. 2008, 602–03 (suggesting that damages are likely to be undercompensatory for this reason).

lies with those who suggest damages tend to be undercompensatory.[173] As discussed above in this section, the direct arguments as to whether damages assessments are biased are not conclusive, so the presumption matters.

One possible reason for presuming that damages are fully compensatory in expectation is that full compensation is the stated goal of the law of damages generally, and the law of patent damages in particular. But in legal scholarship the fact that the courts say that they are doing something is not usually taken as particularly good evidence that they are succeeding.

On the other hand, descriptively, "[t]he standard practice in virtually all legal systems assumes the dominance of property rules over liability rules," except in circumstances in which there is a serious risk of holdup.[174] To the extent one believes that the common law tends toward efficiency, this would suggest that there is one fundamental and general concern reflected by property rights. One candidate for such a general concern is that damages are undercompensatory. That is, rather than saying that property rules are justified by undercompensatory damages, it might be suggested that the prevalence of property rights is itself reason to believe that damages are undercompensatory. However, this inference is not very strong, given that there are plausible alternative explanations for the dominance of property rules, and considering that the theory of property rights and liability rules remains unsettled.

Smith notes that, in the general literature, "[p]ro-liability rule commentators also tend to disagree with those in the pro-property rule camp on the relative magnitudes of both the hold-out and undercompensation problems,"[175] and the same appears to be true in the patent literature. This is even though the two problems are independent; there is no particular reason to believe that holdup will be large if valuation is accurate, or vice versa. It is just as plausible that both problems are large, or both are small. This suggests that intuitions on whether damages assessments are accurate may turn on general intuitions about the desirability of injunctive relief, rather than the other way around.

7.5.5 *Summary*

In the general property rights scholarship, inaccuracy of damages has been a prominent justification for injunctive relief. However, this justification is most

[173] *See, e.g.*, Hovenkamp & Cotter 2016, 903–04 (suggesting that "[a]bsent some reason to believe that courts systematically are likely to err in favor of defendants," there is no obvious reason to suppose that damages are undercompensatory in expectation). *See also* Shapiro 2016 (providing a model in which judicial errors are unbiased, and noting that if this is so, the errors will not affect the incentive to innovate, but without providing any support for the assumption).

[174] Epstein 1997, 2092; *see also* Smith 2004, 1731–40 (describing the "long tradition of preference for property rules in the law," except in situations involving very high transaction costs or holdout and strategic behavior).

[175] Smith 2004, 1746.

powerful when errors in damages awards are systematically undercompensatory, rather than inaccurate but unbiased. It is often assumed that errors in damages are unbiased, but there appears to be no sound justification for this assumption; but on the other hand, neither is there any compelling general reason to suppose damages are systematically significantly undercompensatory. Given the centrality of this issue to property rights generally and the question of injunctive relief for patents for complex products more specifically, the issue warrants further research.

7.6 HOLDOUT/REVERSE HOLDUP

7.6.1 *General*

Holdout, or reverse holdup, refers generally to efforts by an implementer to pay a royalty that is unfairly low. In contrast to holdup, holdout is generally undertheorized.[176] The holdout argument is typically stated informally, leaving considerable ambiguity as to the precise mechanism, with consequent lack of clarity as to the circumstances in which holdout is likely to be a problem. Once unpacked, the factors are generally ambiguous. A notable exception is Langus et al. (2013) who provide a very detailed model of holdout in the context of European law. The drawback of their model is that its very specificity makes it unclear how widely their results can be generalized.

The difficulty of enforcing patents is commonly suggested as the primary source of holdout, on the view that when damages are compensatory, the threat of an order to pay damages (and costs) does not act as an effective deterrent, because the implementer will be no worse off if it resists and is ultimately held liable than if it licenses *ex ante*. For example, Epstein et al. (2012) say that if reasonable royalty damages are capped at the amount the infringer would pay in *ex ante* negotiations,

> the blithe infringer – the infringer who for any reason falls short of "willful" – is to pay no more, if identified, sued, and defeated, than he would have had to pay if he had in fact negotiated a license at the time the standard was set. The situation is difficult enough if the patentee is in a position to identify and pursue, often at great cost, the large number of infringers. But, these assumptions ignore the high costs in the detection and enforcement of these rights.[177]

[176] *See* Chien 2014, 20 (noting that holdout is "arguably undertheorized").

[177] Epstein et al. 2012, 26–27. Kieff & Layne-Farrar 2013, 1113 argue that if a RAND commitment were interpreted as preventing SEP holders from ever seeking an injunction, "infringers would rationally consider the benefits of simply avoiding any up-front offer to take a license on any terms, RAND or not, knowing that on the back end they will not have to face an injunction for any patent that makes its way into any RAND commitment from within an SSO." However, they do not explain why it would be rational for an implementer to avoid an *ex ante* license on RAND terms if the probability of detection is high and they would eventually be required to take a license on RAND terms and pay the same RAND royalty for pre-license infringement. *See also* Wright 2014, 807 (stating that "it is well understood that weakening the availability of injunctive relief for infringement . . . may increase the

This is sometimes referred to as the "catch-me-if-you-can" problem.[178]

Comments such as these raise three distinct issues: (1) litigation costs; (2) underdetection; and to a lesser extent (3) undercompensatory damages.[179] Oligopoly power on the part of implementers is also sometimes put forward as a fourth distinct source of holdout, particularly in the context of SSOs.

1 Litigation Costs and Resource Constraints

To isolate the role of litigation costs, suppose that detection is certain and damages are fully compensatory. The basic rejoinder to the argument that implementers will take advantage of high litigation costs to force an unfair settlement is that this strategy is expensive for the implementer as well. If litigation costs are symmetric, costs drop out of most formal models, as the parties will settle in order to avoid them, and symmetric costs do not give either party an advantage in the negotiations. Indeed, it is normally suggested that high litigation costs will encourage early licensing, rather than holdout, in order to avoid the litigation costs.[180] This reasoning implies that there must be some kind of asymmetry between the parties before litigation costs can distort the royalty settlement, though when asymmetry does exist, it can result in unfair settlements.[181] The same is true if there are asymmetries in risk aversion, perhaps because of resource constraints.

probability of reverse holdup and weaken any incentives implementers have to engage in good faith negotiations with the patent holder," and that in the absence of injunctive relief "a potential licensee can delay good faith negotiation of a F/RAND license, and the patent holder can be forced to accept less than fair market value for the use of the patent," though without explaining the mechanism); Geradin 2010a, 125 (arguing that without the threat of an injunction "any firm wishing to implement a standard would be invited to begin immediately using the invention without even trying to obtain a license from the IP owner and take its chances in court later," though again without elaborating on the mechanism); Egan & Teece 2015, 13 ("Implementers can simply use the invention covered by a patent and wait to get sued, using as many diversionary tactics in the courts as is possible, knowing that it is hard, time-consuming, and expensive for a patentee to get an injunction."); Sidak 2008, 736–43; Camesasca et al. 2013, 300.

