Neoclassical Labor Economics: Its Implications for Labor and Employment Law

Michael L. Wachter
University of Pennsylvania Carey Law School

Follow this and additional works at: https://scholarship.law.upenn.edu/faculty_scholarship

Part of the Economic Policy Commons, Economic Theory Commons, Labor and Employment Law Commons, Labor Economics Commons, Law and Economics Commons, Policy Design, Analysis, and Evaluation Commons, and the Political Economy Commons

Repository Citation

This Article is brought to you for free and open access by Penn Carey Law: Legal Scholarship Repository. It has been accepted for inclusion in Faculty Scholarship at Penn Carey Law by an authorized administrator of Penn Carey Law: Legal Scholarship Repository. For more information, please contact PennlawIR@law.upenn.edu.
Research Handbook on the Economics of Labor and Employment Law

Edited by

Cynthia L. Estlund

New York University School of Law, USA

Michael L. Wachter

University of Pennsylvania Law School, USA

RESEARCH HANDBOOKS IN LAW AND ECONOMICS

Edward Elgar
Cheltenham, UK • Northampton, MA, USA
2. Neoclassical labor economics: its implications for labor and employment law

Michael L. Wachter

The application of economics to labor and employment law trails its application in virtually all other areas of business and commercial law. Topics such as contracts, torts, corporations, commercial law, and tax have all readily integrated economic reasoning as an assist to legal reasoning. In almost all these cases, moreover, the economics concept of efficiency has been accepted as one of the goals of the law. This is decidedly not the case in labor law or, for the most part, in employment law. Disagreement over fundamental principles serves as an important explanation for this relative lack of emphasis on economic reasoning.

Most labor law scholars believe deeply that collective bargaining is the preferred framework for the employment relationship. They tend to share the original goal of the National Labor Relations Act (NLRA), which remains on the books today, of equalizing bargaining power by enabling workers to engage in collective bargaining with their employers. The NLRA’s goal of equalizing bargaining power encompassed both procedural and substantive ends: it sought to promote the mechanism of collective bargaining and, in doing so, to raise wages above the market wage that would prevail absent unionization.

Achieving both procedural and substantive ends simultaneously turned out to be the challenge. As unions helped workers gain a wage premium, unionized firms became less competitive and nonunion firms fought harder to stay nonunion. Modern neoclassical economic theory offered a simple explanation based on the most basic principle of economics. The inverse relationship between quantity (i.e., employment) and price (i.e., wage), works in the labor market as in every other market. If unions succeed in raising wages above competitive levels by using the economic weapons available to them in collective bargaining, union employment is likely to suffer. The result is a tradeoff between the number of workers covered by collective bargaining agreements and the size of the union wage premium. As a consequence there is an inconsistency within this goal of the NLRA.

Several attempts have been made to formulate alternative theories that erase this dismal tradeoff between the size of the unionized workforce and the union wage premium in order to resuscitate the twin goals of high union employment and a higher union wage. Most variants are built around the efficiency model. Workers are not like other commodities or inputs in that they can respond to incentives. Perhaps higher wages could call forth greater cooperation and higher productivity. To date, however, this line of research has not been successful in resuscitating the normative appeal for the union sector because its assumptions only hold for the nonunion sector.

Research has moved beyond this threshold issue to other labor market issues, particularly those arising in the internal labor market, or, in other words, the employment
relationship that exists inside the firm. Here workers are less mobile and the firm is less constrained by market forces. Consequently, there is at least potentially greater room for policy intervention that would improve the protection accorded to workers. Even here, however, economic theory which emphasizes the transaction costs of collective bargaining offers more caveats than encouragement. The lack of integration of economics into labor and employment law is thus not surprising at all. If one’s agenda is to promote collective bargaining or provide more procedural rights to nonunion employees, neoclassical economics is not a promising starting point.

The task of this chapter is to analyze the tradeoff between wages and employment and to evaluate other policies and policy prescriptions in the context of a neoclassical model of the labor market. That model is different from the simplified textbook portrayal. Yet it is not new, as its basic building blocks have been available for several decades. For example, portions of this chapter are drawn from two articles that I co-authored (Wachter and Wright, 1990; Rock and Wachter, 1996), citing to the original articles that developed the concepts.

The chapter is organized in the following sections. Section I describes the foundations of the neoclassical model, starting with the textbook model of the union and nonunion labor markets. These are the external labor markets (ELM) where workers seek jobs and employers fill vacancies. This section presents the core issue; namely the tradeoff between union wages and union employment. It then explores the fundamental theorems of welfare economics that allow us to attach welfare implications to alternative market outcomes and investigates the various theoretical attempts to refute the existence of the tradeoff between union wages and employment. Section II develops the model of the internal labor markets (ILM), which represents the employment relationship inside the firm. It also presents the connections between the ILM and the ELM and presents the four central economic factors in the ILM; namely job-specific training, risk aversion, information asymmetries and transaction costs. Sections III and IV discuss how the parties deal with the four factors of the ILM either through union contracting (Section III) or nonunion contracting or norms (Section IV). The last section, Section V, concludes the chapter and discusses policy implications.

I. FOUNDATIONS OF THE NEOCLASSICAL MODEL

A. Textbook Model of a Competitive Labor Market

I start with the simple competitive equilibrium textbook model of the labor market (see Figure 2.1). Although simple, it is all that is needed to engage much of the debate. This is the external labor market (ELM), where firms hire workers into new jobs and unemployed workers search for jobs after an existing job turnover or an entry into the labor market. The key prediction is that wages are determined by market forces, labor supply (denoted by S), and labor demand (denoted by D). Market-equilibrating forces lead the labor market toward the equilibrium solution, where S and D intersect, with employment at $E_C$ and the wage at $W_C$. At this point wages cannot be higher without employment being lower. The equilibrium point is thus efficient.\(^4\)

Suppose, however, that the labor market were out of equilibrium, with a low wage of
$W_L$ (where the subscript L denotes the low wage). At this low wage level, demand for labor ($D_L$) would be far in excess of the supply of labor ($S_L$). Since labor markets are competitive, employers would offer higher wages to compete for workers in order to fill the high demand for labor. Wages would rise toward $W_C$, and the labor supply would increase as more individuals joined the labor force seeking jobs at the higher wage levels. At the same time, labor demand would begin to decline because the higher pay of the work force would mean that workers were less profitable. Thus the equilibrating process would continue until the wage had increased from $W_L$ to $W_C$ and demand had declined from $D_L$ to the competitive equilibrium level of employment (denoted by $E_C$).

Alternatively, suppose the labor market were out of equilibrium with a high wage of $W_H$. At this point labor demand would be at $D_H$, which would be much lower than the supply $S_H$. Market-equilibrating forces would again come into operation. Now workers would compete for the jobs at the high wage rate. As a consequence of the labor competition, the wage rate would begin to decline. As the wage declined, employment would begin to pick up, reflecting the lower wage rate. The process would continue until the wage had declined to $W_C$ and employment had increased to $E_C$.

The above discussion of labor supply and demand presents a partial equilibrium view. Labor supply and demand functions fit into a structural model of the overall economy. In the structural model, wage effects ripple through other markets and in turn feed back to influence the location of the labor supply and demand curve.

The labor demand curve discussed above is referred to as a derived demand because it is derived from the demand for the firm’s product. The location of the demand curve depends on the productivity and skill of workers in producing the product or service that the employer is selling in the market. The greater the productivity of workers, the greater is the equilibrium wage. Demand curves slope downward because of substitution effects.
and diminishing returns as consumers switch away from expensive products and employers substitute inexpensive factors of production for expensive factors. At this level of abstraction the diagram refers to the overall labor market; however it can be generalized to any number of different skill levels and geographical regions. This discussion assumes the existence of competitive labor markets.

The supply curve's location depends on the work/leisure choice made by individuals. Supply curves are generally viewed as being upward sloping: that is, more individuals enter the labor market in search of jobs when the wage is high. However, whereas the demand curve is always downward sloping, the supply curve need not always be upward sloping. For the purposes of this chapter, I assume the labor supply is upward sloping, but nothing of importance in the chapter turns on this assumption.

It is the existence of an inverse relationship between wages and employment that is at the heart of the debate. There are some changes in the economy, such as an expansionary monetary and fiscal policy in the midst of a recession, that can allow employment to increase even as wages are increased. But changes of this type are not contested. The debatable issue is a very specific one, namely whether an exogenous increase in union wages above the market-clearing wage, as a result of collective bargaining, will have a negative effect on union employment. Deconstructing the wage/employment tradeoff thus becomes critical for traditional labor scholars who want to undo the claim that higher union wages are inversely related to union employment levels.³

In this chapter I assume, as do most other economists, that external labor markets in the nonunion sector are generally highly competitive. I do not assume that product markets are necessarily competitive nor do I assume that internal labor markets are perfectly competitive. Although the textbook model above assumes perfect competition, perfection is not needed for the efficiency and welfare conclusions to hold. Labor markets that are highly competitive, but not perfectly so, can have frictions and information gaps at any point in time, and can be in disequilibrium in the short run. None of the policy conclusions is changed by acknowledging these short-run market imperfections.

