WHY HOLD-UPS OCCUR: THE SELF-ENFORCING RANGE OF CONTRACTUAL RELATIONSHIPS

BENJAMIN KLEIN*

Court enforcement and private enforcement are not alternative contract enforcement mechanisms, but are used jointly by transactors to define the self-enforcing range of a contractual relationship. Within this framework contract terms economize on the limited amounts of private enforcement capital possessed by transactors, either by directly controlling transactor behavior or by shifting private enforcement capital between transactors to coincide with likely future market conditions. Hold-ups occur when market conditions change sufficiently to place the relationship outside the self-enforcing range. This probabilistic view of hold-ups is contrasted with opportunism more generally and with moral hazard behavior.

One of my most enjoyable intellectual experiences was working with Armen Alchian on the Klein, Crawford and Alchian [1978] hold-up paper. In this paper I extend the basic framework presented in that paper, pointing out what I now consider to be its shortcomings and providing insights into the nature of hold-ups and the form of contracts chosen by transactors to avoid hold-ups. The major analytical extension entails combining hold-up analysis with my work on private enforcement. Because private enforcement capital is limited and written contract terms are necessarily imperfect, transactors must optimally combine court-enforced written terms together with privately enforced unwritten terms to define what I call the self-enforcing range of their contractual relationship. Hold-ups occur when unanticipated events place the contractual relationship outside the self-enforcing range. This probabilistic framework, where transactors enter contractual relationships knowing that a hold-up may take place (but believing that the expected gains from trade outweigh the expected rent-dissipating costs associated with the hold-up risk), is shown to have important implications for understanding the structure of contracts adopted by transactors in the marketplace.

I. WHY DO HOLD-UPS OCCUR?

I begin with a simple example that illustrates the basic economic forces involved in a hold-up. Assume that a builder constructs a house on a piece of land the builder does not own but, rather, only leases short-term. After the initial land lease expires, the landowner could hold up the builder by raising the land rent to reflect the costs of moving the house to another lot. This example illustrates all the hold-up factors emphasized in Klein, Crawford and Alchian—(a) the builder has made an investment that is highly specific to a particular piece of land and (b) the landowner has taken advantage of the incompleteness of the contract that governs the relationship (in particular, the fact that the lease does not cover future years) to (c) expropriate the quasi-

* Professor, University of California, Los Angeles. I am grateful for comments from Armen Alchian, Harold Demsetz, Andrew Dick, Jon Karpoff, John Lott, Kevin Murphy and an anonymous referee.

Economic Inquiry
(ISSN 0095-2583)
Vol. XXXIV, July 1996, 444-463 ©Western Economic Association International
rents on the builder's specific investment. The obvious question is why anything like this would ever occur; that is, why would someone be so naive as to build a house on land for which they had only a short-term lease?

Our primary goal in Klein, Crawford and Alchian was not to explain the existence of hold-ups, but rather the institutions adopted by transactors to avoid hold-ups. For example, we would expect that builders, anticipating a potential hold-up problem, would decide to purchase the land or at least to sign a long-term ground lease before starting construction. However, we do present some examples in the paper of hold-ups that actually occurred. The implicit reason we give for the occurrence of these hold-ups is transactor ignorance. Apparently, transactors are not always smart enough to choose the contractual arrangement that would eliminate the hold-up problem.

Oliver Williamson provides a similar, but much more explicit answer to the question of why hold-ups occur. When defining "opportunism" he states:

By opportunism I mean self-interest seeking with guile. This includes but is scarcely limited to more blatant forms, such as lying, stealing and cheating. Opportunism more often involves subtle forms of deceit. ... More generally, opportunism refers to the incomplete or distorted disclosure of information, especially to calculated efforts to mislead, distort, obfuscate, or otherwise confuse.\[1\]

For example, the hold-up may have occurred in our illustrative house construction example because the landowner deceived the builder with a low up-front land rental price and vague promises about the future.

Relying on the ability of one transactor to take advantage of the naivete or ignorance of another transactor is a highly unsatisfactory way to explain the incidence of hold-ups. Simple examples of deception, such as a builder constructing a house on land that is only rented short-term, rarely, if ever, occur. More complicated and less obvious examples of hold-ups may sometimes involve the deception of an imperfectly informed transactor. However, explanations of hold-up behavior based upon transactor deception are often either not refutable or clearly inconsistent with the facts. For example, the most extensively cited hold-up example presented in Klein, Crawford and Alchian is the Fisher Body-General Motors case, a transaction between two large, sophisticated business firms with no evidence of any pre-contract deception on either transactor's part.

The Fisher Body-General Motors case concerned a contract signed by General Motors and Fisher Body in 1919 for the supply of automobile bodies by Fisher to General Motors.\[2\] Fisher Body, in order to produce the automobile bodies, had to make an investment in stamping machines and dies that was highly specific to General Motors. As a result, a significant potential was created for General Motors to hold up Fisher. After Fisher Body made the specific investment, General Motors could have threatened to reduce its demand for Fisher-produced bodies, or even to terminate its relationship with Fisher completely, unless Fisher reduced its prices.

The Fisher Body-General Motors case appears analogous to our naive house construction on rented land example. However, contrary to our house construction example, the transactors in the Fisher-General Motors case clearly recognized

---

1. Williamson [1985, 47]. Also see Williamson [1979, 234, n. 3].

2. The description of the Fisher-General Motors contract is taken from Klein et al. [1978, 308–10]. The contractual agreement between Fisher Body and General Motors can be found in the minutes of the Board of Directors of Fisher Body Corporation for November 7, 1919.
the hold-up potential and attempted to take account of it in their contract terms before any specific investments were made. In particular, to prevent General Motors from appropriating the quasi-rents from the Fisher investment by threatening to reduce its purchases from Fisher, the contract included a ten-year exclusive dealing clause. This clause required General Motors to buy all of its closed metal automobile bodies from Fisher for a period of ten years.