[178] Golden 2007, 2135.

[179] Note that in actual *ex ante* negotiations the royalty is presumably discounted by the probability of validity, while in U.S. law, at least, reasonable royalty damages are assessed on the basis that the patent was known to be valid and infringed. Thus, it is not strictly correct to say that the implementer who is caught will pay no more than it would have had to pay had it actually negotiated a license in the first place. However, it is true that the expected royalty (if calculated accurately) is the same whether the licensee negotiates a discounted royalty, or gambles on paying a non-discounted royalty.

[180] That is obviously true if each party bears its own costs, but even with full fee shifting in favor of the successful party, expected litigation costs will be positive, and the licensee will strictly prefer to license *ex ante*. Moreover, litigation costs are never fully shifted, particularly if one takes into account business disruption. Camesasca et al. 2013, 300, call costs "more or less irrelevant," but even very small costs are enough to make the implementer prefer to license, all else being equal.

[181] *See* Morton & Shapiro 2016 (discussing the distortion caused if litigation costs are highly asymmetric as between the patent holder and the target firm); Chien 2014 (discussing asymmetry arising when small firms are involved in litigation); Denicolò et al. 2008, 594 (noting asymmetric litigation costs may lead to holdup problem in either direction). In Lemley & Shapiro 2007a, 1999 n.16, litigation

Litigation costs may undoubtedly be asymmetric in particular cases, but there seems to be little reason to believe that litigation costs systematically favor the accused infringer either in general, or in the category of patents for complex products. Indeed, it may be that patent holders have a systematic cost advantage because litigation may impose substantial discovery costs on alleged infringers without an equivalent burden on the patent holder.[182] Similarly, patent assertion entities likely have a cost advantage over end users.[183] Nor is there any particular reason to believe infringers have a systematic resource advantage over patentees.[184]

With that said, there is no doubt that cost or resource asymmetries may cause significant distortions in some individual cases, and potentially in some categories of cases. However, it is not clear that granting injunctive relief in such cases will effectively address the problem in those cases where costs consideration favor the implementer. Litigation cost asymmetries tend to lead to unfair settlements when those costs are high relative to the value of the invention, so that the implementer's main leverage is the threat to impose high litigation costs on the patentee. The prospect of injunctive relief as a remedy, as opposed to an ongoing royalty, will shift that balance only when the extra costs imposed on the implementer by injunctive relief as opposed to an ongoing royalty – that is, the holdup costs – are large enough to counterbalance the litigation cost asymmetry. This means granting injunctive relief would not help the patentee in those cases in which the holdup threat is relatively small. With extremely high costs and delay, injunctive relief becomes entirely irrelevant.[185]

Injunctive relief might tilt the balance substantially, even in the face of high litigation costs, if there is a very large potential for holdup. But allowing holdup may not be a proportionate response, for example if the implementer was not aware of the patent when it infringed, or if the implementer had a good-faith belief that the patent

costs drop out of the analysis because of their focus on a percentage overcharge. However, in their model asymmetric litigation costs still result in an absolute over/undercharge.

[182] Morton & Shapiro 2016, 13; Golden 2007, 2133.

[183] Chien 2014, 13 (noting specialized PAEs have been able to drive down the costs of bringing patent cases without a corresponding reduction in the cost of defense, and "[t]he resulting gap between the cost of defense and cost of assertion has created compelling patent nuisance fee economics").

[184] Golden 2007, 2132, suggests that "a patent holder's resources for litigation might also be substantially less than those of the potential infringer," but without noting the opposite is also plausible. Golden goes on to say that the infringer enjoys an additional advantage because it will, "if it chooses, likely be able to enjoy the benefit of the invention for years before the typically tortuous process of patent litigation can produce favorable returns for the patent holder." However, if the successful patentee is fully compensated for the past infringement, and it has the resources to fund the litigation, then the fact that the patentee was not receiving royalties during the litigation period will not affect the bargaining outcome.

[185] *See* Golden 2007, 2134–35 ("The potential infringer may very well have a plausible claim that the threat of a permanent injunction is no real threat at all – that by the time a permanent injunction could issue, the accused product will have long since, and in the regular course of business, been either discontinued or substantially redesigned in a way that nullifies any possible claim of ongoing infringement.").

was invalid. This suggestion is more justifiable where *ex ante* negotiations were feasible, so that the result of granting injunctive relief is to induce negotiation. But in that case the holdout argument is primarily a supporting rationale for the view that injunctive relief should be preferred when *ex ante* negotiations are possible.

Most of the analysis of litigation costs in the holdup context has assumed the American rule that each party bears its own costs. Fee shifting may be a more effective way of addressing the holdout problem raised by asymmetric litigation costs, though it raises its own problems.[186]

2 Asymmetric Stakes

Golden (2007) suggests that there is an inherent asymmetry in the amount at stake in patent litigation because a patentee who is unsuccessful in litigation will lose not just the revenue from that one deal, but also from other potential licenses if the patent is held to be invalid.[187] However, this does not reflect undesirable leverage; it merely reflects the point that the negotiated royalty should reflect the probability that the patent is invalid.

7.6.2 *Underdetection*

Denicolò et al. (2008) note that implementers may infringe intentionally without seeking a license "hoping that patent holders do not have the will or the resources needed to detect or pursue each and every instance in which their patents are infringed."[188] If the probability of detection is sufficiently small, the expected royalty may be undercompensatory even in the presence of some degree of holdup; the royalties that are paid will be too high, but many will not be paid at all. Consequently, if there is a significant likelihood of underdetection, a holdout problem may arise in the sense that the implementer may choose not to negotiate a license *ex ante*, even though it anticipates that it will be held up for an excessive royalty if it has to negotiate *ex post* under the threat of an injunction. If injunctive relief is routinely denied, then the problem is exacerbated because the downside to the implementer of holding out is reduced, and so there will be more situations in which it is rational to hold out.