The model I use here, the neoclassical theory, is the standard model used in economic analysis. It can be criticized for ignoring many institutional features of the labor market, but its goal is testable predictions. And that it does with excellence. Wages do generally rise at a faster rate during expansions and decrease or rise at a slower rate during recessions. More skilled workers will generally have higher wages than less skilled workers. Growing regions will tend to have faster wage growth than declining regions. There are discrepancies, of course, because no theory is perfect, but the overall predictive picture is strong. Moreover – and this is key – there is no alternative theory which offers a similar range of predictions. Certainly there are many variants of the neoclassical model, but all share the assumption that prices and quantities are determined in markets by pressures of supply and demand. Moreover, whenever markets are out of equilibrium, there are equilibrating forces that push the market toward equilibrium. The equilibration is not instantaneous and it can even take years to recalibrate, as was the case in the early 20th century, where workers from the southern regions of the United States moved north in search of newly created manufacturing jobs.

Another key component of the neoclassical model consists of the fundamental general equilibrium theorems. The first fundamental theorem, stated in its most simple form,
says that competitive markets are efficient because they maximize the total income that the economy can produce. The second fundamental theorem, again in its simplest form, is that there are many different income-maximizing states, each corresponding to a different distribution of income across individuals. The government, through tax and expenditure policies, is given the job of choosing among the alternative income-maximizing states to determine the welfare-maximizing income distribution. The policy implication drawn from these theorems is that policymakers should first attempt to make markets as competitive as possible, and then should seek to improve income distribution to reach the optimal distribution of income or welfare.

A number of interesting, albeit highly stylized, policy implications can be drawn from the above description of the labor market. First, there is no need for government policy (or for unions) to ensure that wages rise over time with worker productivity. Competitive markets will lead to that result. Instead, if unions raise wages above competitive levels, the economy will be less productive. Second, when changes in the income distribution are deemed desirable, they should generally be achieved through tax and government expenditure programs rather than by dictating noncompetitive outcomes in markets through government regulation. Industrial policy should be confined to making markets as competitive as possible. Thus policies that improve the efficiency of particular labor markets are generally favored in neoclassical economics.

Although no one believes that markets are perfectly efficient, the neoclassical textbook model assumes that markets are perfectly competitive as a simplifying assumption (Ehrenberg and Smith, 2011) and as an ideal against which policy improvements can be judged. Notably, over the last several decades neoclassical models have incorporated job-search frictions, job-specific training, information asymmetries, and other market realities that might allow for policy improvements.

The neoclassical model has also been extended to account for monopsony, in which one firm can exercise labor market power. Unlike in the competitive labor market, in which firms are “wage takers” when they hire workers, in the monopsony model the firm is a wage setter. It needs to raise its wage if it wants to increase hiring and can lower its wage when it is reducing employment. I will not illustrate this model or discuss it in any detail, however, because there is little evidence of material monopsony power in U.S. labor markets so that whatever monopsony power exists is unlikely to change any of the conclusions reached in this chapter. If monopsony were dominant in the labor market, higher union wages could lead to higher employment.

Labor markets tend to be more competitive than product markets. The reason is that employers, regardless of their product market, still hire workers from the same external labor market. Prospective workers have no firm-specific training and their skill level depends mainly on their level of education and general training, whether they are high school graduates, college graduates, or professional school graduates. If employers are hiring locally, then each would hire blue-collar workers from the same high schools and entry-level engineers or managers from the same colleges. In other words, even if labor markets were entirely local, they would be more competitive than the local product markets. However, labor markets are not local. While workers incur costs in moving from one locality to another in search of jobs, in today’s society labor markets are more national than local.
says that competitive markets are efficient because they maximize the total income that the economy can produce. The second fundamental theorem, again in its simplest form, is that there are many different income-maximizing states, each corresponding to a different distribution of income across individuals. The government, through tax and expenditure policies, is given the job of choosing among the alternative income-maximizing states to determine the welfare-maximizing income distribution. The policy implication drawn from these theorems is that policymakers should first attempt to make markets as competitive as possible, and then should seek to improve income distribution to reach the optimal distribution of income or welfare.

A number of interesting, albeit highly stylized, policy implications can be drawn from the above description of the labor market. First, there is no need for government policy (or for unions) to ensure that wages rise over time with worker productivity. Competitive markets will lead to that result. Instead, if unions raise wages above competitive levels, the economy will be less productive. Second, when changes in the income distribution are deemed desirable, they should generally be achieved through tax and government expenditure programs rather than by dictating noncompetitive outcomes in markets through government regulation. Industrial policy should be confined to making markets as competitive as possible. Thus policies that improve the efficiency of particular labor markets are generally favored in neoclassical economics.

Although no one believes that markets are perfectly efficient, the neoclassical textbook model assumes that markets are perfectly competitive as a simplifying assumption (Ehrenberg and Smith, 2011) and as an ideal against which policy improvements can be judged. Notably, over the last several decades neoclassical models have incorporated job-search frictions, job-specific training, information asymmetries, and other market realities that might allow for policy improvements.

The neoclassical model has also been extended to account for monopsony, in which one firm can exercise labor market power. Unlike in the competitive labor market, in which firms are “wage takers” when they hire workers, in the monopsony model the firm is a wage setter. It needs to raise its wage if it wants to increase hiring and can lower its wage when it is reducing employment. I will not illustrate this model or discuss it in any detail, however, because there is little evidence of material monopsony power in U.S. labor markets so that whatever monopsony power exists is unlikely to change any of the conclusions reached in this chapter. If monopsony were dominant in the labor market, higher union wages could lead to higher employment.

Labor markets tend to be more competitive than product markets. The reason is that employers, regardless of their product market, still hire workers from the same external labor market. Prospective workers have no firm-specific training and their skill level depends mainly on their level of education and general training, whether they are high school graduates, college graduates, or professional school graduates. If employers are hiring locally, then each would hire blue-collar workers from the same high schools and entry-level engineers or managers from the same colleges. In other words, even if labor markets were entirely local, they would be more competitive than the local product markets. However, labor markets are not local. While workers incur costs in moving from one locality to another in search of jobs, in today’s society labor markets are more national than local.
B. Textbook Model of a Union Labor Market

The primary goal of labor unions is to “take wages out of competition.” Labor is not just another commodity and, in this line of reasoning, workers should not be treated as a commodity and have their pay determined by market forces. A key question for labor economics is what consequences will follow union efforts to raise wages above what the market would otherwise dictate.

The textbook union labor market uses a simple variant of the supply and demand model used in Figure 2.1. Figure 2.2 starts by duplicating the wage and employment levels that exist in a competitive market. As noted above, the equilibrating forces of competitive markets push the labor market toward the intersection of the S and D curves. The union market works very differently because the equilibrating market forces are replaced by the collective bargaining mechanism. The wage set through collective bargaining, denoted $W_u$, is higher than the market-clearing, or competitive, wage, $W_c$. With its legally protected strike weapon, the union can extract a wage premium from employers in return for not striking.

As a result of the higher union wage, firms set labor demand at $D_u$. The societal cost of the union wage being higher than the competitive wage is a lower level of employment ($E_u$ rather than $E_c$). Furthermore, at the higher union wage, not only does the number of jobs decrease, but the number of workers seeking those jobs increases. The result is unemployment, measured by the horizontal distance at $W_u$ between the supply of workers to the union sector, $S_u$, compared to labor demand, $D_u$. Without the equilibrating force of competitive markets the resulting social loss of reduced output and higher unemployment can persist in the union labor market.$^{10}$

The union labor market is sometimes referred to as a cartelized market in that the union as the agent of the workers allows labor (workers or suppliers of labor) to act in

![Figure 2.2 The union market](image-url)
concert so that the wage is set above competitive levels (Posner, 1984; Wachter, Chapter 15 in this volume). From an economics vantage point, there is nothing controversial about calling the unionized labor market a cartelized market. The use of the term simply fits the economic definition of a cartel. Since cartels in product markets, however, are generally viewed negatively, the term cartel may have acquired negative connotations. Its use here, however, is entirely descriptive. Cartels seek to take prices, or in this case wages, out of competition, and that is perhaps the best single sentence stating unions’ broad economic goal.\footnote{11}

The problem for the union sector is to maintain the cartel and prevent the equilibrating market forces from operating. This was a more likely event under the original Wagner Act, which favored unionization.\footnote{12} After the Taft-Hartley Act was passed it became easier for nonunion firms to form, remain nonunion, and thus compete, using their lower labor cost structure, with the unionized firms. The result was the growth of nonunion sectors in virtually all United States product markets. The emergence of international competition has only worsened an already bad situation for unions (Wachter, Chapter 15 in this volume).\footnote{13}

In terms of Figure 2.2, this shift of work from the union sector to the nonunion sector means that the labor demand curve shifts gradually to the left over time, resulting in steadily declining union employment. The union firm can either lower its product price to the nonunion price (to the detriment of profits) or try to maintain a price premium to cover the union wage. In either case, the union firm is likely to cut employment as it loses market share. Where the competition is greatest, the unionized firm is unlikely to earn a competitive return on its capital so that it is unlikely to continue to invest in updating its capital equipment in its high cost plants. As its capital becomes obsolete, the plants’ production becomes uncompetitive so that output falls, causing a continuing erosion in union employment (Linneman, Wachter, and Carter, 1990; Hirsch, Chapter 4 in this volume).

