6-24-2005

Theories of the Employment Relationship: Choosing between Norms and Contracts

Michael L. Wachter
University of Pennsylvania Law School, mwachter@law.upenn.edu

Theories of the Employment Relationship: Choosing between Norms and Contracts

MICHAEL L. WACHTER

University of Pennsylvania

The employment relationship is the construct at the heart of any industrial relations system. Most workers are employed inside firms, and their dealings with their employer are thus partially outside the protections offered by whatever competitive forces operate in the labor market. This appears to leave nonunion workers without much protection against unfair outcomes and unfair dealings to the extent that the firm has an unequal bargaining advantage. Although there are many commentators who believe that the system works badly, it can be argued that nonmarket mechanisms have evolved that provide substantial, if incomplete, protection to workers. In this paper I analyze labor market theories that address this issue, particularly how an employment relationship can work when it appears to be entirely one-sided and stacked in favor of the firm. In so doing, I analyze the choice of norms versus contracts as a method of forming agreements to guide the relationship and the extent to which these methods are either self-enforcing or require judicial enforcement.[1]

To address this question, it is necessary to analyze the three types of labor market relationships that are prevalent in the economy. The first is the labor service market, an external labor market in the sense that its activities take place outside the boundaries of an individual firm. Within this market are very different relationships in terms of their formality, spanning the subcontracting market at one extreme and the markets for personal services and for spot labor on the other extreme. The second type is the employment relationship inside the nonunion firm. This involves a firm and its employees, rather than two independent agents as in the prior case. The third type is the employment relationship inside the union firm, which is distinct from the prior case because of the substantial regulatory apparatus of the National Labor Relations Act (NLRA).[2]

The parties' relationships in each of these markets are markedly different from one another. In the labor service market, two independent parties reach an agreement to transact, that is, to provide labor services in return for a package of terms and conditions of employment. The terms and conditions reflect an agreement, and since the relationship is entirely voluntary, it is assumed to represent a joint profit-maximizing point for the parties. With respect to enforcement, these agreements are intended to be primarily self-enforcing. However, since the agreements are contracts in the legal definition of that term, state enforcement by courts provides the ultimate enforcement mechanism. The governance structure for the labor service market is thus contracting based.

In the employment relationship inside nonunion firms, the agreements are frequently determined by the firms rather than bargained over term by term by the parties. Although labor contracting in the external market sometimes features take-it-or-leave-it terms, the nonunion sector operates almost entirely on this basis.[3] As in the labor-contracting markets, the terms are

intended to be self-enforcing. However, unlike the former case, the terms are essentially norms of behavior[4] and thus are enforced nonlegally. Since the norms are effectively determined by the firm, the governance structure of the nonunion employment relationship is hierarchical. There are a few exceptions where legal enforcement is available to enforce government-mandated terms, such as the rules governing employment discrimination, the funding of retirement plans, and occupational safety and health.[5]

In the union labor market, the parties bargain over the individual terms of the contract rather than having them be determined by the firm. Not only are the agreements legally enforceable contracts, but the relationship between the parties is also heavily regulated by statute in a manner quite different from the law of contracts. Of particular importance is that the relationship is not entirely voluntary. Workers have a right to unionize, and if they do, the firm must bargain with the union over the mandatory terms, including wages, hours, and other conditions of employment.

Why do these very distinct forms exist? This paper provides an answer to this question. I presuppose that the primary purpose of each of the alternative structures is to maximize the value of the wealth available to the parties. To be successful, each of the structures has to resolve the four key features of industrial organization theory: match-specific assets, asymmetric information, risk aversion, and transaction costs (Rock and Wachter 1996, 1999).[6] The central question then is, How do the terms of the employment relationship work to protect the parties' agreements with respect to the four characteristics? Are the terms of the agreement interpretable as maximizing joint profit? Are the enforcement protections adequate?

Historical Antecedents to the Enforcement Issue

Historically, labor market analysis addressed the bargaining process, albeit indirectly, when it inquired whether the treatment of workers by firms was "fair" by some metric. Typically, such inquiries have discussed the difficulties of reaching fair results given the disproportionate bargaining power available to firms and the potential for the arbitrary use of that power.[7]

Numerous reasons have been offered for workers' not receiving their just deserts. In discussing this question, the important distinction is between firms and markets. Some commentators bemoan the outcomes of competitive markets that generate wages and working conditions below what those commentators believe to be fair. But I take this line of inquiry to be limited, given the many positive attributes associated with competitive outcomes. Moreover, policy cannot improve on those outcomes.

A more useful approach is to criticize labor market outcomes as being noncompetitive, which many commentators have historically done. This criticism was best captured by classical monopsony theory; monopsony would generate equilibrium market wages that are below competitive levels.**QQ AU: OK as edited? I don't think you intended to say that the theory generates low wages. XQQ*\ In classical monopsony markets, individual firms exercise market power and can set the wage below the value of the worker to the firm (Boal and Ransom 1997). If this is the problem, the optimal solution is to make the market more competitive. Antitrust enforcement policies that make markets more competitive, rather than monopsonistic, improve outcomes and result in higher wages and employment.

Unions are another solution to classical monopsony. Unions raise wages above the level that firms would voluntarily pay on their own accord. If labor markets are monopsonistic, union-

inspired wage increases move the labor market closer to a competitive equilibrium and increase employment. Historically, when classical monopsony theory was more popular, unions were viewed as a procompetitive market force (Kaufman 2002; Wachter 2003). It is difficult to maintain this position today.

Current models of monopsony recognize that the source of the upward-sloping supply curve is labor market friction such as costs associated with recruiting and retaining workers. In these models, it makes no sense to think of the monopsony wage as being “too low” because of market power, and there is no reason to assume that a higher wage results in an efficiency gain (Manning 2003). Consequently, union-inspired wage increases do not represent a countervailing power that improves the functioning of the labor market.

The most important line of inquiry, which is at the heart of the debate over labor market outcomes, is to focus on the problems that occur at the level of the firm, not the market. At the firm level, a failure to achieve just deserts can potentially arise because of the very nature of the corporate form. The corporate form creates and enforces an organizational structure of centralized management. The corporate directors or the executive officers they designate wield hierarchical governance powers. By design, management calls the shots in the sense of organizing economic activity inside the firm. In the nonunion firm, centralized management allows the firm to dictate the terms and conditions of employment and to create an enforcement mechanism that resolves intrafirm disputes in an approved manner (Rock and Wachter 2001).

It is typically this one-sided hierarchical structure that is criticized when commentators discuss issues of unequal bargaining power, arbitrary authority, and hierarchical governance. Inside the firm, workers may not be adequately protected by competitive forces acting at the market level. First, employees are asked to make commitments to their jobs, which makes them

relatively immobile. Second, workers are likely to have less information than the employer about labor market conditions, including the value of the worker to the firm and prevailing conditions elsewhere in the labor market. Finally, employers can exercise their power in ways that take advantage of workers' lack of bargaining power. While some workers receive the optimal market wage, others may not.

The primary policy mechanism for dealing with hierarchy inside firms has been labor unions and the collective bargaining mechanism that they provide and champion. Unions bring intrafirm power sharing so that both the terms of the labor agreement and the enforcement mechanism are a result of collective bargaining, which reduces the hierarchical power of management over labor matters. By writing enforceable collective bargaining contracts, unions can prevent employers from exercising their power arbitrarily, harming individual workers. By using their rights to collect information related to the collective bargaining process, unions can equalize the information available to workers.

In analyzing these three forms, I will make use of the labor-contracting literature.[8] This literature is efficiency oriented, asking whether stylized employment practices can be interpreted as a way of dealing with market problems such as the arbitrary use of power, informational asymmetries, relative risk aversion, and the need to make match-specific investments. More generally, it asks whether the observable stylized features of labor market relationships are interpretable as the parties' resolution of these market problems in a joint profit-maximizing manner. In this literature, issues of fairness and unequal bargaining power are recast into different terms with more precise meanings, namely, opportunistic behavior and asymmetric information.

To date, the labor-contracting literature has paid little attention to the enforcement

question. Specifically, enforcement issues are viewed as unimportant because the stylized labor-contracting arrangements are largely self-enforcing. To the extent that third parties assist in enforcement, it is through reputational effects on other firms or workers. Judicial enforcement, although mentioned, plays no interesting role. In this paper I pay particular attention to the enforcement question.