In the general remedies context, Polinsky & Shavell (1998) argue that the problem of underdetection justifies an award of enhanced damages under which the multiplier reflects the probability of the infringer escaping detection.[189] However, as Blair

[186] *See, e.g.*, Chien 2014, 40–41, for a brief discussion with citations to some of the general fee-shifting literature.

[187] Golden 2007, 2134 (noting also that even short of invalidation, failure to reach agreement might make agreement with others less likely).

[188] Denicolò et al. 2008, 591.

[189] Polinsky & Shavell 1998, 887–96.

& Cotter (2005) note, calculating the multiplier with any degree of accuracy may be impossible.[190] And as Cotter (2013a) notes, "most nations generally do not authorize awards of enhanced damages," and in the United States, which does, the availability of enhanced damages depends upon state-of-mind criteria that have relatively little to do with the underdeterrence rationale.[191]

Less attention has been focused on the implications of underdetection for injunctive relief.[192] In principle, the holdup value that a patentee, armed with the prospect of an injunction, might extract could serve as a kind of enhanced damages that would counterbalance the problem of underdetection. Even though the individual implementers who were detected would be held up, in principle this would not adversely affect implementer behavior, because the expected rate of return would not be depressed below that which would be expected if there were no holdup and no underdetection. However, there are no evident structural or institutional considerations that suggest that the problems of holdup and underdetection are likely to balance each other, even roughly, and in contrast with enhanced damages, there is no adjustable multiplier, which might, at least in principle, allow the court to balance the two factors, even if the court could assess the probability of underdetection.

7.6.3 *Undercompensatory Damages*

To isolate the issue of undercompensatory damages from that of litigation costs and underdetection, suppose that detection is certain and litigation costs are symmetric, but damages are undercompensatory. In that case, it will only be in the interest of the implementer to hold out by delaying trial if injunctive relief *is* routinely granted, in which case it will be in the interest of the implementer to delay proceedings because the effective royalty paid prior to trial, in the form of damages, will be less than the royalty it pays after trial when it has to bargain under the threat of an injunction. On the other hand, if injunctive relief is routinely denied, and the same reasonable royalty is granted post-trial as an ongoing royalty as for pretrial damages, then the implementer has no reason to delay trial, because its liability is the same before and after. On the contrary, in that case the implementer would prefer to settle early – for the undercompensatory rate that both parties anticipate being awarded in litigation – in order to avoid litigation costs. Thus, if the only concern is undercompensatory damages, routinely granting injunctive relief is the source of holdout, not a cure for it.

[190] Blair & Cotter 2005, 45–49 (analyzing the issue), 58 (summarizing by noting that "calculating the appropriate amount of the multiplier may be impossible").

[191] Cotter 2013a, 73.

[192] Denicolò et al. 2008, 592, raise the issue in the context of an article on injunctive relief, but they conclude only that "policy should be concerned not only with the possibility of holdup, but also with manufacturers' incentives to behave opportunistically, purposefully infringing a known patent or failing to adequately search for patents."

7.6.4 Oligopoly Pricing in SSOs

There is a substantial literature addressing the possibility that implementers, particularly when operating through the framework of SSOs, may exercise oligopoly power to depress royalties that would otherwise be obtained by patentees.[193] These concerns are addressed primarily through competition law. Addressing that literature is beyond this scope of this chapter, as it does not have direct implications for patent remedies.

7.6.5 Summary

The basic intuition behind the catch-me-if-you-can argument is that without the threat of injunctive relief, the implementer has no particular incentive to seek a license, and the burden of seeking out the implementer and initiating negotiations lies with the patentee. Injunctive relief levels the playing field (or tilts it the other way), by giving the implementer an incentive to seek out a license early on, or risk being held up. This argument is most powerful when *ex ante* licensing is feasible, in which case it supplements other arguments for injunctive relief, such as the valuation problem and the desirability of reducing transaction costs.

When *ex ante* negotiations are not feasible, so that the catch-me-if-you-can argument must stand on its own, it is less persuasive as a justification for injunctive relief as it is not clear that the specific mechanism at issue systematically favors the implementer. Holdout and holdup are normally portrayed as opposing arguments, in favor of or against injunctive relief. But as Chien (2014) argues, in many respects both can be seen as consequences of transaction costs and asymmetries in the patent litigation system, which implies that both can be addressed simultaneously by reforms that target those fundamental problems. Consequently, reforms aimed directly at these problems are desirable, such as early dispositive rulings, institutional coordination, and fee- and cost-shifting, along with other procedural reforms.[194]

7.7 ROYALTY STACKING

7.7.1 Introduction

Royalty stacking refers generally to any mechanism by which the total royalty burden is unduly increased by the presence of multiple patentees.[195] The term may refer to two distinct phenomena: first, where the presence of multiple patentees exacerbates the effect of one of the forms of holdup described above; and second, the Cournot

[193] *See, e.g.*, Sidak 2009; Farrell et al. 2007, 632; Gilbert 2011; Kieff & Layne-Farrar 2013, 1107–09; for a review of some of the literature, *see* Cotter 2009, 1200–06.

[194] Chien 2014; Morton & Shapiro 2016; Golden 2007, 2125.

[195] *See* Lemley & Shapiro 2007a, 1993 ("Royalty stacking refers to situations in which a single product potentially infringes on many patents, and thus may bear multiple royalty burdens.").

complements problem, which may arise even in the absence of holdup. The term is also commonly used to refer to any situation in which the cumulative royalty seems too high. However, a high aggregate royalty is not problematic in itself, as it may simply indicate that the licensed technologies are valuable.

7.7.2 *Cumulative Effect of Holdup*

Lemley & Shapiro (2007a) note that "the existence of such 'royalty stacking' exacerbates the holdup problem," and "[a]s a first approximation, the magnitude of the [holdup] problem is multiplied by the number of patents that read on the product."[196] However, Denicolò et al. (2008) point out that this is true only if the cost of redesign is independent. If the cost of designing around two patents at once is less than the sum of designing around each of the patents separately, then the holdup problem is less than additive. In the extreme case where the cost of designing around two patents at once is the same as the cost of designing around one of them, then any cumulative effect is due only to the difficulty of bargaining with two patentees rather than one, and not due to an increase in holdup itself.[197] It is not clear whether the costs of redesign are generally independent. Moreover, even if two patents could be designed around as easily as one, if the implementer faces sequential demands, independent redesign costs may arise.