Union wage setting in the labor market does not repeal product market pressures: wages are not actually taken out of competition in competitive product markets. The extent of the union wage premium, which is the percentage difference between $W_U$ and $W_C$, will affect the fortunes of the firm, and if there are nonunion firms in the same product market, the union firm will be disadvantaged. That is invariably the case today: there are essentially no product markets that lack a nonunion sector.

One of the explicit goals of the Wagner Act was to equalize bargaining power. As Wachter (Chapter 15 in this volume) points out, the substantive element of that goal was to push the collectively bargained wage higher than the market wage. If the Wagner Act had also succeeded in the procedural element of that goal of spreading unionization throughout the economy, then the general equilibrium result of unionization would be not only to raise wages throughout the economy, but also to generate unemployment. The results would be akin to those shown in Figure 2.2.

The framers of the Wagner Act did not anticipate that the goal of raising wages above market levels and increasing union employment at the same time would prove to be inconsistent. The neoclassical economic model was not developed enough during the 1930s for it to be a useful policy tool. The three components of the theory not available to commentators in the 1930s were, first, a partial equilibrium theory of the labor market; second, a general equilibrium theory that pointed out the efficiency gains achieved by competitive outcomes; and third, a business cycle model.
The partial equilibrium theory of the labor market, which posits that labor demand moves inversely with the wage rate, was first sketched out by Marshall (1890)\textsuperscript{14} at the turn of the last century, but its complete statement awaited the work of Sir John Hicks. Although Hicks’ book was published in 1932,\textsuperscript{15} Hicks was a 28 year old, hardly known budding star at the time. Consequently, the theory did not make its way out of university communities until the 1940s and it is not clear what Congress believed to be the appropriate microeconomic theory to use in evaluating the effect of an increase in wages resulting from collective bargaining.

Since the competitive wage was not a well-understood concept, it is fair to believe that the theory adopted by Congress may have been more one of Commons than of Hicks. Commons, one of the great labor economists of his time, thought that unemployment would be a constant problem and that, absent either government intervention or collective bargaining, the unemployed would bid down wages until they were at or immediately below a living wage (Commons and Andrews, 1927).

Until the advent of the modern neoclassical theory of labor markets, there was little support for the notion that society could be wealthy enough to generate a derived demand for labor that, in intersection with labor supply, would generate and maintain a socially acceptable wage. Yes, real wages had risen substantially over time in the 1800s, and took off with the industrial revolution. But a series of cyclical crises culminating in the Great Depression suggested an unpleasant dynamic story that capitalism was prone to excess competition which would inevitably generate depressions and “reserve armies” of unemployed workers. If the Great Depression had been the story of the future, then Commons rather than Hicks might still be taught in Econ 1.

The economic inefficiency of cartels was also largely unknown at the drafting of the NLRA. Showing that unions may cause unemployment in one market does not necessarily prove that the economy is worse off: Adam Smith (1776) conjectured that that was true in 1776, and Leon Walras (1874) developed the initial theory for it (if you could read French); but it was not until the 1950s that general equilibrium theory, and the proposition that competition produced efficient results, took hold through the works of Arrow and Debreu (1954), Debreu (1959), and McKenzie (1981), among others.

A piece of the puzzle was still missing, and that was the macroeconomic analysis of cyclical unemployment. This was provided by Keynes (1936) in the midst of the Great Depression. What Keynes showed and post-Keynesians have further developed is that a combination of monetary and fiscal policies could dampen the severity of business cycles. Although the Real Business Cycle model has raised questions about Keynesian theory and even the efficacy of monetary and fiscal policy, the state of the debate between New Keynesian versus Real Business Cycle theories leaves policymakers with far greater countercyclical tools than were available during the Great Depression (Mankiw, 1989; Abel, Bernanke, and Croushore, 2011).

The current macroeconomic debate has both the Keynesians and the Real Business Cycle theorists agreeing upon one critical feature of modern business cycles, namely, wage rigidity. That nominal wages are rigid in the union sector is not surprising. They are a key feature of the collective bargaining agreement. But wages also do not decline during recessions in the nonunion sector, at least since the Great Depression (Dobrescu, 2012).\textsuperscript{16}

The finding of wage rigidity during recessions in the nonunion sector puts to rest
concerns raised by the old economic greats like Malthus and Commons. Even during a recession and a period of high unemployment, market wage rates do not fall. One explanation for this wage rigidity is that firms are less risk averse than workers and are thus willing to offer insurance against income instability due to the business cycle. The other argument, and the one advanced in this chapter (see Section IV), stresses that wage cutting is not done even by nonunion firms because wage reductions in the face of declining demand is not a self-enforcing adjustment to declining product demand.

The upshot of this brief historical analysis is that, from the perspective of modern labor economics, the theory available to policymakers at the time of the passage of the Wagner Act in 1935 was flawed or not yet developed. It lacked the microeconomic foundations of labor markets being developed by Hicks and others; it lacked the fundamental welfare theorems of general equilibrium theory; and it lacked a modern business cycle theory. The Act’s fear that absent unions there was nothing that could prevent recurring depressions was out of date by the late 1930s, and there is nothing in modern business cycle theory to suggest that unions help to prevent recurring business cycles.¹⁷

Indeed prevailing economic opinion had already moved away from this position by the late 1940s, when the Taft-Hartley Act was passed, moving labor policy from favoring unionization to neutrality.¹⁸ Unions might sometimes be part of the problem rather than being part of the solution.

Still, many labor law scholars ignore the economic effects of unions even though the neoclassical model is so widely accepted in other areas of legal scholarship. What accounts for the widespread resistance to this conclusion in labor markets? One answer is the hope that a competing theory, the efficiency wage theory, might be used as an economic argument for favoring unions.

The heart of the efficiency wage model is the idea that if you pay workers a higher wage they will work harder and be more productive. It is, after all, a plausible outcome. Workers are not like any other commodity because they can adjust their behavior to incentives. As a consequence, instead of worker productivity explaining the wage, the wage determines worker productivity by providing incentives to work more productively.

C. Efficiency Wage Model

The efficiency model can be viewed loosely as repealing the harsh predictions of the inverse relationship between the union wage and union employment. In the labor law literature, this has always been a favored line of thinking. The original example of the efficiency wage model was Henry Ford’s decision to pay wages far above competitive levels in order to secure a more reliable and productive labor force.

The seminal efficiency wage model centers on the idea of a “gift-exchange” (Akerlof, 1984). The gift is a wage above the market-clearing wage. The basic theory is that the worker understands that she is receiving a gift in the form of a wage above the market-clearing wage. What is often neglected when scholars apply this model to the union sector is the importance of identifying the gift-giver. If the gift is coming from the firm, workers reciprocate by becoming more loyal to the firm, putting out more work effort and thus
becoming more productive. Workers who are being treated well by their employer may, for example, respond to a busier than usual schedule by putting out more effort. The converse is also true: workers who think they are being badly treated may respond by putting forward less effort.

To see how the theory is likely to operate it is useful to describe the paradigmatic example of the efficiency wage: that is, Henry Ford's $5 a day wage. There are various stories of the famous episode. If one is to take Henry Ford at his word, which was cited in the famous case of *Dodge v. Ford Motor Company*, he indicated: ""My ambition . . . is to employ still more men; to spread the benefits of this industrial system to the greatest possible number, to help them build up their lives and their homes. To do this, we are putting the greatest share of our profits back into the business."" In this story, Henry Ford is just a good, very rich capitalist who wants to share part of the good fortunes of the corporation with the workers rather than just with the shareholders.

The historically more grounded story is more complicated and less encouraging. Henry Ford was one of the first of the great entrepreneurs of the industrial age. He was rapidly expanding the business to meet a burgeoning demand for the low-cost cars he was producing. This required a concomitant rapid increase in employment. The problem was that, in 1920, he was primarily hiring workers off the farm who were unaccustomed to the discipline and on-time performance of manufacturing jobs. Workers were often not on time or strayed from the assembly line. Others just left to go back to the farm. The $5 a day wage, which was roughly double the competitive wage at that time, was an attempt to reduce turnover and make existing workers more productive.

The $5 a day wage required no elaborate monitoring, for if the boss did not think workers were meeting the standard, they could be freely fired under the then and now existing employment-at-will rule. In other words, the efficiency wage model requires that workers increase work effort as a response to the higher wage. The efficiency wage is thus an incentive story. Ford paid a higher wage, but the cost of that higher wage was offset by a more productive workforce. Besides requiring no monitoring, Ford could set the wage at whatever level he thought maximized profits. The goal of the $5 wage was not to lose money; it was to make greater profits.

How does this story play out in a unionized firm? First, to stay faithful to history, I note that Ford was vehemently antiunion. The violence that erupted when the United Auto Workers organized the Ford Motor Company is well known in labor history. Henry Ford hired a security force that engaged in battles with union organizers. Obviously, the creator in real life of the efficiency wage concept did not take kindly to having his facilities organized. And that fact is important.