The Four Factors of the Labor-Contracting Relationship

In this section I rely on the labor-contracting literature to examine how the parties attempt to resolve problems arising from their need to make investments in their relationship (i.e., match investments), their differences in risk aversion, asymmetric information, and transaction costs.[9]

The first of the factors is investments in the relationship, that is, match-specific investments. Investments in the match provide the basic rationale for long-term attachments between a firm and an employee or between a contractor and a subcontractor. Effectively, the employees or subcontractors become more valuable in their current job than they would be if hired anew in a different job. The original explanation for this job-specific value was that such employees had job-specific training, whether formal or learning by doing, so that their increased productivity was a result of knowing more about how to perform the job. But in ongoing relationships, much more is involved in making established workers more valuable than new workers. For example, when circumstances change, an understanding of how the parties have responded to unanticipated events can lead to a greater ability to resolve these difficulties. Many other examples of the intuitive capital that workers attain with tenure can be given. It is best to refer to the broader scope of knowledge and understanding as *match investment* rather than the

traditional term *job-specific training.*

The difficulty that arises when match investments are present is the holdup problem. Workers are worth more in their current job than they are in the external labor market. The gap between an employee's value on the current job and the employee's value in the external labor market creates a quasi rent that is subject to *ex post* expropriation by one party. Either party could threaten to terminate the relationship if not given a larger share of the profits. Since match-specific investments create value for both firms and workers, there is an efficiency gain if the parties can make the appropriate level of investment without fear of opportunistic behavior by the other party.

The general solution is for both parties to invest in the match so they incur sunk costs. Once sunk costs are incurred, both parties lose if the ongoing relationship is terminated. The result is that neither party can credibly threaten to opportunistically terminate the relationship, since doing so would result in a loss to the threatening party. Joint investments are thus a self-enforcing agreement.

The second of the four factors is asymmetric information. Information is asymmetric when it is relatively more costly for one of the parties to observe or monitor the quantity of inputs or outputs, the state of technology, or product market. In order to maximize the parties' joint profits, the information needs to be used by the parties so that they can adjust to whatever conditions actually exist. The difficulty is that since the disadvantaged party cannot verify the informed party's claim, the informed party can misstate the actual information to put itself in the best position possible. That is, the informationally advantaged party can use the power opportunistically to improve its return, even if it means decreasing the size of the joint surplus available to both parties.

This creates a dilemma. Efficiency requires that the party with the correct information be given the right to collect the information and to use it on behalf of the parties' joint interests. But how can the disadvantaged party be protected from the misuse of the information advantage? The general solution is to restrict the channels through which the informationally advantaged party can use the information. While the advantaged party can use and profit from its information, it cannot do so in a manner that harms the uninformed party.

Are such mechanisms available? The answer is generally yes, although the mechanisms may not be perfectly self-enforcing. As an example of such an arrangement, when product market conditions are known to one party (say, the firm) but not to the other (say, the worker), the agreement allows the firm to alter the amount of labor it purchases but not the wage rate. For the mechanism to be self-enforcing, the firm cannot reduce the wage rate. If the firm were able to reduce the wage rate, it could falsely claim that product demand had decreased and hence that it needed to reduce wages to reduce labor costs. However, since demand had not declined, the firm could maintain employment and output levels. Profits would increase because the cost of labor would be lower as a consequence of the wage reduction. On the other hand, if the firm is forced to reduce employment or hours of work in response to a proclaimed downturn in demand, its output and thus its profits fall. This type of agreement is self-enforcing because the firm does not have an incentive to misstate market conditions, as this would result in a decrease rather than an increase in profits. Thus, the firm will not act opportunistically.

Risk aversion is the third of the four factors. Risk aversion arises when one of the parties, typically the worker, is more risk averse than the other. Efficient risk bearing thus requires that the workers' returns be smoothed so that their own income is affected only by their own performance and not by exogenous (to them) fluctuations in the revenue and profits of the firm.

Although problems of risk aversion are partially resolved by smoothing income, there is no perfect solution because of the problem of moral hazard. Specifically, if the worker's return were entirely guaranteed, the worker could reduce work effort without fearing a reduction in income.

Solving the problem of risk aversion is made difficult by the presence of asymmetric information. Since worker's behavior is imperfectly known to the firm, the firm cannot be certain when the worker is shirking or the external workplace environment is adverse. The resulting solution leaves the worker with a greater variance in income than would otherwise be desirable.[10]

The solution to risk aversion points out a problem with all the self-enforcing solutions, namely, that they are all second best in the limited sense that none would be adopted in a world where behavior such as opportunism, asymmetric information, and moral hazard were not present. For example, absent information asymmetries, the firm facing a decline in labor demand might best maximize its own profits and the workers' utility by a reduction in both hours of work and the wage rate. Forcing the firm to make the entire adjustment through a reduction in demand is thus second best. However, second-best solutions are not necessarily market failures that give rise to policy improvements. Information asymmetries, potential opportunism, and moral hazard are real economic costs just like any other economic cost, such as workers' insistence on being paid to work. Consequently, the self-enforcing arrangements worked out by the parties are arguably first best, given the restricted set of solutions available to them.

Transaction costs are the costs associated with negotiating, writing, and enforcing contracts. High transaction costs occur when the parties interact frequently, when the interactions are connected rather than independent events, and when the environment over which the parties interact evolves over time. These conditions are all present in an ongoing relationship.

The greater the number of contingencies that affect the relationship over time, the greater is the cost of contracting. High transaction costs pose a threat to the potential surplus that the relationship can generate.

The contracting problem is exacerbated when the value at stake in each individual contingency is low. When the transaction is a low-value event, the benefit of contracting to protect the transaction is low, and hence even moderate contracting costs are detrimental to joint surplus. Transaction costs are also higher the more match assets and asymmetric information in the relationship. Match investments and asymmetric information give rise to the potential for opportunistic behavior, and regulating this behavior requires a more detailed and hence costly contract.

Overall, the main problem created by the high transaction costs of contracting is that they can reduce or erase the joint surplus available from creating and maintaining the ongoing employment relationship. It is the presence of transaction costs that make the enforcement issue complex. If match investments, asymmetric information, and risk aversion are present but transaction costs are low, the parties can write a complete contract. This means that every possible future state of the world is known and addressed by the contract or that the parties can fully allow recontracting when the environment changes. When a contract is complete, disputes are easily resolved, since the courts can enforce the terms of the contract.

When transaction costs are high, the parties face a discrete choice. Writing contracts to address all potential circumstances becomes costly. If the costs are high enough, the parties may decide on an alternative mechanism for protecting the potential surplus from the relationship. Each of the three institutional structures discussed is best suited for a very different transaction-cost setting. Consequently, as I argue later, when the parties decide on how to handle transaction

costs, they are effectively choosing an institutional structure, and since each of these structures carries its own enforcement mechanism, the parties are also choosing among alternative enforcement mechanisms.

The Choice of Using Markets or Firms

From the perspective of enforcement issues, the choice among alternative organizational structures is critical. In this section, I analyze the possibly most important organizational decision: whether to conduct the transaction inside the firm or outside the firm. To understand the striking differences between the enforcement mechanisms protecting intrafirm and extrafirm transactions, it is worth looking briefly at the existing theory of the firm.

Two Prevailing Models

There are two prevailing models of the firm: the property rights theory and the transaction cost theory. Although the theories were developed separately, they are highly complementary and, indeed, for our purposes can be viewed as a single theory. The property rights approach focuses on the role of physical and intangible capital and posits that the core of the firm is best defined by the physical and intangible capital over which the firms has residual control rights. The transaction cost approach focuses more on human capital and the optimal degree of vertical integration, that is, which labor suppliers should be brought inside the firm and into the employment relationships and which should be left outside the boundaries of the firm. In their respective domains, these approaches share critical assumptions and develop complementary insights.

The critical insights shared by the two models are the need to regulate residual control

rights and the impossibility of doing so by contract. Even in a relationship where the parties have contracted over a number of predictable and verifiable contingencies, there will inevitably be contingencies over which the parties cannot contract or have not contracted. Particularly in an ongoing relationship, in which the environment evolves over time, this residual class of contingencies can become large and can include many of the central problems facing the relationship. By its very nature, the residual class of contingencies contains those contingencies that make contracts very expensive to write. Consequently, any existing contract is necessarily incomplete (Holmstrom 1999).