7.7.3 *Cournot Complements*

1 Theory

The problem of Cournot complements arises in principle whenever multiple independent suppliers with market power sell complementary inputs; Cournot's example was suppliers of copper and zinc, which is combined to make brass.[198] The price decisions of each supplier impose a negative externality on other suppliers; as one supplier raises its price, demand for the product decreases, thus decreasing revenue for the other suppliers. If the suppliers price independently they will not take this externality into account, and the resultant aggregate price will be higher

[196] Lemley & Shapiro 2007a, 2011; *id*. at 1993 ("As a matter of simple arithmetic, royalty stacking magnifies the problems associated with injunction threats and holdup, and greatly so if many patents read on the same product.").

[197] Denicolò et al. 2008 assume symmetric bargaining power, so that the implementer would get only one-third of the total rent if it negotiated with two patentees, whereas it would get half if it negotiated with them individually; but these assumptions about bargaining power and the split of the surplus are not theoretically robust. This is not to dispute the basic point made by Denicolò et al. 2008, but rather to reinforce it; in their example and with other plausible assumptions regarding bargaining power, there might not be any stacking effect at all.

[198] Cournot 1838, 99–116.

than would be charged by a single supplier that owned all the inputs.[199] Consumers will be worse off, and the suppliers (patentees) themselves will also be worse off than if all the inputs were supplied by a single firm.

The Cournot problem does not arise unless there are multiple input owners, and it becomes worse as there are more independent input suppliers. In the patent context, the problem does not turn on the number of complementary patents, but on the number of independent price-setting owners of those patents. This implies that there will be no Cournot complements problem in an industry in which a single entity owns all the complementary patents. By the same token, the problem is mitigated or eliminated if some or all of the input owners coordinate their prices. That is, the extent of the problem depends on the number of patent owners who are independently price-setting.[200]

The Cournot complements problem does not require that the inputs are strict complements; it arises to some degree whenever the demand for one input depends on the demand for the other, so that an increase in the price of one affects demand for the other.[201]

Nor does the problem of Cournot complements turn on the presence of sunk costs holdup. However, in the absence of sunk costs holdup, Cournot price-setting alone cannot result in prices greater than the value of the patented technology. If the inputs are strict complements, then the aggregate royalty cannot exceed the combined value of the patented technology to the product to at least some users,[202] though other users will be priced out of the market. The royalty is nonetheless excessive in the sense that it is higher than the price that would be set by a single firm holding all the relevant patents.

In principle then, the loss to society comes from reduced output, rather than implementers refusing to enter the market at all. However, to the extent that

[199] A fortiori, it will be higher than the competitive price (marginal cost); but competitive price is not usually used as a benchmark in the patent context, as that will not provide an adequate incentive to invent.

[200] Geradin et al. 2008.

[201] For example, copper and zinc are not strict complements in making brass as they can be combined in varying proportions to create brasses with different properties.

[202] Lemley & Shapiro 2007a, 2048, do suggest that the royalty charged by an individual patentee will exceed the value of its contribution "if and only if" the value of the product without the patented technology, minus the marginal cost of the product, is greater than the value of the patented technology, which they describe as "a relatively weak condition." However, Lemley & Shapiro 2007a, 2047–48 qualify this by stating that "in the presence of holdup and opportunism, each patent has the ability to charge a royalty that exceeds the value of its patented technology," and "there is no reason why the constraint $r_i \leq v_i$, must hold if redesign costs are significant." Thus, Lemley & Shapiro have incorporated sunk costs holdup into their royalty stacking discussion. Elhauge 2008, 565, critiques Lemley & Shapiro's suggestion that where there are multiple patent owners facing one downstream firm "a 'royalty stacking' problem will be created in which each patent owner charges more than the value of its product." Elhauge 2008 says the source of this error is their failure to recognize that the implementer can "simply decline to use the overpriced technologies at all," but it is perhaps more accurate to say that Lemley & Shapiro assume sunk costs.

Cournot price-setting exacerbates sunk costs holdup, then it may result in implementers not entering the market at all.

2 Mitigating Mechanisms

A) INPUT PRICE COORDINATION. As just discussed, the extent of the Cournot complements problem depends on the number of rights holders who set their prices independently. Consequently, the problem is mitigated if owners of the complementary inputs can coordinate prices, so that they are no longer setting prices independently. Under Cournot price-setting the input owners cumulatively make less than would a single monopolist who owned all the inputs, so it would be in the interest of the input owners to coordinate prices so that the cumulative price for the inputs will be the same as would be charged by a monopolist, assuming the collective action problem can be overcome.

Some coordination mechanisms include cross-licensing among vertically integrated firms and patent pools.[203] When vertically integrated firms cross-license on the basis that each will charge the other the same rate for equivalent patents, then if firm A raises its rate, it knows that firm B will raise its rate in return, and the demand-effect externality will be internalized. If all firms are vertically integrated and symmetrical, the Cournot complements problem will be solved.[204] More generally, the complements problem depends not on the number of entities holding patents on complementary inputs, but on the entities who are independently setting the input prices, and vertically integrated firms that cross-license are effectively not independent input price-setters. However, nonintegrated upstream firms will have no interest in cross-licensing and "prefer a royalty rate that is somewhat higher than the monopoly rate," which means that the Cournot complements problem will persist if there are nonintegrated upstream firms.[205]

Coordination can in principle also be achieved by a patent pool, even in the presence of non-vertically integrated upstream patent holders.[206] A pool will license the pooled patents at a rate that maximizes profits by balancing higher royalties against lower volumes. That is, the pool internalizes the externality in the form of reduced volumes, which gives rise to the Cournot complements problem. Because the price with Cournot stacking is higher than the profit-maximizing price, it is advantageous for all patent holders to solve the problem, whether or not they are vertically integrated. However, because of high up-front expenses associated with their formation collective action problems, a pool will not necessarily be formed