The efficiency wage story simply does not fit in a unionized company. First, the idea that workers will work harder because the job pays better than their next best job prospect depends on low-cost monitoring. The efficiency wage is itself an implicit contract: higher wages for higher effort. For this deal to work, the company must be able to discharge a shirking worker quickly and at low cost. The efficiency wage pays for itself by encouraging workers, who find the above-market wage sufficiently compelling, to work hard to keep the job, whether this means on-time arrival, few absences, and no shirking on the job. That mechanism will be disrupted if the worker found to be shirking can mount and sustain a claim of unfair dismissal. From the worker's standpoint, the claim makes sense: for some workers caught shirking, it may
have been a one-time lapse or the supervisor may have made a mistake in rendering the discharge. The presence of a union with a grievance-arbitration process in the collective bargaining agreement protects workers from these mistakes; but at the same time it forces the employer either to engage in costly monitoring or to tolerate a certain amount of “shirking,” thus reducing the benefits to the employer that are promised by the efficiency wage theory.

Second, “who gets credit” for the gift? In collective bargaining, the gift-giver is the union. After all, it is the union that fought for the higher wage. The press rings with stories in which union leaders claim that their gains were “won” at the bargaining table and had to be wrung out of the boss. The stories are quite plausible. However, for the gift-exchange version of the efficiency wage theory to make economic sense, the gift has to be seen to come from the employer, for it is the employer’s willingness to pay the above-market wage that encourages workers to be more loyal and grateful.

The efficiency story is alive and well, however, in many nonunion sectors of the economy. None of the economic models of the efficiency theory sees the efficiency wage as the product of collective bargaining where the parties can use their economic weapons to achieve the best contract for their side. To generate greater efficiency to cover the cost of the higher wage, the efficiency wage is calculated based on offsetting the higher cost with greater effort. It is entirely an incentive device. Nonunion firms may pay higher than market wages, both to reduce their turnover and monitoring costs but also to avoid unionization.

Elements of the efficiency wage model can be found in Freeman and Medoff in their book *What Do Unions Do?* (1984). They argued that higher union wages might encourage higher productivity among workers by providing workers with a voice that could affect the workplace environment. As a consequence of voice, workers might gain improvements in job seniority and security that would encourage workers to pass on skills to junior workers and to work more productively in general. Freeman and Medoff provided little evidence that unions could raise productivity high enough to offset the union wage premium. The only empirical support for the effect was provided by Brown and Medoff (1978). Indeed Freeman and Medoff acknowledged that any positive effects of unions on productivity would probably not be sufficient to undo the detrimental effect of the union pay premium. Consequently, they argued that the gains from the good face of unions could only be achieved by policy intervention such as amending labor law.

The vast literature on the subject of the efficiency effects of unions, much of it stimulated by the Freeman and Medoff book, showed a negative rather than a positive effect on productivity (although not statistically significant). This is summarized in Chapter 4 of this volume by Hirsch.

In sum, unions raise pay, and consequently union firms are at a cost disadvantage when competing with nonunion firms. There is no discernible offsetting increase in worker productivity. The great loss of union employment has been more than offset by the growth in domestic nonunion employment. This is not a story about low wages in emerging countries. Instead it is about the existence of a sizeable nonunion sector in every American industry that is able to compete successfully with its union competitors on the basis of lower labor costs. (See Hirsch, Chapter 4 in this volume.)
II. UNDERSTANDING THE INTERNAL LABOR MARKET

Thus far, I have stressed that the external labor market is generally efficient. The external labor market is where individuals search for jobs when they either enter the labor force or leave their current employment, and where firms look for new employees. It is “external” in the sense that it is outside of the firm. What makes the external labor market competitive is that firms in numerous different industries are all looking for entry workers with generalized skill sets. A high school graduate could become a semi-skilled worker at Caterpillar, or at JP Morgan, or at a small firm. A college graduate could become a management trainee at General Electric or a web designer for a small firm. The workers’ general skills may be more or less specialized and better suited for some industries than others; but they are not firm specific. Hence these workers can work in a range of industries or for any firm within the industry. The same story can be told of the MBA graduate who can be hired into management at Caterpillar or at JP Morgan, or at a small firm.

Once a worker is hired into a firm, she leaves the external labor market and enters the internal labor market (ILM). The internal labor market is the relationship between the firm and its employees. The ILM consists of a set of explicit or implicit rules governing wages, promotion opportunities, grievance procedures and other terms and conditions of employment. In large firms, the ILM can be highly formalized; in small firms, it may barely exist.

A. Basic Principles of Internal Labor Markets

The ILM comes in two broad types: unionized and nonunionized. In the union ILM, the rules governing the workplace are largely a matter of express contract; they are written in the form of a collective bargaining agreement and are enforceable before an arbitrator, a court or the NLRB. In the nonunion ILM, the agreements are not contractual; that is, most of their terms cannot be enforced by a third-party arbitrator or judge. Rather, the nonunion ILM relationship is generally one of employment-at-will, and most of its terms are subject to modification at will.

Both types of ILMs are generally governed by similar broad principles, whether legally enforceable or not. Most hiring takes place at entry-level jobs which connect the ILM to the ELM. Based on performance or job tenure, workers can then be promoted into higher level jobs. Earnings typically rise with job tenure. In a product market decline, adjustments are typically made by reducing hours or laying off workers rather than by reducing wages; and resulting layoffs or discharges generally are inversely related to seniority. The difference between the union and nonunion ILMs is that these characteristics are contractually mandatory in the union ILM, whereas they are non-mandatory, frequently followed norms in the nonunion ILM.

What is described here is a somewhat idealized version of an ILM – or perhaps a historically contingent ILM, such as that which existed at “blue chip” companies like IBM and AT&T until the 1980s. Many observers have described changes in the nature of ILMs since the 1980s, characterizing them as an erosion or even an outright collapse of the long-term employment relationship and mutual expectations that are central to the idealized ILM (Stone, 2004). But none has convincingly shown that ILMs function anything like external “spot” markets for labor. So we will stick with the idealized ILM
for analytic purposes, while recognizing that this overstates the difference on the ground between ILMs and ELMs.

From a policy perspective there is a huge difference between external and internal labor markets. Whereas ELMs are generally efficient because market forces discipline the parties ex ante, there is no reason to assume that internal labor markets are efficient. Most importantly, whereas workers are mobile in the ELMs, they are less mobile in the ILMs, in part because of the firm-specific training they undertake. Consequently, market discipline is muted in ILMs.

The analysis of ILMs began during the 1950s when Kerr (1954) and Dunlop (1958) described the institutional realities of internal markets. This pioneering work did not attempt to integrate the ILM as an institution into the neoclassical theory of labor markets. But that integration has come about since 1970 as the efficiency aspects of the ILM were developed by Doeringer and Piore (1971); Williamson, Wachter and Harris (1975); Okun (1981); and Freeman and Medoff (1984). The ILM is thus no newcomer to neoclassical labor economics. It has been an integral part of labor economics for over 30 years.

B. Relationship between Internal and External Labor Markets

An important feature for both the union and the nonunion ILM is the manner in which they connect with the ELM. As noted, the ELM is differentiated in worker skills largely by the amount of education workers have received, and workers are hired into the ILM at entry-level jobs. This is true for all levels of education, though the ports of entry differ. Upon hiring, workers receive job-specific or firm-specific training, which enables them to move up the promotion ladder. Employees tend to receive higher wages based on seniority and the amount of specific training they have undertaken. Hourly pay tends to fluctuate little over the business cycle. Instead, the stage of the business cycle is manifested in hours worked.

Whereas the market discipline of the ELM controls entry wages for entry-level jobs, it has only indirect control over the great variety of jobs filled by internal promotions in the ILMs. Over the innumerable job titles that exist in the United States economy, very few are determined solely by the ELM. Yet the ELM does discipline all the ILM jobs because employees always have the right to quit and are likely to do so if job opportunities are more favorable at other firms. As expected, job turnover is higher for workers with little job tenure since starting over with a new employer typically means starting at that firm’s entry-level position and entry pay; that usually represents a significant loss for more senior workers. Since young entry-level workers are mobile, they are likely to earn a market-based wage, while the older, less mobile workers are vulnerable because market discipline imposes little direct constraint on the employer.

Nevertheless, the external markets matter a great deal in controlling the pay and working conditions in internal labor markets. Few workers need to be mobile. As is true in economics generally, it is the worker on the margin of stay versus quit who wields influence. The worker on the margin will have the best information on alternative job opportunities and is likely to share it with other workers. The employer whose ILM is inconsistent with the relevant ELMs, in terms of wages, benefits and other terms and conditions of employment, will not be able to hire and retain a qualified workforce.
With the exception of wages, benefits, and other terms of conditions of employment, employers retain discretion over other matters whether the firm is unionized or not. Employers retain broad discretion, for example, to choose the firm’s technology and to construct job descriptions in accordance with that technology. Virtually all product market decisions are decided unilaterally by the employer. Relations with the firm’s other stakeholders are also typically determined by the employer without any input from employees.

C. Four Central Economic Factors that Affect the ILM

The ILM has four foundational factors that distinguish it from ELMs and that create potential inefficiencies: firm-specific training, risk aversion, asymmetric information and transaction costs.  