Obviously, such a contract had to set the price at which Fisher would supply bodies to General Motors. The transactors agreed upon a formula where the price was set equal to Fisher's "variable cost" plus 17.6 percent. An upcharge over variable costs, rather than a formula based on Fisher's total cost, was probably used because Fisher was selling automobile bodies to many different companies and it was difficult to isolate and measure the capital and overhead costs associated with General Motors shipments. The 17.6 percent upcharge presumably was designed to cover Fisher's anticipated capital and overhead costs.

The Fisher-General Motors contract, therefore, was not totally unsophisticated, as was the short-term land-lease contract in our hypothetical example. However, the Fisher-General Motors contract, as it turned out, was similarly inadequate in preventing a hold-up, albeit to the advantage of Fisher rather than General Motors. After the contract was signed, the demand for automobiles rose substantially. Fisher took advantage of the contract in the face of this large demand increase to adopt an inefficient, highly labor-intensive production process and to locate its body-producing plants far away from the General Motors assembly plant. From Fisher's point of view there was no economic reason to make capital investments when, according to the contract, they could instead hire a worker and put a 17.6 percent upcharge on the worker's wage. In addition, there was no economic reason for Fisher to locate their plant close to the General Motors assembly plant when, according to the contract, they could profit by locating their plant far away from the General Motors plant and put a 17.6 percent upcharge on their transportation costs. The result was automobile bodies that were very costly for General Motors to purchase and highly profitable for Fisher to produce.

The Fisher-General Motors case illustrates why transactors are concerned about hold-ups. When automobile bodies are produced and sold inefficiently, as they were by Fisher, the total gains from trade are reduced. We can expect in such cases that ex post renegotiation of the contract will occur so that, after a lump sum is paid to the transactor engaging in the hold-up, price and cost will return to the efficient level. In the Fisher-General Motors case the contract renegotiation took the form of a General Motors' side payment to the Fisher brothers along with purchase of the Fisher Body company.

Since the probability of such ex post lump-sum transfers will be taken into account by transactors in their ex ante contract terms, these hold-up lump-sum transfers may appear to be of no significance if transactors are risk neutral. However, as the Fisher-General Motors case vividly illustrates, the transactor placed at a disadvantage during a hold-up does not immediately costlessly renegotiate the contract and make a lump-sum payment to the transactor engaging in the hold-up. Real resources are wasted during the hold-up process, as transactors attempt to convince their transacting partners that a hold-up potential does exist and of its magnitude. It is these dissipative, purely redistributive costs associated with hold-up behavior, not the lump-sum transfer

itself, that are wasteful. Because of these costs it is efficient for transactors to design contractual relationships that reduce the likelihood of a hold-up occurring.

The obvious question in the Fisher-General Motors case is why a hold-up occurred—that is, why did General Motors use such an imperfect or incomplete contract which placed it in a position where it could be held-up by Fisher in the way it was? It is much too unlikely an explanation to rely on General Motors' naivete or on Fisher's deception. General Motors and Fisher Body were aware of the hold-up problems inherent in their relationship, and both Fisher and General Motors had to have been aware that the contract they adopted to solve their hold-up problem was "defective" in the sense that it contained obvious malincentives. Yet General Motors and Fisher adopted this incomplete and imperfect contract because they believed it would have been more costly to write a more complete and perfect contract.

II. THE USE OF INCOMPLETE CONTRACTS

General Motors and Fisher knowingly entered into their incomplete contract because they believed that this contract, while imperfect, was optimally designed to minimize the probability of a hold-up occurring. Unfortunately, conditions developed that permitted Fisher to use the contract to hold up General Motors. If General Motors and Fisher had known ahead of time what was to happen, no doubt they would have written their contract to take account of the problems that developed. In that sense the Fisher hold-up of General Motors was unanticipated. However, in an uncertain world where complete contractual specification is costly, transactors use incomplete contracts that deliberately do not take account of every contingency. As a result, transactors knowingly leave themselves open to the possibility of hold-ups.

The costs associated with contractual specification that lead transactors to use incomplete and imperfect contracts involve much more than the narrow transaction costs of writing down responses to additional contingencies. In addition to these extra "ink costs," complete contractual specification entails wasteful search and negotiation costs associated with discovering and negotiating prespecified contractual responses to all potential contingencies. Because most future events can be accommodated at lower cost after the relevant information is revealed, much of this activity involves largely redistributive rent dissipation with little or no allocative benefit. Transactors are merely attempting to obtain an informational advantage over their transacting partners, hoping to place themselves in a position where they will be more likely to collect on (and less likely to pay for) hold-ups. Therefore, rather than attempting to determine all of the many events that might occur during the life of a contractual relationship and writing a prespecified response to each, the gains from exchange are increased by the use of incomplete contracts.

Transactors also use incomplete contracts because writing something down to be enforced by the court creates rigidity. Since contract terms are necessarily imperfect, once something is written down transactors can engage in a hold-up by rigidly enforcing these imperfect contract terms, even if the literal terms are contrary to the intent of the contracting parties. This is what occurred in the Fisher-General Motors case, where the written contract terms that were meant to prevent General Motors from holding up Fisher were actually used by Fisher to create a much greater hold-up of General Motors.

4. These rent dissipating costs during the contract negotiation process are analogous to the costs associated with the purely redistributive oversearching for an informational advantage analyzed in Kenney and Klein [1983].
It may appear that this type of hold-up, where a transacting party uses the court and the threat of litigation to enforce an imperfect contract term that is contrary to the intent of the contracting parties, is different from the type of hold-up that occurred in our house construction example, where the landowner took advantage of the absence of a contract to hold up the builder after the short-term land lease expired. We may wish to think of the court as unable to protect the builder in the house construction case, whereas actually the court is effectuating the hold-up by strictly enforcing the written contract terms in the Fisher-General Motors case. However, although this distinction may be important for contract law, the hold-ups are analytically similar. Both hold-ups are caused by a transactor using the court to take advantage of an imperfection in the contract that governs an economic relationship. In the Fisher Body-General Motors case, court enforcement of the imperfect cost-plus contract sanctions Fisher's attempt to charge General Motors arbitrarily high prices. Similarly, in the house construction case, court enforcement of the obvious imperfection in the contract (namely, that the contract only covers the short term) sanctions the landowner's attempt to charge the builder an arbitrarily high price after the short-term land lease expires.