[203] Layne-Farrar & Schmidt 2010, 1132–36; *see also* Contreras 2015a (discussing patent pledges as another coordination mechanism).
[204] Layne-Farrar & Schmidt 2010, 1135–36.
[205] *Id.* at 1136.
[206] *Id.* at 1135; Geradin et al. 2008; Lemley & Shapiro 2007a, 2014–15.

even if a successful pool would increase the patent owner's revenue.[207] These problems in pool formation mean we cannot be confident that pools will generally form so as to solve the Cournot complements problem.

b) tacit coordination through generalized bargaining strategies. Tacit coordination through more general bargaining strategies may also solve the Cournot complements problem. In the model generating the Cournot complements problem, each patentee sets its own per unit price while taking the prices of other patentees as given, and implementers choose quantities based on the offered price. In contrast with Cournot's single-stage price-setting model, Spulber (2016) develops a two-stage quantity-setting model. In the first stage, each input supplier (e.g., patentee), makes a binding commitment to provide whatever quantity of its input the implementers demand, up to a specified maximum. In the second stage, implementers and patentees bargain over price, resulting in a price that clears the market at the specified quantity. The result in Spulber's model is that the quantity of complementary inputs supplied is equal to the quantity that would be offered by a monopolist selling the inputs as a bundle; in other words, the royalty stacking problem disappears. The reason is that quantity-setting results in tacit coordination between patentees. Because inputs are complementary, each patentee can unilaterally set the maximum total output quantity by limiting its own input quantity offer. Because each patentee recognizes the effect of its offer on overall output, it will offer the quantity that maximizes joint profits in order to maximize the total value to be bargained over in the second stage.

The general insight from Spulber's work is that the Cournot complements problem arises because Cournot's model restricts the available strategies, and not simply from the fact of complementary input monopolies. It is of course likely that real-world licensing does not follow either Cournot's model or Spulber's, both of which assume an equilibrium outcome. For example, if patentees in fact approach the implementer sequentially, rates negotiated in earlier deals may be effectively taken as given in subsequent negotiations, and the overall equilibrium outcome will only be achieved if in the early negotiations the implementer correctly anticipates the subsequent royalty demands and bargains accordingly.

More generally, the extent of the Cournot complements problem depends on how patent holders set royalty rates in practice. A more detailed understanding of real-world royalty negotiation practices would help build a more accurate model of royalty stacking and would help identify industries in which the Cournot complements problem is likely to be important.

[207] *See, e.g.*, Contreras 2013, 76–77 (describing high upfront costs associated with pool formation); Lemley & Shapiro 2007a, 2014 (noting that potential pool member might try to hold out for a larger share of the pool, thus preventing the pool from forming at all).

7.8 EMPIRICAL EVIDENCE

7.8.1 *General*

Given that there are mechanisms that could plausibly mitigate the effects of both holdup and royalty stacking, at least in some circumstances, it is an empirical question as to whether these problems are "common enough and costly enough in actuality to warrant policy changes."[208] Three types of evidence are advanced: case studies, testing of quantitative models, and analysis of the industry structure.

On the whole, there is little evidence that holdup and royalty stacking are systemic problems, but there are some individual cases that are strongly suggestive of attempted holdup. Presumably there are other such cases that have settled and remain confidential.

7.8.2 *Case Studies*

1 Overview

Case studies in which arguably excessive royalties were demanded are often advanced as evidence of holdup or royalty stacking. There are two general concerns with case studies. One is that without a sound benchmark for the optimal royalty on the facts of a particular case, it may be difficult to say whether any particular royalty is too much. A second concern is that even if a particular case does illustrate pernicious royalty demands, a single example does not establish that there is a systemic problem. With that said, identifying what might be isolated instances of holdup or royalty stacking remains important as the courts may wish to respond to holdup or royalty stacking if established on the facts of a particular case, even if the problem is not systemic.

2 Distinguishing Holdup and Stacking

In case studies it can be difficult to distinguish holdup from royalty stacking. We may be able to conclude that a particular royalty is excessive because it implies an aggregate royalty for multiple patented technologies that would be excessive. But the royalty may be excessive, even without the Cournot complements problem, because the individual royalties are excessive due to holdup, or it may be excessive even without holdup as a result of the Cournot complements problem. And of course a combination is possible, in which the Cournot complements problem exacerbates individual holdup. The fact that the aggregate royalty is excessive does not in itself allow us to distinguish between these cases.[209]

[208] Geradin et al. 2008, 145.

[209] For example, the court in *Microsoft Corp.* v. *Motorola, Inc.* (W.D. Wash. 2013, p.73, 86) (U.S.) found that royalty stacking, rather than holdup, was the primary constraint on the upper bound of a RAND

The main problem in identifying royalty stacking in a particular case is to establish a sound benchmark for what is a reasonable aggregate royalty on the facts. The appropriate benchmark to address the Cournot complements problem is the royalty that would be charged by a single patentee holding all the relevant patents. This benchmark can be approximated by a successful patent pool. Like a single patentee, a pool will seek to maximize its revenue by considering the trade-off between a high royalty and widespread adoption of the standard. But, as noted above, pools face significant hurdles to their formation, and a relevant pool often doesn't exist. Moreover, patents that are excluded from a pool may be systematically different from those that are included. For example, if the pool in question distributes the royalties to individual patentees purely on a numerical basis, without consideration of the value of the particular patent, patents that are particularly valuable to the standard may not be adequately compensated by the pool rate. When a patentee stays out of a pool and demands a higher than pool rate, this might be because it had an average or weak patent and it was seeking to hold up implementers, but it might also be because it had a particularly valuable patent and the pool rate was not adequate. An assessment of patent quality is needed to distinguish between these possibilities.

3 Case Studies

Lemley & Shapiro (2007a) provide two examples of holdup. The first is Rambus charging "a 0.75% royalty rate for patents that do not cover industry standards and 3.50% for patents that do cover industry standards."[210] However, as Denicolò et al. (2008) point out, this misunderstands the facts in the Rambus litigation; both sets of patents covered standards, and the difference in royalty rates was due to the fact that the latter incorporated more patented components.[211]

Their second example, RIM's settlement with NTP for $612.5 million, is more persuasive.[212] The settlement was eighteen times the jury award, and the parties

royalty, but the evidence of stacking was simply an intuitive assessment that the cumulative royalty was excessive.