1. Firm-specific training

Firm-specific training is the core rationale for long-term attachments and thus for the ILMs (Wachter and Wright, 1990). Without firm-specific training, there would be no advantage to the firm in retaining a current employee as compared to hiring a new employee each day. Firm-specific training makes workers more valuable to their current employer than to other employers. The result is a wedge or surplus between their value to their current employer versus their value in the ELM. When this surplus is shared between the firm and the employee, the firm has an incentive to retain workers with job-specific training and workers have a similar incentive to stay with the employer. The surplus from firm-specific training is also perhaps the most important single factor enabling firms to pay efficiency wages or wages above competitive levels.

The problem is that, while the employer and employee are disciplined initially by the usual ELM forces, firm-specific investments in training create a lock-in because the investments are sunk and have no value outside of the match. This creates a bilateral monopoly bargaining situation and the prospect of inefficient rent-seeking as each party attempts to take more of the surplus than might have been agreed to. This lock-in is a clear break with competitive markets that require that workers be mobile. The ILM can thus be efficient, where the surplus is made as large as possible by cooperative behavior, or it can be reduced or eliminated by adversarial behavior. Although the parties have an incentive to maximize the size of the surplus, there is no market mechanism that guarantees this behavior. Consequently, the ILM is vulnerable to market failures.

Ironically, for those who identify the Chicago School with the assumption of perfectly competitive markets, the lock-in problem created by ILMs was originally identified by some of the greats of the Chicago School; namely, Becker (1964), Mincer (1962) and Oi (1962), among others.

The nature of the market failure is easily described. If the parties could trust each other and prevent rent-seeking, each would have an incentive to make the optimal level of investment in the relationship. However, each party has an incentive to act opportunistically once the training is completed. The incentive to act opportunistically is great if the costs of the investment are not shared. Suppose, for example, that the employer could get the employees to bear all the training costs. Once trained, the employees would be in a vulnerable position. The employer could attempt to divert almost the entire surplus to
itself by threatening to fire the workers unless they accepted a minimal return. Pushed against the wall, the employees would have little choice but to remain with the employer and take any wage above the market wage paid in the ELM. Hence, the employer has a potential gain from acting opportunistically. Similarly, if the employer paid the entire training costs with the expectation of getting all of the resulting surplus, the employees, once trained, could threaten to quit unless the employer agreed to give up its return on the joint investment. Pushed against the same wall, the employer would be willing to take a very low return on its investment if the workers agreed to stay, because it could not secure equally productive employees on the external market.

In the face of the risk that opportunistic behavior might undo the mutual gains from investment in firm-specific training, the parties can develop self-enforcing norms that prevent opportunistische rent-seeking. The key characteristic of the self-enforcing norm or contract was described by Becker (1964). Becker's solution was for the employer and the employee to share in the costs of the investment. In a simple model with one employer and one employee, the cost sharing would create an incentive-compatible contract where the incentive for opportunistic behavior might be eliminated. Cost-sharing takes the form of a "training wage" that is lower than what the employee could earn in an ELM without job-specific training, but higher than the employee’s current productivity which is very low as he first acquires specific training. With the current costs shared, the payoff for both parties would come after the training is complete. In this case, the employer would have no incentive to fire the worker and the worker would have no incentive to quit because each would only receive a return on their investment if the relationship were maintained.

Unfortunately, a threat of employer opportunism remains even if the job investment costs are shared. Assume there are a number of employees. If the workers are not organized and cannot act in concert, the employer could pick off one vulnerable employee at a time, and threaten to fire the employee absent a wage reduction. For each individual worker, the threat to fire that worker is real because the employer might secure its expected return on the joint investment from the continued employment of the other workers. Each worker could be put individually in that position, and would both stay on the job and accept a lower wage, as long as the employer allows the worker to receive some benefit from the investment. That is, workers are better off by not quitting as long as they are paid any premium, however small, over the wage available in the ELM. Since each worker believes that the others will be offered the same bargain, each worker knows that quitting is suboptimal. Hence, the employer can successfully act opportunistically by getting the employees to accept a lower, though still positive, return on their investment.

This inequality in bargaining power is reduced if the workers can act in concert by becoming unionized. Forced to deal with the workers as a single unit, the employer’s threat to fire the workers does not ring true because the threat would also cause the employer to lose whatever it invested in the relationship. One of the important rationales for labor unions is that they can protect workers from opportunistic behavior by employers who might otherwise seek to divert the portion of the surplus that belongs to workers.

Can the same results be achieved without a union? Arguably they can. Where the employment relationship is ongoing or where the employer must regularly hire workers
in the ELM, the sharing of costs can be a self-enforcing norm. If the employer and its employees need to make ongoing investments in their relationship, employer opportunism in the initial period will discourage the employees from agreeing to the second set of investments. Hence, repeat play reduces the potential for opportunism. Similarly, if the employer must return to the ELM to hire additional workers, the employer’s concern for its reputation will deter it from behaving opportunistically; otherwise it will be difficult to hire new employees or to get these employees to invest in their relationship. Consequently, the importance to the employer of retaining a reputation for fair play, both externally and within an ongoing relationship, can make the arrangement self-enforcing. The self-enforcing feature of the arrangement will be greatest, obviously, where both are present – the employer needs to return regularly to the ELM and the relationship with each employee is ongoing (or repeat-play).

The option to unionize provides additional protection against opportunism by a nonunion employer. As noted above, the evidence is overwhelming that controlling for skill, job tenure, education, demographics, and location, unionized workers are paid more than comparable nonunion workers. This suggests that, in the unionized workplace, a large share of the investment in firm-specific training is paid for by the employer in the form of above-market wages. Given the higher labor costs associated with a unionized labor force, it may not be necessary for the nonunion workers to actually form a union, for the ever-present threat of unionization can be particularly effective in discouraging employer opportunism.

Ironically, although the union sector is struggling due to nonunion competition, the nonunion sector has benefitted from the existence of the union sector because of a deterrent effect on nonunion employers who might otherwise act opportunistically if not for the fear of being unionized.

2. Risk aversion
If the parties to the employment relationship were risk neutral, the great bulk of pay as well as promotions would be strictly performance based. That is because pay-for-performance creates the greatest incentives for employees to work at their highest level of performance, and is therefore wealth maximizing. The evidence, however, is that workers are more risk averse than are employers. That is, employees prefer to receive a stable income, rather than an amount that varies with the fortunes of the firm, even if the variable pay has a higher expected value than the steady stream. This risk aversion may be a function of employers’ better access to financial markets to smooth out the consequence of revenue fluctuations; but it simply may be due to the fact that workers don’t like fluctuations in their income. Stable does not mean constant; in fact, workers tend to prefer a steadily rising wage profile. That is, workers make more investments in the form of reduced income when they are younger, and receive the returns on those investments when they are older. The value of firm-specific investments thus only partly explains the widely observed fact that workers’ income tends to be higher as they get older.

If employees were not risk averse, the great bulk of pay would be performance based. Wage increases, bonuses, etc. would all be based on performance. Promotions too would be based on performance. The preference for pay-for-performance is that it is wealth maximizing. It creates the greatest incentives for employees to work at their highest sustainable level of performance.
Almost all businesses work to find the best tradeoff in their compensation arrangements between stable wages and the high-powered incentives provided by pay-for-performance. This is the same problem that the firm solves to determine the best efficiency wage level. Pay-for-performance may take the form of stock options or bonuses, or it may take the form of annual salary increases or promotions based on performance. In this regard, unionized firms tend to have the least amount of pay-for-performance. This might be due to a greater distrust of employers, higher-than-normal risk aversion, or a preference for minimizing pay disparities across workers.

3. Asymmetric information
Another important feature of ILMs is the presence of information asymmetries. Asymmetric information exists in the ILM when it is relatively more costly for one of the parties to observe or monitor the quantity and quality of worker inputs. Although the employee is involved in an ongoing relationship with the employer, the firm has information advantages over employees. In particular, the employer has superior information about the productivity of its workers. Remember that the labor demand curve depends on the marginal productivity of the worker. Except in the unusual case where the worker produces a good or service from scratch and sells it herself to the customer, the worker’s marginal productivity will be unobservable to the worker. The worker, however, has an information advantage over another key item: namely, the worker’s effort. Because monitoring is costly, employers can only estimate their employees’ work effort.

One answer to the problem of asymmetry of information has been provided by government regulation of the ILM in cases where employers have a significant information advantage and the stakes for employees are large. That is the case for both workplace safety and deferred compensation in the form of retirement pay. Moreover, employer reputation may not be a sensitive control in such cases because of the difficulty and complexity in observing existing safety conditions and deferred compensation, and in separating out the role of employer opportunism from changes in technology or the product market. The problem of asymmetric information is resolved here by government regulation in the form of the Employee Retirement Income Security Act (ERISA) and the Occupational Safety and Health Act (OSHA).

Outside of these areas where government information fills an existing gap, what explains the lack of complex contracting to control other information asymmetries more generally? The same question can be asked about job-specific training, which might also benefit from contracts that protect each party’s vulnerable investments. The union contract, although lengthy, hardly covers all possible contingencies. In the nonunion sector, there is no contracting at all. The answer to the existence of limited contracting rather than complex contingent claims contracts to deal with every contingency has to be the transaction costs inherent in contracting in the ILM.