How then, are disputes resolved when they involve residual contingencies over which the parties have not contracted? The answer is that one of the parties will bargain and effectively purchase the rights of residual control. The party that purchases the residual control rights will then use its own mechanism for resolving problems involving contested residual contingencies.

Property Rights Theory

Let us first look at the problem of residual control rights**QQ AU: OK? Else what problem? XQQ*\ from the perspective of physical (tangible) and intangible capital. The conclusion of the property rights theory is that a firm is defined by the physical and intangible capital over which it has residual control rights. Alternatively stated, at the core of the firm are these residual control rights.

Take the case of the development and production of a network switch by the hypothetical company NetSwitch. Numerous tasks need to be accomplished before the product can be marketed. For example, machinery has to be built**QQ AU: OK? Else please confirm that you mean that the switch-producing firm itself would not purchase the machines but rather**

that some other firm would buy them. XQQ to produce the switch, and if the switch is an intermediate product, it has to be integrated**\QQ AU: Cut OK? NetSwitch is producing the switch, right, not leaving production to some other company? Or do you mean that they'd leave the integrable *design* to some other firm and simply produce according to some other firm's design specs? XQQ** into other equipment. Although NetSwitch may be able to do all the tasks, it is likely that some tasks, including the two just mentioned, will be left to other firms. More specifically, few firms fully integrate the process from production of intermediate good through the distribution system to final consumers. In dealing with the various cooperating firms, some rights of control are necessarily contracted away. For example, when NetSwitch deals with the equipment seller, it may agree on specifications for the equipment and give the equipment maker some "control" rights, such as the decision of how best to build the machine.**\QQ AU: The machine that produces the switch or the machine into which the switch is installed? XQQ** This major transaction is typically protected by contract. Similarly, NetSwitch may contract with the distributor or end user of its product. Here again, certain control rights would be transferred to the distributor or end user.

With several firms involved, ambiguities can arise as unanticipated contingencies occur. If a contingency is unanticipated, it cannot be contracted for in advance. Who gets to decide what happens in this circumstance? Alternatively stated, who has the control rights to resolve the unanticipated contingent-state problems? This is the core problem that the property rights model addresses. The solution proposed by the model is that one of the parties must buy the residual control rights. The party that owns the residual control rights then gets to decide the outcome when a dispute arises within the range of contingencies where the contract is incomplete (Hart and Moore 1990).

The “owner” of the residual control rights is best described as the owner of the asset. The term *owner* usually refers to the individual or entity that gets to “call the shots.” The owner is thus the person or entity that has the right to direct the use of the asset as long as that use does not infringe on those rights that have previously been contracted away. The core of the firm is thus the assets over which the firm has residual control rights. Indeed, at the core of each of the firms in the network switch example, whether the supplier of machinery,**\QQ AU: Again, production machinery, or end-product machinery? XQQ** the end user, or NetSwitch, are the assets over which each of the firms has residual control rights.

If a dispute arises between parties having contractual rights to an asset, the dispute can be referred to the courts for resolution. The threshold issue for the court to determine is whether the dispute is over matters that are covered by the contract. If the matter is covered by the contract, the court resolves the dispute by applying the terms of the contract. However, if the dispute involves residual control rights that are not covered by the contract, the owner of the residual control rights decides the dispute by fiat, that is, by exercising the power that comes with ownership.

In my example, we can assume that NetSwitch will purchase the residual control rights because it has the most at stake and the best overview of the switch’s potential. At the core of NetSwitch are the residual control rights over the creation, production, and sale of the switch.

Transaction Cost Theory

The transaction cost theory is similar to the property rights theory in that it focuses on residual control rights in a world of incomplete contracting. The difference is that the transaction cost theory deals with labor rather than capital inputs. With this basic difference, the theory is

otherwise remarkably similar.

The firm that owns NetSwitch**QQ AU: Is NetSwitch not the firm? Do you mean that NetSwitch is a subsidiary of some parent firm, or do you mean the firm that owns the network switch, that is, NetSwitch? XQQ***** will want to use labor inputs on NetSwitch's owned capital. The problem posed by organizing the human capital is similar to that involving the physical and intangible capital. Some labor services can be contracted for. But here again, not all contingencies can be anticipated and described in a contingent-state contract. Residual issues will inevitably arise. Consequently, whereas the firm will contract for some labor services, other labor services will be brought inside the firm.

The manner in which the parties cope with inevitable contractual incompleteness provides the basis for a positive theory of the employment relationship and provides its distinctiveness and differences from the labor service and collective bargaining modes.[11]

The theory posits that the determination of which relationships are brought inside the firm depends on the transaction costs involved in the relationship. When transaction costs are low, the parties can write contingent-state contracts to protect the integrity of their transactions. Transactions can thus be left in the market, with the market providing the parties with unequaled high-power incentives for joint profit-maximizing**QQ AU: OK? Else what is maximized? XQQ***** behavior. In addition, the parties can rely on market information to estimate asset values and opportunity costs. With information symmetrically available to the parties, the potential for opportunism is reduced, and the reliability of third-party enforcement, should that prove necessary, is increased (Williamson 1996).

Transaction costs are high in the employment relationship due to a full range of factors such as large numbers of match-specific assets and the high degree of information asymmetry

(Williamson, Wachter, and Harris 1975). When transaction costs are high and contract governance is too expensive, the relationships are brought inside the firm, where they are governed by the intrafirm hierarchical governance structure. From the perspective of transaction cost theories, the decision to bring relationships within the firm is the decision to opt for the intrafirm governance structure over market governance.

Central to the transaction cost approach is the use of hierarchical organizational structure to direct the overall activity of the various components, including employees, brought inside the firm. The hierarchy directs activity using self-enforcing rules and standards. It is this apparatus that replaces market and legal contracts as the organizational mechanism for transacting (Williamson 1996).

Transaction cost theories have played a larger role in labor economics and industrial relations than property rights theories, presumably because of the focus on labor as the central actor. However, incorporating property rights theory resolves important problems that are present when the transaction cost model is considered alone. For example, placing control over access to specialized nonhuman capital at the center of the theory of the firm resolves a critical weakness with the TCE**QQ AU: Please spell out TCE. XQQ**\ theory. Since employees cannot constrain their basic right of job mobility, at least to any meaningful extent, human capital cannot be at the core of the firm, that is, the defining feature of the firm.

What prevents the specifically skilled supplier or employee from holding up the firm even after becoming an intrafirm employee? The property rights answer is that employees follow the orders of the asset owner because the owner can deny continued access to the assets in which the employee has made investments or can grant access to even more valuable assets if the employee is a loyal and productive agent. Because the value of the specific human capital is tied

to particular and transferable nonhuman capital, the nonhuman capital within the firm serves as the glue for the nontransferable specific human capital (Holmstrom 1999; Rajan and Zingales 2001).

Integrating transaction cost and property rights theories also enriches our understanding of the role of hierarchy. In the transaction cost theory, hierarchy is needed to associate the workers with their connected, match-specific tasks. But if specifically trained workers are at the core, why not have them own the firm and appoint a manager to act on their behalf to coordinate them?[12] The answer to this question is provided by the property rights theory. When nonhuman capital is present, the corporate hierarchical structure comes into focus. The value of the corporation is the free cash flow generated by the company's physical and intangible assets. In this model, it is the shareholders who choose the directors and, indirectly, the executive officers who manage the business.[13] Again, employees follow the directions of the hierarchy because of the potential access it can provide to value-enhancing nonhuman capital.

The property rights theory also clears up another ambiguity present in the discussion of asset specificity. The parties who jointly make asset-specific investments do not jointly own the assets. One party buys the asset in the sense of buying the residual rights of control not otherwise ceded to others by contract. The owner can then direct its use. Consequently, in our hypothetical network switch example, when the specialized supplier remains independent, she is the owner of the assets that create the semi-finished goods, but NetSwitch owns the residual rights of control when it purchases her assets. Similarly, employees do not jointly own the assets in which they and the firm make specific investments. Common ownership is inferior to sole ownership because of the efficiency gain when the party that prizes the residual control rights most highly gets to purchase those rights. If employees want to become residual claimants, they have to tie

their compensation to the profits or free cash flow that the firm's nonhuman assets can generate. Employees do not regularly become the suppliers of capital because of risk aversion and the nondiversification that follows when an individual has her human and nonhuman capital assets in the same firm.