[210] Lemley & Shapiro 2007a, 2009 (citing Patterson 2003, 2001 n.33). Patterson in turns cites Smith 2001 as reporting that Rambus was charging a royalty of 3.5 percent of sales for rights to patents that had been incorporated in a standard, as compared with a 0.75 percent rate "for some of its other patents." Neither Patterson nor Smith stated that the other patents did not cover industry standards.

[211] Lemley & Shapiro 2007a, 2016 n.57 (citing *Rambus, Inc.* (FTC Feb. 23, 2004, ¶¶ 1262, 1390) (U.S.) (Initial Decision)). While the factual findings of the Initial Decision were vacated by the subsequent FTC Liability Opinion, *Rambus, Inc.* (FTC Aug. 2, 2006) (U.S.) (Opinion of the Commission), the Commission would still have granted a higher royalty in one respect of one standard: *see Rambus Inc.* v. *Fed. Trade Comm'n.* (D.C. Cir. 2008, p.462) (U.S.) (noting two standards were at issue, with a higher royalty for one than the other).

[212] Lemley & Shapiro 2007a, 2009 (citing *NTP, Inc.* v. *Research in Motion, Ltd.* (E.D. Va. 2003) (U.S.) (awarding reasonable royalty damages in the amount of about $33.5 million) and noting the 2006 settlement of $612.5 million).

would have had to anticipate a twelve-fold increase in sales going forward for the settlement to correspond to the reasonable royalty damages awarded by the jury.[213] Unless an extremely rapid growth in sales was plausible, or if the jury had grossly underestimated the value of the patented technology in its reasonable royalty award, this is very suggestive of holdup.

Lemley & Shapiro (2007a) also provide two case studies relating to standards, in addition to Rambus. The first relates to 3G Cellular Technology, in particular the WCDMA (3GPP) and CDMA2000 (3GPP2) standards.[214] They note the large number of patent families associated with each standard, owned by at least forty-one different companies.[215] This indicates that the structural requirements for royalty stacking are satisfied. Lemley & Shapiro then cite estimates in the range of 20 percent of the price of the phone as the total cost of the relevant licenses. Denicolò et al. (2008) dispute the accuracy of the aggregate rate, citing sources suggesting it is close to 5 percent.[216] More fundamentally, they note that even if 20 percent were the true aggregate rate, this figure in itself does not tell us that the royalty stack is excessive. Much of the value of a cell phone lies in the patented technology, and it is not obvious that 20 percent is too high for the central functionality of a phone. The value of the intellectual property in a book is not excessive simply because it is a multiple of the value of the physical medium in which it is embodied, even if that multiple is very large. The rates themselves, without any objective estimate of the value of the patented technology, are not helpful. Further, the 3G technology at issue was widely licensed and achieved substantial market penetration,[217] which suggests that holdup and stacking did not have serious adverse effects; though, as always with case studies, it might be said that uptake would have been even greater in the absence of stacking.

The second case study provided by Lemley & Shapiro (2007a) is of the IEEE 802.11 family of Wi-Fi standards.[218] Again they note that numerous patents held by multiple companies are essential to this standard, which suggests that stacking is potentially a problem, but the only evidence they give that the stacked royalties are actually excessive is that one patentee was awarded a 6 percent royalty after litigation.[219] Geradin et al. (2008) point out that without knowing how important the patent was to the standard, we can't say from the rate alone whether the royalty was excessive.[220] More generally,

[213] This is after adjusting for the fact that the jury award covered approximately six years and nine years were left on the patent. Denicolò et al. 2008, 597, argue that the settlement might have anticipated increased sales, but a twelve-fold increase seems implausibly high on its face.

[214] Lemley & Shapiro 2007a, 2025–27.

[215] *Id.* at 2026 (noting 732 patent families for WCDMA and 527 for CDMA2000; and noting that there are probably other unlisted SEPs).

[216] Denicolò et al. 2008, 599–600.

[217] Geradin et al. 2008, 160–61.

[218] Lemley & Shapiro 2007a, 2027–28.

[219] *Id.* at 2028 (referring to an award in favor of Symbol Technologies).

[220] Geradin et al. 2008, 161.

this verdict may have been an outlier.[221] Courts, and juries, sometimes make mistakes. As Lemley & Shapiro (2007b) point out, only systematic errors will affect negotiating incentives.[222] A single error, even if it is a significant outlier, will not substantially affect expected outcomes or negotiated royalties.

Cotter (2009) provides several other possible examples of "patent ambush," in which patentees were alleged to have induced an SSO to adopt a standard that incorporated patented or soon-to-be patented technology, "and then, once lock-in has occurred, demanding higher royalties than the patentees would have been able to negotiate ex ante."[223] It seems clear that in these cases the patentees were attempting to get a higher royalty by negotiating after the standard was adopted, but this does not necessarily illustrate sunk costs holdup, as opposed to network value appropriation.[224] As discussed above, the value of a patented technology increases after it is adopted as part of a standard even in the absence of any sunk costs, simply because the technology is more likely to be widely adopted. Siebrasse & Cotter (2017a) argue that allowing a patentee to capture some part of this network effect value is unobjectionable from a policy perspective. It is, in any event, a distinct effect, as holdup may allow a patentee to capture more than the value of its technology to the implementer, while the network effect does not. Without a more detailed assessment of the facts, we cannot say whether the *ex post* increase in royalty demanded was due to network effect appropriation or sunk costs holdup.

Other suggestive examples are provided by recent litigation. In *Microsoft Corp.* v. *Motorola, Inc.*,[225] Motorola had asked for a royalty of 2.25 percent of the end-product selling price for licenses to its patents that were essential to Wi-Fi and video standards. This would have amounted to a royalty of $5.85 for an Xbox, for the Wi-Fi SEPs alone.[226] Judge Robart found that a reasonable royalty was only 3.5 cents per

[221] Geradin et al. 2008 make the distinct point that "this one rate may be an outlier in comparison to non-litigated rates" because "court awarded royalty rates often include an element of punishment to ensure that future infringement is deterred." This point is speculative, and in any event it is misplaced as a critique of Lemley & Shapiro; if courts systematically add a deterrent sanction on top of the true value of the patent, this will exacerbate the holdup problem, unless the deterrent sanction is imposed only in those cases in which *ex ante* licensing was feasible.

[222] Lemley & Shapiro 2007b, 2172.