4. Transaction costs
Transaction costs involve the costs associated with writing and enforcing agreements that govern the relationship between employers and employees. As noted in the section above, the presence of asymmetric information leads one to expect complex contingent-state contracts. As described by Wachter and Wright (1990),
[s]uch contracts would specify what happens in the face of potential exogenous changes in technology or in the demand for the firm’s output, and hence inputs. Combined with risk aversion and match-specific investments, such contracts would also describe the parties’ agreed-upon tradeoffs between income smoothing and the provision of appropriate incentives for correct reporting of asymmetric information.

This is no simple task; it would challenge even the best drafter of contracts. The term “bounded rationality” can be used to describe the inability of contract drafters to anticipate all the potential states of the world that might emerge, and how the parties should respond to each particular state of the world.

Several factors make contracting in the ILM particularly costly. First, many jobs are idiosyncratic to the ILM. Besides the large number of different jobs, workers within the same job will differ in their amount of firm-specific training. Second, the number of interactions is large. ILM contracts are not for a single delivery of a product on a date and time to be contracted for. The employee’s performance is continuous over the work time and over the days of employment. There is not one deliverable; there are many. Third, jobs are also interconnected. One worker’s performance affects the potential performance of workers who are in the same team, the same office, the same assembly line, etc. Finally, jobs evolve over time, fundamentally with changes in technology, or in more minor but continuous ways with changes in product demand.  

24 The importance of transaction costs makes labor and employment law central to the modern theory of the firm, which begins with Coase (1937) and is elaborated by Williamson (1985), Hart (1989), and other scholars. In the theory of the firm, simply stated, the firm faces “make or buy” decisions at every turn. Take the firm that makes computers. It must decide as to each part, for example, the microprocessor, the operating system, and the keyboard, whether to make it internally or purchase it through suppliers. Much of that make-or-buy decision is based on technology, patents, and similar factors. For example, the firm will normally produce for itself computer parts over which the firm has patent protection and has developed the technology.

In this decision-making process, a single question reappears: how easy is it to contract for any individual part or input? The prevailing view is that holding other factors constant, a firm is likely to buy in the market when the input being purchased is standard rather than idiosyncratic. It is likely to make rather than to buy when the input is more specialized or idiosyncratic. Also, the firm is more likely to make rather than buy when information asymmetries are important. Information asymmetries make third-party enforcement of a contract difficult and error-prone because key facts are not observable or verifiable by an arbitrator or judge. Finally, the higher the cost of specifying terms in advance – as in the case of repeated interactions, interconnected tasks, and evolving tasks and technology – the more likely the firm is to make rather than to buy.

All of these make/buy decisions turn on the amount of transaction costs, which is why transaction costs are central to the boundaries of the firm. What results is the following conclusion: buy when the transaction costs of contracting are low. When transaction costs are low, the parties can write a contract that is likely to be largely complete. Holes in the contract will still occur when an unexpected event happens over which the parties have not contracted. But, given sufficient existing contractual completeness, the court can usually fill the hole by inferring a contract term based on what the parties themselves would have done if they had anticipated the event. On the other hand, if an event occurs
for which there are no relevant contract terms, the court will assume that the event is outside of the four corners of the contract and is therefore not covered by the contract.

Hence, the theory of the firm resolves a paradox that is created by the four factors discussed above. Specifically, the economic theory suggests that, with idiosyncratic products (or, in labor market terms, workers with different amounts of job-specific training on tasks that may be unique to the firm) and high information asymmetries, the parties have an incentive to write complex state-contingent contracts. Yet these very same factors that contribute to high transaction costs make the contract more costly to write, more incomplete than complete, and hence more likely to be subject to judicial error. Consequently, the theory of the firm provides a different answer: instead of using complex state-contingent contracts, use no contract at all. Instead, “make” rather than “buy” by bringing that activity inside the firm and use the firm’s hierarchy to control that input (Coase, 1937; Williamson, Wachter, and Harris, 1975).

A transaction cost theory of the firm and the make/buy distinction is generally successful. One almost never observes complex contracting inside the firm, whether in the ILM or elsewhere. The firm might write simple employment contracts for some executive officers or specialized employees; but these contracts are usually short and straightforward, and mostly deal with the possibility of termination.

The collective bargaining contract is the one great exception to the non-contractual nature of the firm. It is the only detailed enforceable contract that the firm writes to deal with activities and with actors that are wholly inside the firm. Its policy implications, discussed in the next section, are extremely important.

III. CONTRACTING IN THE UNION ILM

As noted above, one way to deal with the complexity of the ILM and the potential for opportunism is to meet it head on; that is, to write or at least attempt to write the complex state-contingent contract that describes each potential state of the world that might occur and, for each possible state of the world, what the parties wish to happen. Bounded rationality and the many contingencies that the parties to an employment relationship face make that task impossible. But that is what union contracts attempt to do. And that is why those contracts often run to hundreds of pages. Even so, they are notoriously incomplete. The proof of the degree of contractual incompleteness in union contracts is the number of grievances that often occur and the costs of resolving them. These costs can include the fees of outside arbitrators, the time of lawyers (or managers and stewards) representing the parties, the work-time loss, and the ill-will that often continues beyond the conflict’s resolution.

Beyond dispute resolution costs, however, a frequently cited cost of union contracting is the rigidity imposed by the collective bargaining agreement. All of the factors cited above make rigid contract terms expensive, especially in a world where there are constant changes in technology, competitive pressures on the firm, and the state of the economy. Fixed contract terms cannot be the low-cost answer to these conditions.

Take the seemingly most easily resolved type of dispute: was the employee’s performance or behavior acceptable or sufficiently below contract levels as to merit the employee’s dismissal? Each of the elements that make contract writing difficult comes into play.
First, what level of performance is to be expected? Jobs are idiosyncratic as is the amount of specific training undertaken by the employee. In the extreme, standards may have to be set on an employee-to-employee basis. Second, the employee’s performance may be acceptable, even consummate some of the time, while other times performance is below that contemplated by the contract. It is no easy task to write a contract that defines what percentage of below-par performances merits discharge. The third factor, job interconnectedness, makes the task that much more difficult. Was unsatisfactory performance truly the fault of the employee or did another employee’s behavior, or even the supervisor’s behavior, lead to the unsatisfactory performance? Finally, how is the question of discharge to be resolved if the technology has changed to make the job more difficult, or if the product market has declined to make the employee’s performance less productive even for the same level of work effort?

All of these factors contribute to making dispute resolution difficult even for the core question of worker performance. The employee may claim that her work effort was high. The employer may complain that the employee’s productivity was low. Both may be true in the complex ILM, where work effort and productivity are not one and the same. Even apart from the difficulty of specifying the weight that each should have in judging performance, a problem is that the employer cannot easily determine the worker’s effort and the worker cannot easily determine her productivity.

If this were not complex enough, add yet another factor. Since effort and productivity are not easily observable and verifiable, they are subject to the parties’ manipulation. At the moment of dispute resolution, the neutral arbitrator or judge must rely on the parties’ representations about the facts. Even if the parties were acting in good faith, the dispute would be difficult to judge reliably. If the parties can manipulate the data to their own advantage with little risk of detection, the task of reaching the correct judgment is very difficult indeed.

Another interesting feature of the union ILM, to which I have already alluded, is that it handles the tradeoff between incentives and risk aversion with a near polar solution. Maximizing the value of the corporation and making workers as productive as possible is best accomplished by making pay dependent on performance. But in union ILMs, pay for performance is almost entirely absent. Pay increases are based almost entirely on job tenure and not on performance; even promotions may be based largely on seniority. In addition, the union “just cause” provision and grievance procedures make it very costly for the employer to penalize workers who are underperforming. These two factors alone – no extra pay for good performance and little penalty for poor performance – suggest a very negative effect of unions on firm productivity.

Union preference against pay-for-performance is partly ideological and partly based on distrust fed by asymmetric information. Unions typically favor narrow wage differentials on ideological grounds. Indeed a general finding in the union premium literature is that the most skilled workers get the smallest union wage premium, while the least skilled receive the highest premium. Distrust also plays a key role. Data on productivity are available to the company, but not the union. Even if the company makes information available to the union, it cannot be verified. If the parties are not in a relationship of trust and confidence, it is unlikely that they will agree to base pay on information that cannot be verified by both parties.

Another factor making the union ILM atypical in the contract world is that the
relationship is not a voluntary one as far as the employer is concerned. Generally, contracts are entered into voluntarily by the parties. Contract law serves as “the handmaiden of the parties” and supplies a set of default terms, most of which the parties can change voluntarily. Parties can restrict or expand the scope of the contract to fit their needs. There are few mandatory terms, with the exception of an implied obligation of “good faith and fair dealing.” Commercial contracts typically run for a limited term so that, if the parties do have a falling out, the contract can be terminated.

The collective bargaining contract is also unlike any other commercial contract. First of all, once a union is selected by a majority of bargaining unit members, the company is required by law to bargain with the union even if it prefers to bargain with another union or with no union at all. The employer must bargain in good faith over a set of mandatory topics, including wages, hours, and other terms and conditions of employment. The parties can also deploy “economic weapons” – the union can strike or the employer can lock out the workers – to pressure the other party should they reach an impasse in their bargaining. Although the law’s goal may be peaceful labor relations, that peace is achieved and punctuated by the use of these “weapons.”