Enforcing the Arrangements of Market Players and Employment Relationship Players

In this section, I focus directly on the contracting issue in the three different types of organization structures: the contracting model and the employment relationship in the nonunion firm and the union firm. In so doing, I show that the four industrial factors—match-specific investments, asymmetric information, risk aversion, and transaction costs—largely explain the choice of contracting mechanism. I also relate these models to the theory of the firm and the distinction it draws between extrafirm, or market, transactions and intrafirm transactions. I address the normative question: If the prototypical labor market problem is unequal bargaining power and the unilateral and arbitrary use of power, how are these difficulties controlled in the various theories? Ultimately, what is the enforcement mechanism, and how does it work?

Contract Governance in the Labor Service Market

The labor-contracting model applies to transactions between firms and between workers and firms in the external labor market. It is the market type that has been extensively modeled by economists since it contains few institutional features and no embedded theory of the firm. Since labor market transactions take place outside the firm, the environment is one where transaction costs are low compared with alternative organizational structures. The governance structure is

contract based, and the final authority for resolving disputes is the judicial system rather than a firm's hierarchy.

The firm and the worker (or another firm) set out to resolve one or more of the typical problems involving match investments, risk aversion, asymmetric information, and transaction costs. Their agreement is codified in either an explicit or implicit agreement. As Hart (1995) pointed out, this is essentially the principal–agent relationship, which assumes that the actors can write a complete contract that includes appropriate penalties should anyone deviate from the contract terms.

The contract between the parties, although enforceable in court should that be necessary, is intended to be largely or entirely self-enforcing. This recognizes the central goal of deterring opportunistic behavior while recognizing that enforcement costs can be large. Much of the contracting research, in fact, focuses on the types of arrangements that have strong self-enforcing characteristics, and it is assumed that the parties consciously choose labor contracts that have this property.

If the self-enforcing features of contracts are strong enough to deter all opportunistic behavior, any remaining enforcement issues are trivial. In this world, the role of third-party enforcement is uninteresting, whether it involves private third parties or the courts and the law. Indeed, in much of the economics literature, enforcement issues are rarely explicitly mentioned outside the discussion of the self-enforcing features of the agreement.

However, in a complex world, self-enforcing features rarely fully protect the vulnerable party. In this case, third-party enforcement enters the picture. Third-party enforcement includes the role of private actors, such as other firms and potential workers. Reputational effects are generally seen as the first line of protection should self-enforcement mechanisms fail. Contracts

that are enforced by this method are not perfectly self-enforcing because of their reliance on third-party effects.

Firms or employers are deterred from acting opportunistically because their bad play would eventually be discovered by the labor market and they would suffer reputational losses greater than their potential opportunistic gains. The opportunistically acting employer would eventually be forced to pay higher wages in order to induce new workers to join the firm and thus would have higher labor costs. Also, other firms might be less willing to act as customers or suppliers of the firm because of concerns that its bad treatment of its own employees makes the firm less generally trustworthy.

But what if judicial enforcement is required? At this point the labor-contracting literature goes silent. However, if the contract is complete, as is typically assumed, the legal solution remains trivial. Take the simplest case, when one party decides not to live up to the terms of the agreement. If breach were to occur, whether intentional or inadvertent, the court's role would be to read the contract and enforce the agreed upon penalty. Of course, this situation rarely arises: each party is deterred from breaching because it knows that the court will award damages to the breached-against party.

The judicial enforcement issue becomes more interesting when the contract between the parties is incomplete. This is also typically the boundary where the field of law and economics interfaces with standard economics. There is extensive literature dealing with two primary questions that frequently arise. First, what legal rules should be adopted to deal with partially incomplete contracts resulting from unintentional gaps in the contract? Second, how should the courts respond when contracts are largely incomplete so that there is more gap than content?

The first of these questions has largely been resolved. It is now generally accepted that

the normatively appropriate legal response to partially incomplete contracts is to assist the parties in their profit-maximizing goals while protecting any otherwise unprotected societal interests. This goal is accomplished by adopting the legal rule that is efficient with respect to the problem at hand. The general assessment of contract law is that, as a positive matter, it largely acts in accordance with the normative goal of maximizing the surplus available to the contracting parties. For example, contract law acts as a set of default terms that the parties can adopt by leaving the terms unspecified or that they can overwrite with a term of their choosing (in circumstances where externalities are not present). The result is to provide a standard-form contract that can be adopted so as to minimize the transaction costs of contracting or to allow the parties to adopt whatever terms satisfy their particular profit- or surplus-maximizing goals. Consequently, when the parties omit terms, the courts fill the gaps with the default terms of contract law.

In addition, there is widespread agreement that when a contract is inadvertently incomplete, the court should and, in fact, does fill the gap by adopting the term that the parties themselves would have written had they appreciated the contingency. This is an important conclusion of the legal contract literature: the courts play the role that the labor-contracting literature would want them to play when contracts are inadvertently incomplete. Since the gaps are filled by terms that the parties themselves would have chosen had they known of the gap, the resulting contract still works in a manner that maximizes the surplus available to the parties. Consequently, even though the labor-contracting literature generally ignores enforcement issues, the omission is not a problem because it works entirely in the spirit of the underlying literature.

For these reasons the labor-contracting literature does a good job of describing the theory and practice of the external markets for labor services. More generally, it both describes and

predicts what law and economics scholars call relational contracts, which are pervasive not only in labor service markets but in many commercial markets where the agents are in a continuing, long-term relationship.

Less attention has been paid to the second question: how to address agreements that are primarily incomplete. What should be done when the parties intentionally leave wide gaps in their agreement? The position consistent with this paper is that the courts' response should be to treat the existing contract as complete and to treat issues that arise in the gaps as not covered by the contract.[14]

A simple example illustrates the point. Suppose a builder hires a self-employed laborer to work for him for one week. The laborer meets all the criteria to qualify as self-employed. The two agree to the laborer's hourly wage, the hours to be worked, and the right of the laborer to exit the relationship at will. After a few initial jobs are performed, the builder assigns the laborer to a task that she refuses to perform. The builder responds by withholding all pay.

The laborer grieves in court, claiming that the builder breached their contract by assigning her to the task and asks to be compensated for the time worked. Since the contract is entirely incomplete with respect to work assignment, the court has no guidance as to how to fill the gap. Typically, in such circumstances, the court will treat the contract as complete as to its few terms, ruling that the work assignment is outside of the contract. Since the assignment issue was not contracted for, it cannot be breached. However, the court will enforce the contract as to its hourly pay term, allowing the laborer to recover for hours worked.

More generally, contract law fully protects the interests of parties to the extent that they can do the foundational work of constructing a contract that accomplishes their goals. It is not protective otherwise. Except in the very rare situations where the courts apply defenses such as

unconscionability, duress, or coercion, the mission of the law of contracts is to enforce the terms of contracts. >From the perspective of labor relations, which often sees workers as a vulnerable class in their dealings with corporations, the law of contracts does very little to equilibrate the power of the parties. This is one of the arguments used in the labor relations literature to defend the statutory protection accorded unions by the NLRA.

A final relevant issue regarding the contract law of labor services is that contract case filings and litigations are actually quite rare.[15] This in part supports the ability of self-enforcing mechanisms and private reputational effects to protect parties from opportunistic behavior. The finding is also consistent with the idea that judicial enforcement is important as a deterrence. Critically, however, deterrence appears to be sufficient so that the litigation costs are typically avoided.

Litigation is expensive and wastes resources. In this sense, contract law works best if it works at the deterrence stage. Moreover, when they cannot be resolved by the parties themselves, disputes are often handled by alternative dispute resolution methods, including arbitrators familiar with the parties' relationship. Although evidence about the frequency with which cases are litigated in this venue is sparse, the anecdotal evidence suggests that usage is infrequent.