I am assuming in this discussion that the court only enforces written terms and does not enforce unwritten terms. This is, of course, an oversimplification. Courts interpret both written and unwritten terms when enforcing contractual agreements. However, we can assume that the amount of discretion exercised by the court with regard to unambiguous written terms is limited, and that as transactors add additional things to their contracts the likelihood that the court will effectuate a hold-up by rigidly enforcing these imperfect contract terms increases.

This does not mean that writing down contract terms is not beneficial to transactors. Writing down binding contract terms has the obvious benefit that the court can be used to enforce performance. The idea that court enforcement of explicit contracts may be the mechanism by which a transactor engages in a hold-up merely recognizes that contractual specification not only has benefits but also has associated costs. For some elements of performance there may be no trade-off in terms of added rigidity associated with writing down contract terms. For example, contractual specification is costless when desired performance is measured accurately by the contractually specified term and the term is costlessly observable by the court. However, when transactors must use a less than perfect proxy for performance in a contract there is a trade-off. Including the proxy in the contract not only may help in enforcing the understanding but also may do harm by making the contractual arrangement more rigid.

It is the very benefit of contract specification, i.e., that transactors' hands can be tied with respect to certain variables that might otherwise be used to effectuate a hold-up, that creates the harm of contractual rigidity. As the Fisher-General Motors case illustrates, once an agreement is formalized in a written contract, it cannot cheaply be breached if unanticipated changes occur in the market. The only limit on the cost to General Motors of not performing to the literal terms of the imperfect contract when market conditions deviated substantially from ex ante expectations was essentially General Motors' declaration of bankruptcy.

If, on the other hand, a contractual understanding is not formalized in a written contract, transactors can more cheaply opt out of the agreement if subsequent market conditions deviate substantially from expectations. The understanding is much more flexible because, without the
court forcing transactors to perform to the literal terms of the contract, transactors can renege and only lose the value of whatever transactor-specific investments are present in the relationship. Therefore, at some point transactors may decide to avoid the rigidity associated with court enforcement of written contract terms by intentionally leaving many elements of intended performance unspecified and enforcing these terms instead by a private enforcement mechanism.

III. THE SELF-ENFORCING RANGE OF CONTRACTUAL RELATIONSHIPS

The privately imposed sanction that permits transactors to enforce the unwritten terms of their contracts can be thought of as consisting of two parts. One part is the future loss that can be imposed directly on the transactor if the relationship is terminated. Given the presence of non-salvageable transactor-specific investments, the threat of termination of the relationship implies a potential capital loss equal to the discounted value of the quasi-rents from these investments. For example, if General Motors had terminated (or failed to renew) its relationship with Fisher, they could have imposed a capital cost on Fisher for non-performance equal to the specific investments made by Fisher in the General Motors specific tools and dies.

The other part of the private sanction that is imposed on a transactor who is engaging in the hold-up is the damage to the transactor’s reputation in the marketplace. If the violation of the contractual understanding is taken account of by other transactors in their dealings with this transactor, the transactor engaging in the hold-up will face increased costs of doing business in the future. Potential trading partners will become less willing to rely upon the transactor's promises and demand more favorable and/or more explicit contract terms. For example, if General Motors had held-up Fisher and this was communicated in the marketplace, General Motors would have found it more expensive to purchase inputs in the future.

Each transacting party compares the potential hold-up gain from breaching the contractual understanding with the capital loss from the private sanction. If the hold-up gain is less than the capital cost, then the transactor cannot credibly threaten breach of the contractual understanding. Therefore, although transactors could take advantage of the fact that all the elements of a contractual understanding are not perfectly specified in the written contract, they will not do so and will instead perform in a manner that is consistent with the mutually understood contractual intent.

The magnitude of the private sanctions that can be imposed on each transactor who attempts a hold-up defines what can be called the self-enforcing range of the contractual relationship. The self-enforcing range measures the extent to which market conditions can change without precipitating a hold-up by either party. Changes in market conditions may alter the value of specific investments and, therefore, the hold-up potential, yet as long as the relationship remains within the self-enforcing range where each transactor's hold-up potential gain is less than the private sanction, a hold-up will not take place. Only when changes in market conditions move transactors outside the self-enforcing range so that the one-time gain from breach exceeds the private sanction will the hold-up threat, i.e., the threat of breach of the contractual understanding, become credible. When

5. The private enforcement mechanism upon which the following analysis is based is presented in Klein and Leffler [1981]. Lott [1988] extends the Klein and Leffler model in the spirit of the present paper by introducing random changes in cost or demand which alter the incentive of transactors to perform. A firm’s decision to cheat is also considered to be stochastic in Darby and Karni [1973] and Karpoff and Lott [1993].
this occurs the transactor will not be deterred from breaching even if the transactor expects to be terminated and knows that everyone in the marketplace will think he is a "cheat." This is what occurred in the General Motors-Fisher Body case. Fisher and General Motors found themselves outside the self-enforcing range because of a very large increase in demand by General Motors for Fisher-produced bodies. This increase in demand increased the Fisher hold-up potential so much that it became larger than the private sanction that could be imposed on Fisher by General Motors and Fisher found it profitable to violate the intent of the contractual understanding by taking advantage of imperfect terms of the agreement.

The change in market conditions that permitted Fisher to take advantage of General Motors in this way was presumably unanticipated. When the contract was entered into in 1919 the dominant production process for automobiles consisted of individually constructed, largely wooden open bodies; the closed metal bodies supplied by Fisher were essentially a novelty. After 1919, demand for closed metal bodies grew dramatically, and by 1924 they accounted for about two-thirds of General Motors' automobile sales.6

This unanticipated shift in demand increased the extent by which the contract forced General Motors to rely on Fisher and made it profitable for Fisher to take advantage of the contract to hold up General Motors. The large increase in demand increased Fisher's hold-up potential of General Motors so that it became greater than the private sanction that could be imposed on Fisher by the loss of new and future sales to General Motors and to others in the marketplace that learned about its behavior. If this large change in demand had not occurred, the Fisher-General Motors contract would have been self-enforcing and the malincentives associated with the cost-plus contract terms would not have mattered. Fisher would have known that they could not take advantage of the literal terms of the contract without being punished by General Motors and that the punishment would have been greater than their hold-up gain.