None of this is meant as criticism of the NLRA in terms of its overall effect on the economy. In another chapter in this volume I argue that the NLRA was justified, and has in fact been extraordinarily successful. The argument brings in issues of political economy that are beyond the scope of this chapter, but its main point is worth noting here. Before the passage of the NLRA, labor history in the United States, as elsewhere, was filled with stories of violent strikes. At times revolution was in the air. Only the NLRA, and its nationwide legitimization of unions, collective bargaining, and peaceful strikes, was able to bring a close to that violent chapter of history.

Putting aside the social value of labor law, however, the union ILM has elements that are clearly costly to employers. Apart from raising wages and benefits above competitive levels, unions also bring about a degree of formalization in contract writing and enforcement that is inherently costly.

IV. ILM CONTRACTING IN THE NONUNION FIRM

Whereas the union sector takes on the complexities of the ILM by writing explicit contracts, the nonunion sector does not. This lack of contracting raises two questions. First, why does the nonunion ILM do virtually no contracting at all, versus, for example, some contracting but less than the union sector? Second, how can the nonunion ILM be successful in mitigating or eliminating the potential for employer opportunism?

Before answering these questions it is important to explain how I am using the word “contract.” Economists use the term contract to mean an agreement by the parties, which can be implicit or explicit and is not necessarily enforceable. Indeed the question of enforceability is not raised, nor is it of particular interest to economists, particularly since the contracts are often constructed to be self-enforcing. In this chapter, the difference between judicial enforceability and unenforceability is an important distinction. I will use the term “contract,” whether explicit or implicit, to mean a judicially enforceable agreement, and I will use the term “norm” to mean a mutual expectation that is not intended to be judicially enforceable.
What then explains the difference between the use of enforceable contracts in the union ILM and the almost exclusive use of non-enforceable norms in the nonunion ILM? At a theoretical level, the answer is given by the theory of the firm discussed above. Recall that the boundaries of the firm are drawn specifically so that the activities brought inside the firm are those where formal contracting is too difficult or costly, and where decision making can more efficiently be done unilaterally through the firm’s hierarchical structure. The union ILM is the great exception, in that the activity is carried out inside the firm and the employees bargain collectively with the employer and reach an enforceable agreement governing the employment relationship (Dau-Schmidt, 1992).

The implication of this divide is enormously important. If the transaction cost-based theory of the firm is correct, the unionized firm has a large competitive disadvantage. But there is a big “if” attached to that proposition. In the late 1800s, employees who were dissatisfied with their terms and conditions of employment would, from time to time, attempt to organize a union and engage in concerted activities. Some of those activities were illegal under state law, and the police or National Guard, or even the regular army, was brought in to subdue or crush them (Wachter, Chapter 15 in this volume). Indeed, the most compelling rationale for the passage of the NLRA was to end this industrial strife, which was harmful to the nation’s economy.

The point that can be drawn from the history is that, prior to the NLRA, the nonunion sector was a frequently dysfunctional organizational type. For the affected employer, the costs of strikes were just the tip of the iceberg. Dissgruntled employees had many ways to penalize a firm if they chose to do so. Since monitoring is not and cannot be continuous, employees in manufacturing or construction, for example, could harm the firm by sabotaging equipment or simply ignoring the need for repairs. Although documenting such activity is impossible, it is likely that, even prior to an outbreak of overt concerted activity, disgruntled employees had many ways to hit back at an employer who was acting unfairly or opportunistically. This point was emphasized by institutionalists such as Dunlop (1958) and Kerr (1954) and theorists like Becker (1964).

The militancy of workers in the early organizing drives of the 1930s and before appears to reflect intense worker unhappiness with the existing nonunion ILMs of the time. It is no surprise that worker unrest escalated in the Depression, given that employers have greater potential to act opportunistically during economic depressions, thus provoking increased industrial strife.

Conditions seem to be very different today. At least the extremely low levels of strike activity would seem to suggest that the nonunion ILM operates with less opportunism than in the past. How does the nonunion ILM now operate so as to defuse industrial unrest? At this point, we return to the theory of the firm and the core concept that activities brought inside the firm are both susceptible to norm-based governance and involve activities too costly for effective contract enforcement.

The norm-based governance of the nonunion ILM is integrally tied to the judicial doctrine of employment-at-will (Rock and Wachter, 1996). According to the employment-at-will doctrine, “an employer can fire an employee for good reason, bad reason, or no reason at all.” This doctrine can be criticized for appearing to condone employer opportunism; but this criticism entirely misses the point. The point of the phraseology is to signal a jurisdictional boundary rather than a legal rule that is applied in its literal meaning (Rock and Wachter, 1996). By stating the employer’s prerogatives
as broadly as possible, the employee who believes herself to be wrongfully discharged cannot sustain a claim. The only exception is a worker who falls into a policy exception such as race or gender discrimination, or illegal retaliation.

How can a legal regime that allows workers to be discharged for “no reason at all” be either fair or efficient? The answer lies in the reason behind the stark and unforgiving language.

To understand that reason, compare this actual standard with a counterfactual one where the firm states that “it can lower wages for a good reason, a bad reason, or no reason at all.” The wage reduction rule is not self-enforcing, and any firm adopting it would soon find it impossible to hire workers and difficult to retain them. This rule would be entirely self-serving and not joint profit maximizing. The reason is that a firm can quickly raise its profits by reducing wages; but the increase in profits comes entirely from the lower wages. Workers would obviously respond negatively to such clearly opportunistic behavior. Indeed, much of the industrial strife in nonunion firms of the 1930s arose when companies adopted just this kind of policy. Even though such policies were primarily adopted during severe downturns in the economy, firms could not convince workers of the need to accept the reductions.

The underlying reason for the workers’ distrust of a norm that allows firms to lower wages, even during a recession, goes back to information asymmetries: the firm’s economic condition is known to the firm, but not to the workforce. Even if the workers know that there is a recession and that production has fallen, they do not know how much of a wage reduction is merited. If the firm says that a 10 percent reduction is needed, the workers have every reason to distrust that assertion; maybe a 5 percent reduction would be sufficient to save the plant and the jobs at the plant. The firm has every incentive to overstate its problem to improve profits, while the workers have every reason to distrust the firm. This is precisely why a norm of “reducing wages to deal with a recession” is not an efficient, joint profit-maximizing norm. And, indeed, such a policy is rarely used.27 Instead, the almost universal response to a reduction in product demand is to reduce hours of work. This is an efficient and self-policing norm because the firm does not have a reason to overstate its case or overshoot the mark. If the firm reduces hours more than is merited by the reduction in demand, output will fall and profits will suffer more than they would with a smaller reduction.

Employment-at-will is more like the self-policing norm of “reducing hours in response to a recession” than it is like the inefficient norm of “reducing wages.” Discharging a productive worker for no reason is counterproductive for the firm. It may lower the wage bill, but the firm will lose the efforts of the worker (who is presumably more productive than the next new hire). The rational firm will not discharge a worker for no reason at all because it will suffer in the product market. Hence, a norm of employment-at-will is altogether different from a norm of lowering wages in response to a recession.

To many readers, the idea that product market pressures ensure that the rational firm would not discharge a worker for no reason might not be reassuring enough. However, there are other dynamics endemic to the nature of the ILM itself that tend to ensure that norms work well. There is now an established literature describing the conditions in which non-legally enforceable norms are a better choice for the parties than legally enforceable contracts (Ellickson, 1991). Those conditions, as discussed below, are those of the ILM.
A puzzle remains: if no rational firm would actually fire a worker "for a bad reason or no reason," what is the purpose of the language stating that they can do so? The answer is that the language is intended to be a signal to the courts: namely, it is intended to draw a jurisdictional boundary that leaves a firm's internal business judgments beyond the scope of judicial review (Rock and Wachter, 1996). From a firm's vantage point such a jurisdictional boundary promotes efficiency by leaving internal employment decisions to the discretion of the firm. From a social welfare perspective the jurisdictional boundary is also efficient (putting aside the regulatory protections offered to certain groups and activities) because of the high costs and risk of error associated with adjudicating discharge decisions.

Suppose that the employer needed to prove "just cause." Monitoring is costly, but it would be necessary if the decision were subject to judicial review. As noted above, it is always more difficult to prove a point to a court than to learn the facts. Even flagrant bad play such as theft may be hard to prove, but what about the more prevalent cases where the employee does not perform well enough? How many examples of poor performance would be sufficient? How harmful does each incident have to be to be worthy of discharge for cause? The correct decision would require the judge to learn the norms of the workplace and to sort through facts that are proffered by the firm, and that are often not verifiable by the court. Job specificity makes these decisions even more difficult. In contract disputes, the value of the actual performance is measured against the value paid. Idiosyncratic valuations make this task difficult, but this is exactly the same problem raised in the ILM where workers have firm-specific training. Market values of the appropriate wage rate and productivity are not available.

If judicial review is contemplated, the parties in the ILM would have to write contingent claims contracts to cover the many different states of the world that could result. The contract, no matter how complete, would necessarily be highly incomplete, particularly as more time elapsed. Courts deal with incomplete contracts in two different ways. When the contract is largely complete, the court may plug the gap with a term that the court believes the parties would have adopted had they known of this exact contingency. The more complete the contract, the better the court knows how a gap would be filled. However, if the contract is highly incomplete, the court takes a passive stance, and will typically rule that there is no contract for that event (Schwartz, 1992).