Norm Governance in the Employment Relationship of the Nonunion Firm

In the employment relationship in the nonunion firm, the parties rely on norms to guide their behavior, with the firm's hierarchical governing structure serving as the ultimate authority for resolving disputes. The problems to be solved by the parties to the nonunion employment relationship are similar to those found in the labor-contracting market, namely, match

investments, asymmetric information, and risk aversion.**QQ AU: No transaction costs without contracts? XQQ*\ In much of the theoretical literature, the models do not distinguish the labor-contracting sector from the nonunion employment relationship. Since the parties are dealing with the same types of problems, the inference is that the substantive features of the arrangements are similar. Moreover, both sectors are assumed to be affected by reputational effects that dissuade firms from dealing opportunistically.

There is, however, a critical difference between the contracting mechanism of the nonunion employment relationship and that of the external labor market contract. Hierarchy rather than contracts is used to develop the terms and conditions of employment and to dictate the norms of behavior that will be rewarded or penalized. Should a dispute arise, the firm dictates its resolution. Even though these terms appear to be like contract terms, the courts make no attempt to enforce what appear to be the parties' agreed-upon norms of behavior.

Take, for example, the paradigmatic problem when a firm discharges a worker and the worker believes that the discharge is without cause. Should the employee take the case to court, as she sometimes does, the courts generally apply the employment-at-will doctrine. Employment at will stands for the principle that an employer can fire an employee for good reason, bad reason, or no reason at all (Ehrenberg 1989). If taken literally, this rule seems to promote opportunism.

The employment-at-will rule also appears to contradict the assertion that the courts protect the interests of the parties. Remember that contract theory predicts and the facts suggest that the courts will fill contract gaps using the term that the parties themselves would have chosen. Moreover, courts will often enforce the parties' own practices even if they are not codified in the contract. Note that although employers seem to have enormous arbitrary

discretion under employment at will, relatively few seem to make use of it. Instead, human resource management preaches that firms should follow the self-enforcing norm of discharging employees only for cause and also claims that firms generally do adopt this loftier standard of behavior. Consequently, although one might expect the courts to adopt the discharge-only-for-cause principle since it is the term generally being practiced, the courts adhere to the employment-at-will rule. The courts' paradoxical passive role in these intrafirm disputes is explained by two factors: how the courts treat agreements marked by largely incomplete contracts and the distinction between governance inside the firm as distinct from governance in external market relationships (Rock and Wachter 1996).

If the parties to the employment relationship were to provide numerous contract terms, the court could play the active role of filling in the gaps or applying the parties' own norms. Errors would be unlikely to be large because of the guidance provided by the numerous existing terms about how the parties wished to handle similar disputes. In the employment relationship, however, the reverse is true. The gaps are large, and there are fewer relevant existing terms. Here the probability of judicial error is great, largely because the courts are too uninformed to make educated guesses. And judicial error raises the costs of the parties' relationship rather than lowering them.

In the presence of large gaps, the helpful court points out that the dispute is not a contract dispute because there is no contract and hence the court lacks jurisdiction. The role of the courts in such situations is analogous to their role when they are asked to resolve a dispute when the contract is largely incomplete. As noted earlier, in this circumstance the court narrowly draws the four corners of the contract to include only those terms that clearly meet the conditions required for contract formation. All other issues are assumed to be outside the contract and hence

not accessible to judicial enforcement. In other words, the courts acknowledge and respect the boundaries of the firm. Inside the firm, hierarchy is used is resolve disputes, not the courts. So interpreted, the employment-at-will doctrine is a statement that the court lacks jurisdiction.

If the employer can exercise arbitrary power based on a hierarchical decision-making process, how is the sanctity of employment agreements protected? Unequal bargaining power would surely appear to prevail in this environment. But can or do firms exercise such arbitrary power? The theory of the firm proves especially useful in providing an answer here. The answer has two elements.

The first element explains why the hierarchical structure is needed. Although hierarchical governance and centralized management can create unequal bargaining power, they also create much of the joint surplus available to society. For centralized management to be effective, it cannot be subject to the ultimate jurisdiction of the courts to resolve private matters. Otherwise, judicial enforcement would undermine the legitimacy of centralized management and open numerous disputes to judicial second-guessing. Once the parties have chosen the organizational form, the courts respect the choice and acknowledge that the firm's boundary is a judicial boundary. But what then controls the acknowledged unequal bargaining power?

The second element is that while the intrafirm governance is hierarchical, it has self-enforcing features that exceed even those available in external market relationships. All of the controls available to external market participants are at work within the firm, with the exception of judicial enforcement. These include the strong self-enforcing features of the norms and the reputational effects exerted by third parties. But what offsets the availability of judicial redress?

The exceptional feature of the employment relationship is that it is an intensively repeat-play market. It is this feature that provides workers with considerable bargaining power. It is

now well known that informal norm governance works best in repeat-play situations where the very high frequency of the interactions provides an aggrieved party the opportunities needed to sanction and thus deter bad play (Ellickson 1991).

The reason is that self-help methods are much stronger in such situations. This is particularly true where, as in the employment relationship, high-frequency interactions involve information asymmetries where the employees' day-to-day activities are imperfectly monitored. In this situation, a firm that engages in bad play by not following the norms can be sanctioned by the employees. In the repeat-play, low-monitoring context, the employees can engage in techniques running from work slowdowns to outright sabotage. In this situation it is the firm that lacks bargaining power, since the remedy—increased monitoring—can be prohibitively expensive for the same reason that contract writing is prohibitively expensive (Rock and Wachter 1996).

Whether the norms of the nonunion sector provide a workable resolution to the problems of the employment relation is an empirical question. At least at this point in time, it appears that the system does work, given the rise in the percentage of workers in this sector. Certainly, important questions have been raised at times about some aspects of this relationship, particularly where asymmetric information makes it difficult for employees to determine whether the employer is actively opportunistic. Examples are the statutory interventions to regulate employment discrimination, pension plans, and occupational safety and health. In these cases, as was true in the earlier cases of regulation of work hours, child labor, and minimum wages, the regulations have carved out specific areas for judicial enforcement while leaving the bulk of the employment relationship unregulated and the firm's hierarchical governance mechanism outside the purview of judicial review (Bennett and Taylor 2002).

If exclusions for government regulation prove to be an acceptable policy response to major norm failures when they emerge, the nonunion sector can benefit from a bifurcated enforcement mechanism that allows for very inexpensive nonunion contracting mechanisms in all but those identifiable areas where management opportunism is most likely to occur.

Statutory and Contract Enforcement of Union Employment Relationships

The major exception to the rule of employment at will for resolving intrafirm employment disputes is the union sector of the U.S. economy. The bargaining mechanism in the union sector operates in a manner that can be partially predicted by the labor-contracting literature. The employer and employees reach an agreement that covers wages and other terms and conditions of the relationship. The provisions of the collective bargaining agreement (CBA) are enforceable under contract law. Like many ongoing commercial contracts, the parties resolve disputes by appealing to third-party arbitrators rather than relying directly on the courts. The arbitrators apply the usual standards of contract law to the dispute at hand by interpreting the evidence and the parties' conduct in the context of the agreement.[16]

With respect to the substantive terms of the employment relationship, there are important similarities and differences between the union sector and the two alternative structures. For example, there is evidence that some of the adjustment patterns followed by union firms are similar to those followed by nonunion firms and by parties in the labor-contracting market. These include upward-sloping age–earning profiles, filling many slots through internal promotions, and wages that are relatively inflexible with respect to downward adjustments during periods of economic slack (Rock and Wachter 1996).

On the other hand, there is general agreement that union workers are paid considerably

more than comparably skilled nonunion workers doing comparable work. Although the union wage premium may have declined over the past decade, it is still material.[17] On a normative basis, the substantive terms of the CBA are thus more favorable to employees than are the outcomes in the nonunion sector and the external market for labor services.

The greatest distinctions among these three forms, however, involve the element of power sharing and the manner in which disputes are resolved. While the labor-contracting sector uses contract law and the nonunion sector uses hierarchy as their enforcement mechanisms, the union sector uses the elaborate superstructure of the NLRA that promotes power sharing. The result is that, whereas the formation of an external labor market contract is entirely voluntary and thus inferentially maximizes joint profits, the collective bargaining contract has numerous mandatory features.