IV. AN ILLUSTRATION: THE ALCOA-ESSEX CASE

The concept of the self-enforcing range of a contractual relationship can be further illustrated by the Alcoa-Essex case.7 Essex, an aluminum cable manufacturer, located its cable fabrication plant adjacent to an Alcoa aluminum production facility, thereby permitting shipments of processed aluminum from Alcoa to Essex in molten form. While the Essex plant location lowered costs, it also created an Alcoa hold-up potential. Alcoa could threaten to hold up Essex by increasing the price of delivered aluminum, thereby expropriating the value of the Alcoa-specific element of Essex's investment, namely, the added transportation cost of receiving aluminum from a more distant supplier and the increased cost of reheating cold ingots.

To protect against such behavior Essex entered into a long-term contract with Alcoa, in which Alcoa agreed to process alumina into aluminum for Essex at specified output rates and to be paid in accord with a predetermined price formula. The long-term pricing formula chosen by Alcoa and Essex tied the price Essex would pay over time to the increase in the wholesale price index for industrial commodities.8 This prevented Alcoa from tak-


8. The wholesale price index was chosen for this contract by Townsend-Greenspan (Federal Reserve Chairman Alan Greenspan's old consulting firm). The same index was also chosen for the contract litigated in Missouri Pub. Ser. Co. v. Peabody Coal Co., 583 S.W. 2d 721 (Mo. App.), cert. denied, 444 U.S. 865 (1979).
ing advantage of Essex by arbitrarily increasing the price after Essex had made its highly Alcoa-specific plant investment.

Unfortunately, the wholesale price index which the parties agreed to use in their contract turned out to be a very poor measure to rely upon. Although the wholesale price index had historically tracked Alcoa’s costs, electricity costs (the principal non-labor cost in aluminum production) began to rise much more rapidly than the wholesale price index after the 1973 crude oil supply crisis. By June 1973 Essex was receiving aluminum from Alcoa at a net cost of less than one-half the contemporary market price of aluminum, resulting in what the judge asserted was an estimated windfall gain to Essex of more than $75 million over the life of the contract.

The enforcement by Essex of the literal terms of this imperfect contract can be considered a hold-up since it can be assumed to be contrary to the original intent of the contractual understanding. Like the Fisher Body-General Motors contract, the long-term contract designed to protect Essex against a threatened expropriation of rents by Alcoa resulted, because of unanticipated changes in market conditions, in a much greater threatened shift of rents to Essex from Alcoa.

Figure 1 graphically illustrates the concept of the self-enforcing range of the Alcoa-Essex contractual agreement. The extent of unanticipated changes in market events is measured along the horizontal axis by the deviation of market prices, $P_m$, from contracted prices, $P_c$. For any deviation of market prices from contracted prices, the resulting associated potential hold-up (by the transactor who has gained by enforcing the literal terms of the agreement) is measured along the vertical axis. Let us assume for expositional simplicity that the contractually specified flow of goods implies that each $1 price deviation from the contract price along the horizontal axis creates a potential hold-up gain with a present value of $1 million. For example, if the market price rises above the contract price by $1, the potential hold-up gain to the buyer, Essex, of enforcing the literal terms of the agreement, which we denote by $H_E$, is $1 million; if the market price falls, say, $2 below the contract price, the potential hold-up gain to the seller, Alcoa, of enforcing the contract agreement, denoted by $H_A$, is $2 million. The potential hold-up gain as market price deviates from contract price, therefore, is represented in Figure 1 by the 45 degree line $HH$.

The self-enforcing range of contractual performance is determined by considering the transacting parties’ private enforcement capital. Assume, for example, that Essex’s private enforcement capital, which we denote by $K_E$, is $5 million (say $4 million from the capital depreciation of Essex’s Alcoa-specific investments and $1 million from the future income loss to Essex of operating in the marketplace with a poorer reputation). Therefore, Alcoa could impose a $5 million cost on Essex if it holds up Alcoa by insisting on delivery at contract terms when the value of aluminum has risen above contract terms. As a

9. The actual cost to Alcoa was substantially higher. The judge calculated the loss to Alcoa by considering Alcoa’s accounting costs, including the cost to Alcoa of constructing the additional plant necessary to fulfill the Essex contract, over the period 1977-87. However, if we consider the opportunity cost to Alcoa by comparing what they could have sold the aluminum for in the marketplace with the price at which they were contractually bound to sell the aluminum to Essex, the amount is much higher. For example, in 1979, when Essex received aluminum from Alcoa under the contract at thirty-six cents a pound, Essex resold some of their aluminum in the open market at seventy-three cents a pound, for a difference of thirty-seven cents a pound. Multiplying this underpricing by the seventy-five million pounds Alcoa was committed to deliver to Essex annually yields an opportunity cost to Alcoa of nearly $30 million in 1979 alone.

10. We are assuming that the court will always enforce the written contract terms, not in the sense that the court would require specific performance, but that the court would award money damages to Essex (if market prices increased) or to Alcoa (if market prices fell) based on the written contract terms.
FIGURE 1
The Self-Enforcing Range of the Alcoa-Essex Contract

Essex's hold-up gain ($H_E$) and private enforcement capital ($K_E$) (million $)

Alcoa's hold-up gain ($H_A$) and private enforcement capital ($K_A$) (million $)
result, Essex cannot credibly insist on receiving the goods at the contracted price as long as the market price deviates from the contract price by less than $5. For example, if the market price increased, say, $3 from the contract price, the $3 million gain to Essex of such enforcement is less than the $5 million loss to Essex from termination of the relationship with Alcoa and communication to the marketplace of Essex's failure to adjust contract terms to market conditions.