[223] Cotter 2009, 1188–89 (discussing *Rambus Inc.* v. *Fed. Trade Comm'n.* (D.C. Cir. 2008) (U.S.); *Broadcom Corp.* v. *Qualcomm Inc.* (3d Cir. 2007) (U.S.); Negotiated Data Solutions, LLC, Analysis of Proposed Consent Order to Aid Public Comment, 73 Fed. Reg. 5846–01 (Jan. 31, 2008); *Union Oil Co. of Cal.* (FTC July 6, 2004) (U.S.) (Opinion of the Commission); *Dell Computer Corp.* (FTC May 20, 1996) (U.S.) (Consent Order); and related orders and litigation).

[224] Denicolò et al. 2008, 597 n.80 say that "there seems little doubt that Rambus tried to holdup its licensees, but its attempt was struck down by the FTC," but they are evidently not distinguishing between sunk costs holdup and network value appropriation.

[225] *Microsoft Corp.* v. *Motorola, Inc.* (W.D. Wash. 2013) (U.S.).

[226] *See id.* at 65 (discussing evidence related to the 802.11 portfolio). The Xbox was the only Microsoft product that used Motorola's 802.11 SEPs. *Id.* at 54. The royalty actually proposed by Motorola was $3.00 to $4.50 per Xbox, because Motorola also wanted a cross-license to Microsoft's portfolio. *Id.* at 65. This corresponds to $5.85 when the value of the cross-license is added; that is the appropriate

unit, and an upper bound on a reasonable rate was 19.5 cents.[227] Therefore, Motorola's demand for the Xbox was a minimum of thirty times greater than what Judge Robart found to be reasonable, and perhaps as much as sixteen hundred times greater.[228] Not surprisingly, the U.S. Court of Appeals for the Ninth Circuit said that there was "evidence from which the jury could infer that demanding a 2.25% royalty rate was not a good-faith effort to realize the value of the technology, but rather an attempt to capitalize on the value of the standard itself – that is, to obtain the hold-up value."[229] If we accept Judge Robart's FRAND rate determination as even roughly accurate, it is difficult not to see this as an instance of holdup of some kind. The video SEPs are a particularly compelling example, because the patents related to interlaced video, which is largely obsolete, and so the technology added very little value to the standard.[230]

Lemley & Shapiro also note that a patent pool, Via Licensing, has been set up "[i]n an attempt to deal with the problem of patent stacking for 802.11 products." That is, they cite the existence of a patent pool as evidence of the royalty stacking problem. On the other hand, in their review of the evidence, Geradin et al. (2008) find there is little evidence of systemic problems of royalty stacking within standard setting "that are not already adequately dealt with through existing mechanisms, including . . . patent pools" among other mechanisms. In effect, Lemley & Shapiro cite the existence of a pool as evidence that there is a problem, and Geradin et al. (2008) cite the existence of a pool as evidence that there is not a problem. More accurately, Geradin et al. (2008) do not deny the existence of the problem,[231] but they argue it has been adequately addressed.

Even those who are skeptical of whether holdup and royalty stacking are systemic problems generally do not deny that they may occur in individual cases. It is therefore somewhat surprising that there are not more clear-cut individual cases, though that may be in part because the difficulty of assessing whether a royalty is excessive cuts both ways, and because most negotiations remain confidential. With that said, the individual cases taken together are at least strongly suggestive that excessive royalty demands resulting from holdup and/or royalty stacking do occur, at least on occasion.

comparison, because the FRAND rates found by Judge Robart did not reflect any value for cross-licenses.

[227] *Id.* at 101.

[228] The discrepancy for Motorola's video (H.264) patent portfolio was even greater, as Motorola asked for the same 2.25 percent royalty, and the FRAND rate found by Judge Robart was only 0.555 cents per unit, with an upper bound of 16.389 cents per unit.

[229] *Microsoft Corp.* v. *Motorola, Inc.* (9th Cir. 2015, p.1053) (U.S.).

[230] A caveat is that Motorola's portfolio included twenty-four patents, and the FRAND royalty was based on only eleven that were found to have been used. Motorola's initial demand might have reflected a good faith belief that those patents were also valid and infringed, but even if Motorola had been right, at most this would have doubled the FRAND royalty.

[231] Geradin et al. 2008, 149 ("Certainly the complements theory behind royalty stacking has stood the test of time.").

7.8.3 *Testing of Empirical Models*

1 General

Empirical studies generally do not establish that holdup and royalty stacking are serious systemic problems. Geradin et al. (2008) review the empirical evidence relating to the semiconductor, software, and biomedical device industries, and find no clear evidence that anti-commons and royalty stacking are significant problems.[232]

2 Holdup

The most important recent study is that of Galetovic et al. (2015), which examines SEPs in particular. They examine two empirical implications of the SEP holdup hypothesis. First, if holdup in the standards context is slowing the rate of innovation, then products that are highly reliant upon SEPs will experience slower rates of decrease in quality-adjusted prices than similar products that do not. Second, they consider the quasi-natural experiment resulting from the 2006 Supreme Court of the United States decision in *eBay Inc.* v. *MercExchange, LLC*,[233] which made it more difficult for SEP holders to obtain injunctions against infringers than for the holders of non-SEP patents. They find no evidence of SEP holdup on either test. With respect to the comparison between industries, they find:

> [P]roducts that are SEP-reliant have experienced faster price declines than any other good in the Consumer Price Index (CPI) over the past 16 years . . . The prices of SEP-reliant products have fallen at rates that are not only fast relative to a classic holdup industry, they are fast relative to other patent-intensive products that benefit from Moore's Law but are not SEP-reliant.[234]

On the second test, they use a difference in differences specification to test whether quality-adjusted prices fall faster in SEP-reliant industries after *eBay*, while controlling for industry and year effects. Their analysis does not allow them to reject the null hypothesis that *eBay* did not differentially affect SEP-reliant industries.

These results imply that holdup is not systemically impeding innovation in SEP-reliant industries. There are two caveats to these results that are potentially relevant to remedial issues. First, they do not claim that individual firms never attempt to engage in behavior that can be characterized as holdup.[235] Courts may wish to respond to individual instances of holdup, even if it is not a systemic problem.

[232] Geradin et al. 2008, 155–59. They also consider the examples of WCDMA and Wi-Fi in mobile telephony, that are discussed above, with the same conclusion. *Id.* at 159–63.