Under the NLRA, once the workers have chosen to be represented by a union, the firm commits an unfair labor practice if it declines to bargain with the union in good faith over the terms and conditions of employment. While the firm retains its hierarchical structure to unilaterally implement changes unrelated to the employment relationship, it is constrained during the bargaining process from unilaterally implementing**QQ AU: OK, i.e., it is limited in making these changes? XQQ**\ changes that do affect the employment relationship. In this sense the firm loses some of its residual control rights, which, as described earlier, allow it to call the shots. Moreover, if the parties cannot reach an agreement, they can use the economic weapons of strikes and lockouts against each other.

Consequently, although collective bargaining contracts may maximize joint profit, there are no legal or market forces making this so (Wachter and Cohen 1987).**QQ AU: Should this be 1988? If not, please add to or correct year in ref. list. XQQ**\ This is a very important

distinction. The claim that contract law is efficient assumes that the parties voluntarily enter into their relationship and bargain for terms in an atmosphere where the threat of coercion or duress is absent and where there are few mandatory terms. The bargaining structure in the union sector is very different from this; hence, it would be surprising for the resulting CBA to have the welfare aspects of the unconstrained commercial contract.

An apparent negative upshot of the power sharing is that numerous disputes arise and make the relationship highly litigious. The parties frequently litigate even minor disputes involving the contract rights of individual workers or of management. In addition, the parties frequently litigate "unfair labor practices," that is, allegations that either management or the union has violated the other's statutory rights. Such statutory-based litigation is, of course, entirely absent in labor service contracts in the external labor market. Moreover, while the CBA itself is enforced under contract law, the rights established by the NLRA are enforceable under the act, with the National Labor Relations Board serving as the court of record. The high costs of the current collective bargaining system are without dispute (Gould 1993; Weiler 1990).

The differences in the cost of the enforcement mechanisms are even greater when we compare union collective bargaining with the nonunion employment relationship. Remember that the difference between the labor-contracting sector and the nonunion employment relationship is that the latter operates without the backdrop of a judicial or third-party enforcement mechanism and all its associated deterrence of opportunistic behavior. This makes the nonunion enforcement mechanism even less costly than the labor-contracting mechanism. Consequently, the cost difference in the enforcement mechanisms between the intrafirm union and nonunion sectors is particularly large.

Costly contracting and enforcement is thus a major problem facing the union sector. Can

enforcement be conducted so as to support the joint interests of the parties, or is it inherently costly and adversarial? The theory of the firm and contract theory suggest a pessimistic assessment. First, the activities performed inside firms are those for which residual control issues are prevalent, and residual control rights involve precisely those events that are not easily predicted before the fact and therefore cannot easily be incorporated into contracts. Second, contract theory predicts that contracts work best when the contracting terms are largely self-enforcing and when the threat of litigation is sufficient to deter remaining possibilities of opportunism. Finally, there are no proposals for making it less adversarial.

So why does the union sector engage in explicit contracting, and why are litigation costs so high? There are two answers: either unions use their power to achieve noncompetitive wages and benefits, or management cannot be trusted and acts opportunistically. To an extent, however, the issues dovetail into one explanation.

The fact that unions can and do achieve noncompetitive benefits for their members is one of the most widely supported economic regularities in labor market research. If unions achieve these gains, the use of explicit contracting takes on a special role: it codifies and makes legally enforceable the noncompetitive returns. This special role is magnified by the fact that relationship is not entirely voluntary. A firm cannot refuse to bargain with a union that has been certified to represent its workforce. In this context, it is likely that the unhappy party will do whatever it can to escape from the unfavorable terms that it views as being forced on it.

The sources of the union pay premium, the wage and benefit clauses, are themselves low-cost items that generate little in the way of litigation. The payment of wages and to an extent the payment of benefits can be observed and verified by the parties. It is the contingencies prompted by the wage and benefit clauses that introduce litigation and increase transaction costs.

Specifically, if the workers are being paid more than comparable nonunion workers would be paid, the firm can be expected to engage in tactics to substitute nonunion for union labor wherever possible or to change work rules or assignments to make the wage and benefit premium less onerous. Given the potential scope of management activities to evade the premium, protecting against these contingencies is a major undertaking. The result is a host of contract clauses dealing with work assignment, discharge, relocation, subcontracting, and work rules in general.[18]

This then is the contracting problem faced by the union sector. If management will search for methods to lower labor costs, then the contract has to be elaborate enough to foreclose as many of these as can be anticipated before they occur. This generates two well-known contracting problems. First, management has asymmetric information about the cost and benefits of proposed initiatives, and its information cannot be easily observed or verified by the union. Second, the union is attempting to foresee future management initiatives that are very difficult to predict with any accuracy. These initiatives, after all, are the very residual steps that management itself believes it cannot contract over because of the difficulty of identifying them beforehand. The upshot is that the union must cast a wide contracting net that constrains management behavior over a substantial range.

Expansive contracting provisions to restrict the use of information by the informed party render the union employment relationship less flexible, since many avenues of adjustment normally open to managers cannot be used. Such restrictions include preventing management from taking initiatives whose primary goal is to reduce the union's noncompetitive advantage. But since there is no way of verifying motive, restrictive contract terms also prevent many changes that would maximize joint profit. Indeed, because of the verification requirement,**\QQ**

AU: OK? Else please clarify "because of verification." XQQ even measures that would increase the profits of both sides may be foreclosed. This, for example, supports the widely cited claim that nonunion automakers can pay union wages and benefits and still have lower unit labor costs. More generally, the result of this inflexibility is to raise unit labor costs without generating direct benefits to either the firm or the union (except for the indirect benefit of protecting the noncompetitive results).

The second explanation for the contract morass is that management cannot be trusted and acts opportunistically. The high litigation costs are caused by the need to constrain opportunism. In the context of the union sector, the commentators who follow this line focus on the claim that managers are anti-union in the sense of attempting to keep or make their plant or company nonunion. Specifically, the pointed attacks on management's actions in the union literature involve the fact that management has never accepted unionization as the appropriate organizational structure for dealing with its employees (Weiler 1990). On this claim itself there is little disagreement. Even promanagement commentators agree that most nonunion firms attempt to remain nonunion and that, in some cases, firms with unionized operations attempt to become nonunion. In addition, there is broad agreement that both sides have historically used practices that have been found to be unfair labor practices under the NLRA. Whether management's position is driven more by the noncompetitive agreements achieved by labor unions or by their unwillingness to share power with unions is a topic that is beyond the scope of this paper.

The two stories—that unions achieve noncompetitive results and that management cannot be trusted—are thus just two elements of the same story. If unions achieve noncompetitive results that place the firm at a competitive disadvantage in its product markets or that reduce

shareholders' returns below some anticipated level, then management can be expected to use whatever mechanisms are available to it to escape from that noncompetitive position. The result is the high litigation costs and adversarial context in which the parties often operate.

Given the disadvantages of unusually high contracting and enforcement costs, what advantages are offered by the union form? There are two situations where collective bargaining offers the preferred organizational structure. The first is when society wants to promote noncompetitive results in the labor market, which was arguably the case at the time of the passage of the NLRA (Wachter 2003; Kaufman 2002).[19] The second case is when management cannot be trusted, even absent a union wage premium. Here *untrustworthy* means that nonunion managers use the hierarchical power inherent in the nonunion form to force noncompetitive terms and conditions of employment on imperfectly mobile workers who have made match investments and who may suffer from informational disadvantages that keep them in the dark as to actual conditions or opportunities.

Conclusion

In this paper I have argued that when economic actors in the labor market choose a method for organizing the provision of labor services, their choice is determined by the parties' specific costs associated with the four problems of match-specific investments, asymmetric information, risk aversion, and transaction costs. The costs relate to both labor and capital. The boundaries of the firm, that is, the choice between using markets or hierarchy, are determined not only by the costs associated with organizing labor services in these two different organizational forms but also by the need to control the costs associated with residual rights of control over physical and intangible assets.

The external labor-contracting market, the nonunion employment relationship, and the collective bargaining structure of the union sector are each likely to be cost efficient under some economic circumstances. Two economic actors are more likely to choose the external contracting model to guide their interests when the transaction costs of contracting are low. This is likely to occur when the parties do not need to interact continuously, there are relatively few match investments, and the critical information relevant to the transaction is available to both parties. When this structure is available and markets are competitive, the parties' interests are well protected. Not only can they rely on the power of self-enforcing contract terms and reputational effects, but should these fall, they can rely on judicial enforcement as well. The problems of inequality of bargaining power and the exercise of arbitrary hierarchical powers are less problematic in this environment because the parties are using competitive markets rather than hierarchy to guide their activity.