The analysis is symmetrical for the case of the potential Alcoa hold-up of Essex when the market price falls below the contracted price. If Alcoa's private enforcement capital loss, $K_A$, is, say, $8 million (consisting of, for example, a $2 million loss of Essex-specific investments and a $6 million loss of market reputation capital), the market price can, in principle, fall up to $8 below the contract price and the contract adjustment in price still be made by Alcoa. This is because of the now credible threat Essex can make to cut its losses by terminating its dealings with Alcoa and communicating to the marketplace the breach by Alcoa of the contractual understanding. The entire self-enforcing range, therefore, covers all market price deviations from the contract price between minus $8 and plus $5. Within this range of price deviations, represented in Figure 1 by the flat portion of the SS schedule between minus $8 and plus $5, contract terms will be "voluntarily" adjusted to the market price without any side payments being made by the transactors to one another.

More generally, the contractual understanding is not likely to require contractual adjustments of all deviations of the market price from the contracted price. Given costly information about changing market conditions, it will be wasteful for transactors to devote resources to search for information and negotiate changes for every small deviation of contract terms from market conditions. Consider, for example, a case of a contractor who, after agreeing to build an additional room on your home for $20,000, informs you at the start of construction that the contract price has to be adjusted upward from the agreed upon price of $20,000.00 to $20,010.00 because of a change in his cost of nails which occurred in the two weeks since the contract was bid, negotiated and agreed upon. You would, of course, not be aware of this cost change nor would it be practical for you to verify his claim. More importantly, you would, with good reason, wonder what kind of contractor you were dealing with—apparently one that intended to engage in significant rent dissipating negotiating activities during the life of the contract. Analogously, we should not assume that Essex is holding up Alcoa if it enforces its contract with Alcoa when the market price exceeds the contracted price by a small amount. It is more likely that Alcoa would suffer reputational penalties if it seeks release from its contract with Essex unless a small upward adjustment in price were made. If contractual adjustment is part of the implicit contractual understanding, the understanding generally is that adjustments are not made unless some sufficiently large minimum disturbance occurs. The contract terms that define the self-enforcing range can be

11. I am assuming that Alcoa can credibly threaten to terminate Essex in spite of the fact that it is costly for Alcoa to carry out such a termination threat because they have made Essex-specific investments. Alcoa can credibly threaten to terminate Essex when Essex threatens a hold-up within what Alcoa believed was the self-enforcing range either because of what Alcoa learns about Essex (that Essex has lower private enforcement capital) or because of what Essex and other buyers would learn about Alcoa if Alcoa failed to terminate Essex (that Alcoa has higher costs of imposing the termination sanction).

12. Some contracts explicitly formalize this by including reopener provisions, where the contract is opened for renegotiation after some market price index moves more than a minimum amount. See Goldberg and Erickson [1987]. Crocker and Masten [1991] provide a general discussion of the contractual mechanisms employed by transactors to flexibly adjust prices in long-term contracts.
thought of as a "contractual constitution" that is not anticipated to be frequently amended. 13

Given the constitutional contractual understanding of the parties and the time necessary to negotiate contractual changes, transactors can find themselves outside of the self-enforcing range if surprises take place—that is, if large and sudden unanticipated changes occur in market conditions. When this occurs, a transactor’s hold-up potential is greater than its private enforcement capital and threats to breach the contractual understanding and enforce the literal terms of the contract are credible. For example, if the positive deviation between market and contract prices is greater than the $5 given by Essex’s reputation capital, Essex can credibly threaten Alcoa with litigation to enforce the literal terms of the contract if it does not receive a side payment in return for modifying the contractual arrangement to coincide with market prices. 14

The magnitude of the necessary side payment settlement will be less than the potential transactor gain represented by the $HH$ schedule in Figure 1. For example, if market prices move above contract prices so that $(P_m - P_c)$ is $6$, Alcoa need not pay Essex $6 million to force Essex to adjust the contract price up to the market price. In the real world we do not observe discontinuous behavior such as no side payment being made when the price deviation is $5.00, but a more than $5 million side payment being made when the price deviation is, say, $5.10. Because of the presence of Essex private enforcement capital, Alcoa can impose a $5 million loss on Essex. In the case of $(P_m - P_c)$ equal to $6$, if Alcoa and the market consider payment to Essex of any settlement greater than $1 million as a hold-up by Essex (i.e., as a breach of the implicit contractual understanding regarding settlements), then we can assume that Alcoa will be able to credibly impose the $5 million loss on Essex and Essex will willingly accept only $1 million to adjust the contract price up $6.00 before continuing the business relationship. The hold-up settlement payment schedule is, therefore, represented in Figure 1 by the schedule $SS^{15}$.

This analysis suggests that when parties enter a contractual relationship they can be thought of as buying what amounts to an option representing the probability of a hold-up occurring. In particular, Essex has purchased a call and has written a put, while Alcoa has purchased a put and has written a call. The defining points of the self-enforcing range can be thought of as the exercise prices of these put and call options that Essex and Alcoa have written and purchased along with their contract. As in standard options pricing theory, 16 the values of these options increase (1) as the value of the ratio of the underlying asset price increases relative to the exercise price (in our case, as the value of the hold-up potential increases relative to the private enforcement capital), and (2) as the variance per period of the asset price multiplied by the number of periods increases (in our case, as the variance of underlying market conditions and the length of the contract increases). Therefore, because hold-ups are costly, when the variance of $(P_m - P_c)$ is high, transactors

---

13. Goldberg [1976, 428] has argued that it is useful to think of transactors designing a contractual arrangement as establishing a "constitution" to govern their ongoing relationship.

14. A relationship outside the self-enforcing range does not imply that litigation occurs. For litigation to take place it is necessary, in addition to the parties being outside the self-enforcing range, for the parties to have sufficient informational differences regarding what they have at stake and what their probabilities of success in court are.

15. I am assuming for simplicity throughout the discussion a threshold model of private sanctions, that is, any and all types of hold-ups trigger the same lump-sum private enforcement penalty.