[233] *eBay Inc.* v. *MercExchange, L.L.C.* (U.S. 2006) (U.S.).

[234] Galetovic et al. 2015, 554.

[235] *Id.* at 555.

Secondly, they do not take issue with the view that the theoretical conditions for holdup exist in SEP-reliant industries, which suggests that it is some mitigating mechanism that explains their results. One possibility is that systemic holdup has been avoided as a result of structural factors such as the prevalence of *ex ante* bargaining or repeat play mechanisms. On the other hand, we have seen that it is sometimes suggested that it is legal constraints, such as the FRAND commitment, that mitigate the effect of holdup. That hypothesis is broadly consistent with the result that the prices of SEP-reliant products have fallen at rates that are fast relative to other patent-intensive products that are not SEP-reliant. It is more difficult to reconcile with the result that *eBay* has had no observable effect on holdup, but it is possible that *eBay* was effectively anticipated in the context of SEPs. That is, it may be that even before *eBay*, implementers understood that the FRAND commitment meant what it said and that they would be able to use standards subject to the FRAND commitment without fear of being held up by injunctions or excessive royalties.

From a remedial perspective, it matters what the particular mechanism might be. If structural factors are at play, this would suggest that the courts should be relatively reluctant to withhold injunctive relief to a successful patentee. On the other hand, if it is the FRAND commitment that is avoiding holdup in SEP-reliant industries, the results of Galetovic et al. (2015) show that the FRAND system is working, but it might suggest that the courts should continue to apply the FRAND principles relatively aggressively in order to ensure that the system keeps working. This might also suggest that the courts should apply a similar reluctance to grant injunctions even in respect of patents that are not FRAND committed, if the potential for holdup is otherwise present. The other side of that coin is that it is also possible that the FRAND commitment has been applied too aggressively, resulting in an inadequate incentive to invent. There appear to be no systemic studies addressing that possibility, though it is likely too soon for incentive effects to have manifested themselves.

3 Royalty Stacking

Galetovic & Gupta (2017) empirically investigate royalty stacking, and the Cournot complements problem in particular, in the world mobile wireless industry, focusing on third generation (3G) and fourth generation (4G) wireless cellular standards defined by the third generation partnership project (3GPP). Their paper draws on the fact that the number of SEP holders and the number of SEPs have grown dramatically over the life of this technology: "During the last 20 years the number of SEP holders for 3G and 4G standards grew from 2 in 1994 to 130 in 2013 and the number of SEPs rose from fewer than 150 in 1994 to more than 150,000 in 2013."[236] Cournot complements theory implies that with the increase in the number of SEP

[236] Galetovic & Gupta 2017, 19–20, figure 2.

holders, royalty stacking would have gotten worse. In particular, they note that the price of phones should increase or (if quality increases demand) at least stagnate; that margins of SEP holders and downstream manufacturers will fall; and that the number of device manufacturers will decrease and industry concentration will rise. They find none of these effects. On price, for example, they find that "between 1994 and 2013 and controlling for technological generation, the real average selling price of a device fell between −11.4% to −24.8% per year. Moreover, the introductory average selling price of successive generations fell."[237] They also find no trend in margins, and that industry concentration fell.[238] There are many other variables that might also affect the price of phones. Most obviously, the quality of phones has increased, raising willingness to pay, and manufacturing costs have probably decreased, and other factors such as incomes, substitute prices, and downstream intensity of price competition have also changed.[239] However, in their model, such changes cannot explain the price decrease and other observed effects, because when stacking is severe, the stacked royalty will increase to extract any benefit from cost reductions or increased demand.[240]

Galetovic & Gupta portray these results as indicating that royalty stacking has not been a systemic problem in the wireless industry, despite the large number of SEP owners. This raises a puzzle: How is this result to be reconciled with Cournot complements theory? The general Cournot complements model developed by Galetovic & Gupta (2017) shows that "even with a modest number of SEP holders, the effect of royalty stacking on output is severe and eventually, output collapses."[241] As they observe, the modern wireless industry has a large number of complementary inputs in the form of SEPs, held by independent owners. This implies that the market should "nearly disappear" and yet, as they also observe, the modern wireless industry is very healthy.

Galetovic & Gupta do not attempt to resolve this puzzle. As discussed above, the Cournot complements problem might be mitigated or solved by wide-scale price coordination, perhaps through patent pools, or possibly by specific pricing strategies or practices, but it is not obvious that such factors can explain the apparent lack of royalty stacking in the wireless industry. If Galetovic & Gupta's basic results are replicated, it is of pressing interest to explain why the wireless industry is so robust, as this might shed entirely new light on the Cournot complements problem. While Galetovic & Gupta present their work as challenging the claim that royalty stacking is a problem in complex product industries such as cellular phones, their work can also be seen as a challenge to Cournot complements theory itself.

[237] *Id.* at 5.
[238] *Id.* at 24–25.
[239] *Id.* at 20–21.
[240] *Id.* at 22.
[241] *Id.* at 16, referring to a scenario in which additional SEP holders do not add value. Their model produces similar results when additional SEP holders do add value: *id* at 16–17.

7.8.4 *Industry Structure*

In the general economic literature on holdup, the existence of holdup is often inferred from its institutional effects. For example, vertical integration may be a response to a potential holdup problem.[242] It is possible that the FRAND commitment can be understood as an institutional response to the holdup problem in the standards context. However, there are few studies that explore this analysis in depth, and it is not clear what remedial implications it might have.

7.8.5 *Summary*

On the whole, there is little evidence that holdup and royalty stacking are systemic problems, but there are some individual cases that are strongly suggestive of attempted holdup. The remedial implications of this conclusion are not clear, as the exact mechanism by which holdup is being kept in check is not clear. It may be that holdup is rare because of structural factors, such as repeat play, or because of legal factors such as the FRAND commitment and the threat of intervention by competition authorities; the first hypothesis suggests a general willingness to grant injunctive relief is appropriate, while the latter suggests that the courts should be vigilant to ensure than injunctions do not result in holdup. It is also reasonable to suggest that even though structural factors generally prevent holdup, the courts should be willing to deny injunctive relief in those cases where holdup is attempted. It is therefore important to distinguish these scenarios, and the factors that should consequently be considered in granting injunctive relief.

[242] *See generally* Masten 1996.