Different enforcement issues arise when the activities are brought inside the firm. At least in the nonunion firm, judicial enforcement is absent, so the parties are left to the binding powers of the self-governing norms that they follow, augmented by third-party reputational effects that occur when firms or workers can be labeled as bad players. Why should workers be willing to abjure the protections of market contracting and work inside firms? The theory of the firm provides the answer.

For capital to create value, individual firms must have residual control rights over their key physical and intangible capital. The result is hierarchical governance whereby the executive officers get to call the shots on behalf of the shareholders, who receive the residual income generated by the owned assets. For hierarchical governance to work, the executive officers have to be able to manage the business and affairs of the corporation. Even in corporation law, the

courts give managers great discretion in conducting the business and affairs of the corporation. Although the shareholders get to vote for the directors in their role of residual claimants, they exercise almost no other control rights normally associated with ownership. The reason is to protect the integrity of hierarchical governance and its ability to maximize shareholders' value.

These same concerns devolve into the nonunion employment relationship as well. Hierarchical governance means that the managers get to decide about employment issues with little second-guessing by courts. Workers are willing to relinquish the protections of the external labor market because of the increased wages available to those who are given access to valuable nonhuman capital. In the high-transaction-cost, continuous-interaction setting of the employment relationship, match-specific investments are frequent, and asymmetric information problems are endemic.

For the firm to be successful, it has to decide on the capital stock over which it needs to have residual control. In the nonunion sector, the firm also decides on the norms and specific standards of the employment relationship. Both the firm and its workers have much at stake in making their relationship work. The success of the nonunion sector of the economy suggests that the system must be working, at least tolerably and perhaps much better than that. There is still inequality of bargaining power, which is inherent in the system. But it is at least arguable that the norm-based governance system, policed by the ability of either party to penalize**QQ AU: As intended? XQQ**\ bad play in the transaction-intensive relationship, works reasonably well (Rock and Wachter 2001).

But a hierarchical, norm-based system may not always work. Although the firm has the appropriate incentives to treat workers fairly, the managers may not act this way because of either individual managers' idiosyncratic behavior or firm policy. When this occurs, workers can

seek the heightened protections of the NLRA, whereby the contract formation process itself is protected by contract law and the union certification and collective bargaining processes have distinct statutory protections.

Alternatively, government labor market regulation—narrowly targeted to resolve specific intrafirm problem areas such as employment discrimination, pension regulation, and occupational safety and health—can reduce the need for unions. By carving out problem areas and resolving them while leaving hierarchical governance in place, policy has made the nonunion form much more successful than it otherwise would likely have been (Bennett and Taylor 2002).

The collective bargaining system was originally envisioned to serve the needs of a wide spectrum of workers, not just those dissatisfied with the nonunion sector. Collective bargaining was viewed as a good in itself. However, if collective bargaining is a good, it is an expensive one. The theory of the firm teaches that the hierarchical, norm-based system that eschews contracts and legal enforcement is the low-cost contracting and enforcement mechanism to be used inside firms. The nonunion system has many cost advantages over the union system as long as management opportunism can be adequately controlled by self-enforcing mechanisms combined with reputational effects. If this nonunion system works—in the sense that workers feel adequately protected by it against firm opportunism—then the union alternative is likely to be at a material disadvantage.

Acknowledgments

I am grateful to Bruce Kaufman for many useful suggestions, to Bonnie Clause and Sarah Sisti for research assistance, and to William Draper for library assistance.

Notes

[1] A term of an agreement is self-enforcing if none of the parties to the agreement would find it profitable to use the discretion available to them to redistribute profits to themselves. Agreements that self-enforce thus do not require or benefit from either judicial enforcement or even from private, third-party effects such as reputational effects.

[2] Throughout this article, I refer to the National Labor Relations Act as the regulatory mechanism for protecting unionized workers. Of course, other federal and state laws are also involved in regulating the collective bargaining mechanism (e.g., the Railway Labor Act).

[3] The fact that one party may dictate contract terms to another does not imply that the terms are unfair. The contract terms may entirely or partially reflect competitive market pressures, and the lack of bargaining over the terms may merely reflect the parties' desires to reduce contracting costs (Posner 2003).

[4] See Rock and Wachter (1996). For my purposes, a useful definition of norms is "rules or standards enforced solely by private (i.e., nonstate) actors." The term describes what the parties actually do and is not intended to have any normative content. See, for example, Ellickson (1991).

[5] In addition, the employer and individual employees sometimes write enforceable contracts governing major one-time events such as starting pay, severance pay, or restrictions on competing with the company should the employee quit.

[6] An asset is match specific if an alternative user can redeploy it only with substantial sacrifice of productive value. An asset is general when there exists a ready secondary market so that the asset can be sold at approximately the firm's current use value. Asymmetric information exists when it is relatively more costly for one of the parties to observe or monitor the quantity of

inputs or outputs, the state of technology, or the product market. Risk aversion exists when an individual views risk as bad and is willing to take a lower return or benefit in order to reduce the risk that she faces. Transaction costs refer to the costs of organizing the activities of the inputs and outputs.

[7] Some argue that the efficiency argument is circular. If markets are competitive and workers have choices and can contract freely, then reasonably efficient outcomes will result from the freedom of contract. The problem, according to these commentators, is that workers cannot contract freely, have limited information and limited choices because they are weak in comparison with the firm, and cannot protect themselves. If one assumes this to be true, the collective bargaining apparatus of the union employment relationship would be favored over the nonunion employment relationship (Atleson 1983; Gould 1993; and Weiler 1990).

[8] The economics literature on self-enforcing labor market contracts is extensive. See, for example, Carmichael (1989), Lazear (2000), and Gibbons (1998).

[9] This section draws from Wachter and Wright (1990) and Rock and Wachter (1996, 1999).

[10] This is the efficiency wage problem, whereby increases in wages above competitive levels are actually profit enhancing because they lower monitoring costs by more than the wages increase costs (Ackerlof and Yellon**QQ AU: Yellen in ref. list. Which spelling is correct? XQQ*\ 1986).

[11] The transaction cost theory of the firm was first introduced by Coase (1937) and was developed to its current state by Williamson (1975) and others working in that tradition. For a discussion of its implications for labor market contracting, see Rock and Wachter (1996).

[12] Such a structure exists and is quite prevalent in some sectors. It is the partnership

model in which the partners essentially run the company and all decisions are made by a vote of a majority of the partners unless otherwise specified in the partnership agreement.

[13] The related question is why do the shareholders get to vote for the directors and not the employees, or, alternatively, why not both the shareholders and employees? The answer again turns on residual rights. Only the shareholders are the residual claimants, and thus they alone have the appropriate incentives to maximize the value of the firm.

[14] According to Schwartz (1992), the courts in general do respond in this fashion.

[15] Moreover, this is true for contract law overall. See Galanter (2001).

[16] The range of the collective bargaining agreement is also predicted by contracting theory. Wages, nonwage benefits, hours, and other terms and conditions of employment are covered. Moreover, contracts almost never materially limit the employer's ability to direct the firm through capital expenditure decisions or other nonlabor issues.

[17] There is considerable evidence that unions succeed in achieving a wage and benefit premium, that is, wages and benefits above those paid to comparable workers in the nonunion sector. See, for example, Hirsch and Addison (1986), Kaufman (in press), and Linneman and Wachter (1986). Although there is a claim that unions raise productivity and that the premium is paid out of noncompetitive profits, there is little evidence to support this claim.

[18] It is difficult to empirically verify the effects of contract restrictions on the flexibility of managers to adjust to changing circumstances. This reflects the difficulties of modeling the specific effects of particular contract restrictions. However, there is considerable evidence on the effects of unions on firm profitability. See, for example, Rubak**QQ AU: Ruback in ref. list. Which spelling is correct? XQQ*\ and Zimmerman (1984) and Hirsch (1997).

[19] During the 1930s, the industrial public policy goal was "stabilizing business" or

avoiding "excessive competition," which was understood to mean restricting competition. In this context, unions were viewed as a positive force: a countervailing power to that exercised by corporations (Wachter 2003).