16. See, for example, Brealey and Myers [1991], chapter 20.
will require more reputational capital or reduce the length of their contract or the specific investments they make. The major difference between option analysis and our hold-up analysis is that by writing particular contract terms transactors not only can vary the exercise price but, as we shall see, they also can vary the underlying probability distribution that determines the value of their options.

V. THE ROLE OF CONTRACT TERMS

The role of contract terms within this framework is very different from the standard economic view of contract terms. The view of contract terms that underlies much of the principal-agent and mechanism design literature, for example, is that contract terms are used to create optimal incentives on some court-enforceable proxy for performance. Optimal but not perfect incentives are created by contract terms because the terms are only imperfect proxies for performance and are assumed to represent the sole elements of performance against which transactors maximize. The problems that arose in the Fisher-General Motors case may be considered as an example of the type of imperfect incentives and associated inefficiencies created when imperfect contract terms are used. However, it is a mystery within this standard framework why Fisher and General Motors would have considered it optimal to choose such clearly imperfect contract terms to begin with.

The problem with the standard economic framework is that court enforcement and private enforcement are considered as alternatives—firms will rely upon one or the other, but never both. Principal-agent models, for example, formulate the contracting problem as if transactors do not possess any private enforcement capital. Therefore, it is not surprising that these models have limited predictive value in explaining real world contract terms. On the other hand, standard economic models of reputational enforcement provide no role for contractual specification. However, given the fact that private enforcement capital is limited, transactors can be expected to use written contract terms and, hence, the assistance of the court, as a supplement to private enforcement.

Unlike standard economic models, the probabilistic hold-up framework presented here implies a fundamental complementarity between court enforcement and private enforcement. When employed together the mechanisms are substitutes in demand in the sense of a positive cross-price effect, i.e., an increase in the price of one increases the demand for the other. For example, an increase in the cost of court enforcement increases investments in private enforcement capital. However, the two enforcement mechanisms are complements in production in the sense of a positive cross effect in production, i.e., an increase in the quantity of one increases the marginal product of the other. The two enforcement mechanisms work better together than either of them do separately.

Within this framework transactors use written contract terms not solely to create an incentive to perform with regard to some court-enforced, contractually specified proxy for performance. Rather, transactors use written contract terms to define optimally the self-enforcing range of their contractual understanding. The goal of contractual specification is to economize on the amount of private enforcement costs.

17. A survey of the principal-agent literature is provided in Hart and Holmstrom [1987]. The complex, contractually specified incentive schemes that solve the agency problem in this literature are also generally claimed to be only second-best because of the presence of transactor risk aversion, which creates a tension between the effect of a contract term in optimally rewarding productive work and in shifting unwanted risk to the agent.

capital necessary to make a contractual relationship self-enforcing by merely “getting close” to desired performance in a wide variety of circumstances (without creating undue rigidity) and to let the threat of private enforcement move performance the remainder of the way to the desired level.

Contract terms can accomplish this goal of economizing on private enforcement capital in two fundamental ways. First of all, contract terms can operate directly to control nonperformance. By defining performance with explicit court-enforceable contract terms, such as the quantity, quality and price of a product that must be delivered, transactors control hold-up behavior by legally “tying their hands” with regard to variables that can be manipulated to expropriate rents from a transacting partner. In the Fisher-General Motors case, for example, these contractual restraints took the form of an exclusive dealing clause with a specified price formula.

This is illustrated in Figure 2, which presents the probability distribution of the hold-up potential in the General Motors and Fisher Body relationship, $f(H)$, which is assumed for expositional simplicity to be related to some scalar measure of ex post market conditions that can be measured along the horizontal axis, as in the Alcoa-Essex case. Figure 2 measures General Motors’ hold-up potential and private enforcement capital to the right of zero and Fisher’s hold-up potential and private enforcement capital to the left of zero. (As illustrated, there is no reason that the probability distribution need be centered on zero.) The shaded area to the right of $K_G$ defines the probability of a General Motors hold-up, the area to the left of $K_F$ defines the probability of a Fisher Body hold-up, and the area between $K_F$ and $K_G$ is the probability that the General Motors-Fisher relationship remains within the self-enforcing range.

Panel A of Figure 2 represents the situation after Fisher has made its General Motors specific investment and a significant General Motors hold-up potential of Fisher has been created. Panel B represents the situation after Fisher and General Motors have negotiated a contractual arrangement which attempts to control the hold-up potentials of the parties. Assuming that the rent-dissipating costs associated with hold-ups are proportionately related to the magnitude of the hold-up, transactors will attempt to minimize these costs when setting contract terms by minimizing the expected value of the hold-up, i.e., the sum of the expected hold-up values associated with the tails of the probability distribution. Because the actual hold-up is assumed to be adjusted downward by the private enforcement sanction that can be imposed on the transactor engaging in the hold-up, General Motors and Fisher can be assumed to be minimizing

\[
K_F \int_{-\infty}^{\infty} (H - K_G) f(H) dH + \int_{-\infty}^{K_F} (H - K_F) f(H) dH.
\]

Panel B illustrates that, although the contract terms substantially reduce the probability that General Motors will hold up Fisher for its specific investment, the contract terms also substantially increase the probability of a very large Fisher hold-up of General Motors if market conditions change dramatically. The contract decreases the probability of being outside the self-enforcing range, but also increases the far tails of the hold-up distribution. This corresponds to the rigidity...
A. Fisher makes a General Motors specific investment

B. Contract terms are set

FIGURE 2
The Hold-Up Probability Distribution in the Fisher-General Motors Case
costs associated with literal court enforcement of imperfect contract terms discussed above. It is because of these rigidity costs associated with contractual specification that, after some point, each contract term which the transactors decide to use involves a cost/benefit calculation. We can expect, therefore, that the degree of contractual specification will be lower the greater the private enforcement capital possessed by the transactors.

Where private enforcement capital is larger, contracts will be “thinner,” with transactors writing out only the essential elements of the agreement, or perhaps even proceeding on the basis of a verbal understanding and a handshake; where private enforcement capital is smaller, written contracts will be “thicker,” with transactors attempting to specify more elements of performance and provide for more contingencies. For example, contracts between Japanese companies can be expected to be much less completely specified than a similar contractual relationship between U.S. companies. Japanese companies generally possess large amounts of private enforcement capital in transactions with one another because of the significant sharing of information regarding performance among Japanese companies and the large potential “loss of face” in the Japanese marketplace if it is perceived that one has engaged in a hold-up. When a large amount of private enforcement capital is present, there is a lower likelihood of being outside the self-enforcing range and, therefore, less need to bear the costs associated with contractual specification.

The second fundamental way in which contract terms can reduce the expected hold-up probability is by shifting private enforcement capital between transactors. Rather than directly attempting to reduce the hold-up potential, contract terms can shift the location of the self-enforcing range so that private enforcement capital coincides more accurately with the transactors’ hold-up potentials under likely future conditions. By more closely relating actual private enforcement capital to likely requirements, transactors widen the ex post market conditions that are likely to fall within the range where performance remains assured.

This view of contract terms, as a way to increase the effectiveness of a self-enforcement mechanism by shifting private enforcement capital between transactors, has much greater predictive power in explaining the contract terms we observe in the world than the standard view of contract terms. For example, consider the grant by a franchisor to a franchisee of an exclusive territory. The standard economic view of contract terms would emphasize the effect of the exclusive territory in creating the correct incentive on the franchisee to perform due to the increased customer repeat sale created by the exclusive territory. The exclusive territory thereby reduces the franchisee’s incentive to “free ride” on other franchisees. However, the exclusive territory more importantly also creates a franchisee premium stream and, therefore, gives the franchisee something valuable to lose if it is terminated by the franchisor for non-performance.20

The exclusive territory and the associated payment of a premium stream from the franchisor to the franchisee can be thought of within our private enforcement framework as a shift of private enforcement capital from the franchisor to the franchisee. The franchisee now has more to lose if it is terminated, namely the franchisee loses the discounted value of the expected premium stream associated with the exclusive territory; and the franchisor now has less to lose if it unfairly terminates the franchisee, namely the franchisor saves the discounted value of the larger expected premium stream that it no longer has to pay the fran-

This shift in private enforcement capital, because it more accurately aligns the transacting parties' enforcement capital with likely future franchisee hold-up possibilities, expands the self-enforcing range. That is, the exclusive territory increases the probability that ex post market conditions will fall within the range where performance can be privately enforced.

This example is illustrated in Figure 3 where, once again, we assume there is a market realization of a single hold-up potential, with $H_r$ and $H_f$ representing the franchisee and franchisor hold-up potential and $K_r$ and $K_f$ representing the franchisee and franchisor private enforcement capital, respectively.

If we assume that the shifts in private enforcement capital from the franchisor to the franchisee, represented by $x$ in Figure 3, are dollar for dollar, in the sense that every dollar increase in the franchisor hold-up (in every ex post state where a franchisor hold-up occurs) implies a corresponding saving of one dollar in the franchisee hold-up (in every ex post state where a franchisee hold-up occurs), then this process of shifting private enforcement capital will occur until

$$
\frac{d}{dx} \int_{K_f-x}^{\infty} (H - K_r - x)f(H)dH
$$

$$
+ \int_{-\infty}^{K_r+x} (H - K_r + x)f(H)dH = 0.
$$

This implies that in equilibrium the probability of engaging in a hold-up will be the same for the franchisor and the franchisee, or

$$
1 - F(K_r - x) = F(K_r + x).
$$

Intuitively, if the probabilities are unequal, the overall probability of a hold-up occurring in the relationship could be reduced by shifting private enforcement capital from the transactor with the lower probability to the transactor with the higher probability.

Another example of contract terms efficiently shifting the self-enforcing range is the use of a contract to determine which transactor makes the transaction-specific investment. In general, the transactor who will make the specific investment is determined by comparing the likely future private enforcement capital requirements (i.e., the hold-up potentials) of each party under alternative likely contingencies with the amount of private enforcement capital that each transactor has available. Therefore, it is usually the transactor with the smaller private enforcement capital, such as Fisher Body, that will make the specific investment. The larger firm, because of its increased repeat transaction frequency, generally has more private enforcement capital and hence increased credibility that it will fulfill the contract.

Similar reasoning explains why many contracts may appear "one-sided" or "un-

22. This result assumes that the real costs associated with hold-ups are related to the magnitude of the hold-up and that this relationship is the same for both transactors. For example, the real costs may be proportional to the magnitude of the expected hold-up with both transactors having the same proportionality constant. This will not be the case, however, if any franchisee hold-up gains entail primarily distribution effects with relatively little real costs compared to franchisor hold-up gains of the same magnitude. For example, a franchisee hold-up may entail some costs associated with loss of product reputation on the part of consumers and the cost of replacing the franchisee, while a franchisor hold-up may entail much larger costs associated with the efficiency of the franchising arrangement compared to the next best alternative marketing arrangement. Since contract terms are set to minimize real costs (and not hold-ups), this would imply a higher probability of a franchisee hold-up than a franchisor hold-up in equilibrium.
fair."^{23} For example, if General Motors possessed substantially more private enforcement capital than its suppliers, it could avoid the rigidity associated with contractual specification by not promising anything in writing in its supply contracts. The contracts would appear to be one-sided, but this would not substantially increase the probability that it would hold up its suppliers. Another example is employee termination-at-will clauses. Because it is extremely difficult to specify in a court-measurable way all the conditions of adequate employee performance, and court enforcement of imperfect terms entails all the rigidity costs discussed above in addition to the costs of artificial record keeping, litigation expense and time delay (during which the employee may impose additional costs on the firm), it may pay both parties to use a termination-at-will contract. Although such a contract may seem unfair, with the employee vulnerable to a potential hold-up by the employer, one cannot interpret such arrangements without recognizing that private enforcement capital, in addition to the explicit contract terms, also governs the relationship and that the employer may have sufficient private enforcement capital to define a sufficiently broad self-enforcing range.