References

Ackerlof, George A., and Janet L. Yellen.**QQ AU: Yellon in text citation. Which spelling is correct? XQQ*\ 1986. *Efficiency Wage Models of the Labor Market.* New York: Cambridge University Press.

Alchian, Armen A., and Harold Demsetz. 1972. "Production, Information Costs, and Economic Organization." *American Economic Review,* Vol. 62, no. 5 (December), pp. 777–95.**QQ AU: Not cited; delete? XQQ*\

Atleson, James B. 1983. *Values and Assumptions in American Labor Law.* Amherst: University of Massachusetts Press.

Bennett, James T., and Jason E. Taylor. 2002. "Labor Unions: Victims of Their Own Political Success." In James T. Bennett and Bruce E. Kaufman, eds., *The Future of Private Sector Unionism in the United States,* New York: Sharpe.

Boal, William M., and Michael Ransom. 1997. "Monopsony in the Labor Market." *Journal of Economic Literature,* Vol. 35, no. 1 (March), pp. 86–112.

Carmichael, H. Lorne. 1989. "Self-Enforcing Contracts, Shirking and Life Cycle Incentives." *Journal of Economic Perspectives,* Vol. 3, no. 4 (Fall), pp. 65–84.

Coase, Ronald. 1937. "The Nature of the Firm." *Economica,* New Series, Vol. 4, no. 16 (November), pp. 386–405.

Ehrenberg, Ronald. 1989. "Workers' Rights: Rethinking Protective Labor Legislation." In Lee

Bawden and Felicity Skidmore, eds., *Rethinking Employment Policy,* Washington, DC: Urban Institute Press.

Ellickson, Robert C. 1991. *Order without Law: How Neighbors Settle Disputes.* Cambridge, MA: Harvard University Press.

Freeman, Richard B., and James L. Medoff. 1984. *What Do Unions Do?* New York: Basic Books.**QQ AU: Not cited; delete? XQQ*\

Galanter, Marc. 2001. "Contract in Court; or Almost Everything You May or May Not Want to Know about Contract Litigation." *Wisconsin Law Review,* Vol. 2001, no. 3, pp. 577–627.

Gibbons, Robert. 1998. "Incentives in Organization." *Journal of Economic Perspectives,* Vol. 12, no. 4 (Fall), pp. 115–32.

Gould, William B., IV. 1993. *Agenda for Reform: The Future of Employment Relationships and the Law.* Cambridge, MA: MIT Press.

Grossman, Sanford J., and Oliver D. Hart. 1986. "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration." *Journal of Political Economy,* Vol. 94, no. 4 (August), pp. 691–719.**QQ AU: Not cited; delete? XQQ*\

Hart, Oliver. 1995. *Firms, Contracts, and Financial Structure.* Oxford: Clarendon Press; New York: Oxford University Press.

Hart, Oliver, and John Moore. 1990. "Property Rights and the Nature of the Firm." *Journal of Political Economy,* Vol. 98, no. 6 (December), pp. 1119–58.

Hirsch, Barry T. 1997. "Unionization and Economic Performance: Evidence on Productivity, Profits, Investment, and Growth." In Fazil Mihlar, ed., *Unions and Right-to-Work Laws: The Global Evidence of Their Impact on Employment,* Vancouver: Fraser Institute, pp. 35–70.

Hirsch, Barry T., and John T. Addison. 1986. *The Economic Analysis of Unions: New Approaches and Evidence.* Boston: Allen and Unwin.

Hirsch, Barry T., and David A. Macpherson. 2002. *Union Membership and Earnings Data Book: Compilations from the Current Population Survey,* 2002 ed. Washington, DC: Bureau of National Affairs. **\QQ AU: Not cited; delete? XQQ**

Holmstrom, Bengt. 1999. "The Firm as a Subeconomy." *Journal of Law, Economics, and Organization,* Vol. 15, no. 1 (April), pp. 74–102.

Holmstrom, Bengt, and Paul Milgrom. 1994. "The Firm as an Incentive System." *American Economic Review,* Vol. 84, no. 4 (September), pp. 972–91.**\QQ AU: Not cited; delete? XQQ**

Kaufman, Bruce E. 2002. "The Future of Private Section Unionism: Did George Barnett Get It Right after All?" In James T. Bennett and Bruce E. Kaufman, eds., *The Future of Private Sector Unionism in the United States,* New York: Sharpe.

———. In press. "What Unions Do: Insights from Economic Theory." *Journal of Labor Research.*

Lazear, Edward P. 2000. "Performance Pay and Productivity." *American Economic Review,* Vol. 90, no. 5 (December), pp. 1346–61.

Linneman, Peter, and Michael L. Wachter. 1986. "Rising Union Premiums and Declining Boundaries among Noncompeting Groups." *American Economic Review,* Vol. 76, no. 2 (May), pp. 103–8.

Manning, Alan. 2003. *Monopsony in Motion: Imperfect Competition in Labor Markets.* Princeton, NJ: Princeton University Press.

Posner, Richard A. 2003. *Economic Analysis of Law,* 6th ed. New York: Aspen.

Rajan, Raghuram G., and Luigi Zingales. 1998. “Power in a Theory of the Firm.” *Quarterly Journal of Economics,* Vol. 113, no. 2 (May), pp. 387–432.**QQ AU: Not cited; delete? XQQ**\

———. 2001. “The Firm as a Dedicated Hierarchy: A Theory of the Origin and Growth of Firms.” *Quarterly Journal of Economics,* Vol. 116, no. 3 (August), pp. 805–51.

Rock, Edward B., and Michael L. Wachter. 1996. “The Enforceability of Norms and the Employment Relationship.” *University of Pennsylvania Law Review,* Vol. 144, no. 5 (May), pp. 1913–52.

———. 1999. “Tailored Claims and Governance: The Fit between Employees and Shareholders.” In Margaret Blair and Mark J. Roe, eds., *Employees and Corporate Governance,* Washington, DC: Brookings Institution, pp. 121–59.

———. 2001. “Islands of Conscious Power: Law, Norms, and the Self-Governing Corporation.” *University of Pennsylvania Law Review,* Vol. 149, no. 6 (June), pp. 1619–700.

Ruback,**QQ AU: Rubak in text citation. Which spelling is correct? XQQ**\ Richard S., and Martin B. Zimmerman. 1984. “Unionization and Profitability: Evidence from the Capital Market.” *Journal of Political Economy,* Vol. 92, no. 6 (December), pp. 1134–57.

Schwartz, Alan. 1992. “Relational Contracts in the Courts: An Analysis of Incomplete Agreements and Judicial Strategies.” *Journal of Legal Studies,* Vol. 21, no. 2 (June), pp. 271–318.

Stiglitz, Joseph E. 1975. “Incentives, Risk, and Information: Towards a Theory of Hierarchy.” *Bell Journal of Economics,* Vol. 6, no. 2 (Autumn), pp. 552–75.**QQ AU: Not cited; delete? XQQ**\

Wachter, Michael L. 2003. “Judging Unions’ Future Using a Historical Perspective: The Public

Policy Choice between Competition and Unionization." *Journal of Labor Research,* Vol. 24, no. (Spring), pp. 339–57.

Wachter, Michael L., and George M. Cohen. 1988.**QQ AU: Is this the ref. cited as Wachter and Cohen 1987? If so, which year is correct? If not, not cited—delete? XQQ*\ "The Law and Economics of Collective Bargaining: An Introduction and Application to the Problems of Subcontracting, Partial Closure, and Relocation." *University of Pennsylvania Law Review,* Vol. 136, no. 5 (May), pp. 1349–417.

Wachter, Michael L., and Randall D. Wright. 1990. "The Economics of Internal Labor Markets." *Industrial Relations,* Vol. 29, no. 2 (Spring), pp. 240–62.

Weiler, Paul C. 1990. *Governing the Workplace: The Future of Labor and Employment Law.* Cambridge, MA: Harvard University Press.

Williamson, Oliver E. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications.* New York: Free Press.

———. 1996. *The Mechanics of Governance.* New York: Oxford University Press.

Williamson, Oliver E., Michael L. Wachter, and Jeffrey E. Harris. 1975. "Understanding the Employment Relation: The Analysis of Idiosyncratic Exchange." *Bell Journal of Economics,* Vol. 6, no. 1 (Spring), pp. 250–78.