VI. A PROBABILISTIC VIEW OF HOLD-UPS

This analysis implies that contractual arrangements can only be understood if we recognize that transactors optimally design their contracts to combine court-enforced written contract terms with self-enforced unwritten terms. Given the par-

---

23. See Klein [1980].
ticular contract terms they choose and the private enforcement capital they possess, transactors expect the relationship to remain within the self-enforcing range, where market conditions can change and the parties will perform as intended. However, transactors also know at the time they enter their contractual relationships and make their specific investments that their private enforcement capital is limited, that their written contract is imperfect and incomplete and, therefore, that there is some probability of a hold-up occurring. In particular, transactors know that there is some probability that market conditions may change sufficiently (and the value of the quasi-rents accruing to one of the parties increase sufficiently) so that one party will find it in its interest to engage in a hold-up.

For example, the unanticipated change in market conditions that occurred in the Alcoa-Essex case was the 1973 crude oil “shock” which led to a quadrupling of the price of aluminum. This was a contingency that was not covered in the Alcoa-Essex contract and once this unanticipated event occurred, the short-run gain to Essex from the failure to adjust the price became greater than the depreciation of Essex’s private enforcement capital. It became profitable for Essex to violate the intent of the contractual understanding by demanding enforcement of the contract as written. The contractual relationship had moved outside the self-enforcing range.

This probabilistic view of hold-ups should be contrasted with Williamson’s view of opportunism which, he asserts, is equivalent to moral hazard behavior.24 First of all, if moral hazard behavior is fully anticipated, it should be considered as merely part of the price. For example, if employees sometimes take pencils home for their personal use, the pencils are part of the wage and working conditions of the job. (And if the value of the pencils to the employees exceeds their cost and employers do not expend resources to prevent the taking of them, there is nothing inefficient about this form of compensation.)

More importantly, identifying moral hazard behavior with a hold-up blurs a fundamental analytical distinction. Consider the example of the demand for medical services by individuals with health insurance, a commonly cited case where moral hazard behavior takes place. Health insurers who, after writing the best contracts they can, knowingly accept the fact that their policyholders will take advantage of the low marginal price in these contracts to increase their demand for health services. In spite of this behavior, the transacting parties still find it in their interests to enter the relationship. It is true that the moral hazard behavior is “non-performance,” in the abstract, ideal sense that if sufficient private enforcement capital existed or perfect contract terms could be written, the behavior would not exist. However, the behavior is fully expected and, I would maintain, has nothing to do with a hold-up.

A hold-up is a particular kind of transactor non-performance distinct from moral hazard behavior. Specifically, a hold-up, as opposed to moral hazard behavior, requires unanticipated events. Transactors may recognize that there is a positive probability of a hold-up occurring and may know ahead of time that if particular conditions develop, a hold-up will take place. However, hold-ups are always surprises in the sense that the particular conditions that will lead to the hold-up are considered unlikely and, because it is costly to negotiate and specify contract terms, these unlikely conditions are not taken account of in the contract. If the transactor being held up had expected that market conditions would develop to

24. Williamson [1985, 51, n. 8]. Considering moral hazard behavior as opportunism is inconsistent with Williamson’s definition of opportunism in terms of deception and guile (fn. 1 supra). Williamson attempts to reconcile this inconsistency by claiming that one should consider deception broadly.
place the relationship outside the self-enforcing range, the transactor would have written a different contract or would not have made the specific investments to begin with. Moral hazard behavior, on the other hand, is fully anticipated; presumably transactors already have written their "best" contract. Moral hazard behavior is present not because of any unanticipated events, but merely because measurement and monitoring costs make it uneconomic for the transactors to write and enforce a perfect contract that would yield the idealized behavior that would exist in a costless contracting world.

VII. CONCLUSION

The analytical framework I have presented here, like the framework originally presented in Klein, Crawford and Alchian, should be judged by how well it assists us in explaining the particular contractual arrangements adopted by transactors in the marketplace. I believe that this framework provides an understanding of contracts which has much greater predictive power than the commonly accepted economic framework. Instead of thinking of contract terms as providing transactors with the correct incentives to perform with regard to particular contractually specified margins, this framework suggests that transactors choose contract terms, including vertical integration, in order to economize on their limited (and often unequal) amounts of private enforcement capital and thereby to define an optimal self-enforcing range for their contractual relationship. Rather than sharply distinguishing between two-party versus three-party enforcement mechanisms, i.e., between private reputational enforcement and court enforcement, contract terms can be explained as devices to assist transactors in assuring that sufficient private enforcement capital exists relative to the hold-up potential under the broadest range of likely ex post market conditions.

Finally, it may be argued that the framework presented here provides an economic justification for relational contract law. Because hold-ups are caused by rigid court enforcement of the imperfect and incomplete terms transactors choose to write in their contracts, it may be efficient for courts to use increased discretion in enforcing the terms. The court would appear to be able to provide an effective substitute for the transactors' limited private enforcement capital by taking more explicit consideration of the intent of the contractual understanding, rather than merely rigidly enforcing the written terms of the contract. As a result, relational contract law, by leading courts to flexibly interpret contracts with the goal of avoiding hold-ups, could, in principle, expand the self-enforcing range of contractual arrangements.25

However, while flexible court interpretation may seem appealing in theory, in practice courts cannot employ increased discretion without losing some of the benefits associated with predictable court enforcement of written contract terms, namely the ability of transactors to tie one another's hands with respect to particular behavior and to create rental streams by shifting their private enforcement capital. Therefore, attempts to use increased court discretion to prevent hold-ups may, in fact, have the opposite effect of increasing hold-ups. It is difficult for judges, as it is for economists, no matter how smart and well-intentioned they may be, to understand fully the economic intent and purpose of all the complex contractual terms transactors use in their contracts. Moreover, since many of these contract terms may appear superficially unfair or unconscionable, there may be a temptation not

to enforce them. However, as noted above, these contract terms may be key elements of the transactors' joint attempt to define efficiently the self-enforcing range of their contractual relationship. Therefore, while contract law can, in principle, economize on transactors' limited private enforcement capital, one must proceed with caution down this road to avoid narrowing the self-enforcing range of contractual relationships.

REFERENCES