The labor contract involving the ILM, no matter how great the ambitions of the parties, is likely to be highly incomplete. A court faced with this type of incompleteness is thus likely to say that there is no contract for that event. And that would be the right response. By stating the employment-at-will doctrine in a form that amounts to saying that the employer can do whatever it wants, the firm is simply seeking a judicial finding that there is no contract. This jurisdictional boundary ensures that the courts will stay out of disputes in internal labor markets and leave the resolution to the norms of the workplace.

Clearly there is still potential for opportunism in an employment-at-will world where judicial enforcement is not available to employees. Recall that job-specific training leaves both parties potentially exposed because of their investments in the match. I have already discussed the potential solution to this problem above, but it is worth addressing in somewhat greater detail. The key is that safeguards against opportunism have to be self-enforcing since legal enforcement is ruled out.
The traditional solution emphasized in the neoclassical literature is reputational effects. Most employers have to make frequent trips to the external market. Reputational effects are the first constraint on employer opportunism. Employers with reputations for treating their workers unfairly will find it difficult to recruit and maintain a qualified work force. Even if employees do not learn about an employer’s reputation before hire, they will quickly learn it after being hired. Quit rates are very high in the first year of employment. Newly hired workers have made few investments in the job and lose relatively little by quitting and searching for another job. Even firms with declining employment are subject to this check if they must replace some of their departing workers.

The second check on employer opportunism is that dissatisfied employees can trigger their rights to organize and bargain collectively, thus replacing norms that are intended to be self-enforcing with legally enforceable contracts. This is a very powerful threat given that the newly unionized firm may find its labor costs increasing sharply (assuming a standard union wage premium). In a competitive, largely nonunionized industry, unionization may be fatal to the firm. It can be argued, given low rates of unionization in the private sector, that the threat to unionize is not much of a deterrent to employer opportunism. Perhaps that is true, but an employer would have to be highly risk tolerant to take that bet.

The final deterrent against opportunism is the self-enforcing norms themselves. The literature on norms has identified the characteristics that are most likely to lead to successful norm-based governance (Ellickson, 1991). Workable norms are joint profit maximizing. Since they are voluntary, they have to satisfy both parties. If norms with this characteristic can be developed, they are superior to contracts given the cost of contract writing and enforcement. This is particularly true in contracting over labor relations within a firm, where the frequency of interactions, the connectedness of events, and the evolutionary nature of the relationship make it impossible to write even a moderately complete contract. Fortunately, features that make contracting expensive in the ILM context are very similar to the factors identified by Ellickson as likely to give force to successful norms-based enforcement, such as being part of a closely knit group. Therefore, in the ILM, norms governance has the potential to be very successful, while contract writing and enforcement will be very costly.

In addition, Ellickson notes that the ability to apply sanctions is an important feature of norm governance. In the ILM, employees are empowered to apply sanctions by exercising their ability to quit, as noted in the groundbreaking article by Gary Becker (1964), as well as by participating in negative gossip, which has become increasingly damaging in today’s Internet world. The employee’s ability to apply sanctions in the ILM to deal with employer opportunism highlights the idea that ILMs are likely to generate self-enforcing employment norms that can ultimately trump contracting as a method of workplace governance.

The primary point of the theory of the firm is that contract writing in the ILM is inefficient. If it were efficient to contract over the terms of a particular activity, then it could be left to contracting in the external market, which benefits from the discipline and high-powered incentives of market forces. The ILM is internal to the firm precisely because it is too costly to do otherwise.

The point is underscored by the rising use of “outsourcing” today, which amounts to using the market to purchase labor services. Indeed, outsourcing of the labor function,
as an alternative to a union ILM, has become a major cost-cutting strategy, and thus a major profit producer for many firms. Anything that significantly increases the cost of employing union labor internally (such as judicial imposition of high contract enforcement costs) increases the attractiveness of outsourcing. The nonunion firm rarely finds outsourcing attractive because its norm-based enforcement system is less expensive than writing a contract with a firm that provides outsourcing labor. The “make or buy” choice is a very active area for decision making today regarding labor usage inside the firm.

Hence, norms start with a huge advantage over contract writing inside the firm. Because contract writing and enforcement are so costly and difficult, neither norms nor their enforcement have to be perfect; they just have to get over that low bar to succeed.

V. CONCLUSION AND POLICY IMPLICATIONS

Labor law and law and economics have not been close bedfellows. From one perspective, this is surprising. Labor law is primarily about negotiating and writing collective bargaining contracts, and contract law quickly embraced law and economics when Richard Posner showed that important contract questions could be answered by applying economic contract theory (Posner, 1973). Similarly, labor law is about one of the firm’s key inputs, labor, much as corporate law deals with another of its key inputs, capital. Corporate law found law and economics to be useful for thinking about relations between the firm and its sources of capital, yet labor law remains resistant to law and economics.

The explanation for the hostility of labor law to law and economics is apparent. First, labor law, unlike many areas of law, is inherently normative. The Wagner Act was itself pro-union and many of its supporters believed that unions would become the dominant industrial relations system. The law arguably became neutral between union and non-union industrial relations systems after the adoption of the Taft-Hartley Amendments, but labor law scholarship retained its strong normative edge in favor of unions. Most labor law scholars believe that collective bargaining is the appropriate mechanism for governing the employment relationship for working men and women.

The clash with law and economics was bound to be fierce, or at least spirited. The basic finding of labor economics – uncontroversial in every other area of law and economics – is that the quantity purchased of a good (e.g., labor) varies inversely with its price (e.g., wage). In labor law, such an assertion is a cause for battle, for it puts the central aims of the NLRA at odds with each other. The NLRA’s goal of equalizing bargaining power was to be realized by workers joining unions, and by unions raising wages. However, modern labor economics teaches that if union wages are increased above market levels, union employment falls. The goal is thus internally inconsistent.

The union movement’s hope was that wages were to be taken out of competition, so that the collective wage could be higher than the market wage. But this is unachievable in a competitive economy with a significant nonunion sector. Labor cost differentials between union and nonunion firms matter a great deal.

Over the past several decades, several attempts have been made to controvert the conclusion that if unions, through collective bargaining, force employers to pay more than
the market wage, union employment will decline. The efficiency wage theory suggested that if an employer paid workers an above market wage it could build a more loyal, experienced, and productive workforce. But if the high wage is to be a gift that calls forth greater productivity, it has to be given freely by the firm. In the context of collective bargaining, the union is seen as securing the gains for the workers from recalcitrant management; so in unionized firms, the workers’ primary loyalty is likely to go to the union and not the firm. The efficiency wage story makes much better sense in the nonunion labor market, where the gift of the high wage can be claimed by management. After all, the efficiency wage is simply another form of incentive payment which needs to be set by the employer at the level where the cost of the higher wage is equal to the productivity benefit that results from the higher wage.

The empirical test of the efficiency wage claim for the union labor market has produced negative results. A plethora of academic research found no positive relationship between unions and productivity (Hirsch, Chapter 4 in this volume). While unions raise wages above competitive levels, there is no offsetting productivity effect.

Traditional labor law scholars have tended to view the neoclassical model to be opposed to almost all policy improvements, even in the nonunion sector. There is much to this position, but the neoclassical model itself does allow for a number of policy interventions. To begin with, measures that improve the functioning of markets (i.e., that improve information flows and aid mobility) are supported by neoclassical economics. Moreover, given the problem of asymmetric information, modern neoclassical economic theory can support policy measures to improve occupational health (e.g., OSHA) and the security of deferred compensation arrangements (e.g., ERISA). Policy economists might disagree with each other as to whether specific regulations pass muster under benefit-cost analysis or whether the procedural elements are unnecessarily costly, but, as a theoretical matter, the existence of information asymmetries can justify regulatory oversight.

Regulations such as OSHA and ERISA, as well as the Fair Labor Standards Act, have another feature that makes them acceptable to policy-oriented neoclassical economists: because they apply throughout the economy, they do not cause cost discrepancies among union and nonunion firms.

The anti-intervention tilt of economic analysis comes back into force when policies attempt to improve the lot of workers by providing more procedural rights, such as restrictions on employment-at-will. Here, transaction cost theory is central. In particular, third-party enforcement of procedural rights inside the ILM is very costly as described in this chapter. Process is expensive, especially if it involves third-party review. Talk is not cheap, thus mandating more process is likely to lead to more outsourcing of the labor function.

A broader question is whether traditional neoclassical arguments miss the positive redistributational or welfare-enhancing effects of unions. This raises several complex claims. One is that, when labor unions raise wages, the effect is redistributive toward the bottom end of the income distribution. Putting aside the empirical question as to the position of union workers in the income distribution, neoclassical theory has a potent response. It tells us that using unions to redistribute income is a bad idea. The fundamental theorems of welfare economics tell policymakers to adopt policies that improve efficiency and then redistribute income through taxes and government expenditures.
Hence, even if unions have a positive effect on income distribution, those goals are better and more efficiently handled by other policy mechanisms.

There remain, however, several critical areas for policy reform and welfare improvements. The first concerns laws to combat discrimination or to give special protection to certain vulnerable groups. Whether or not there is a strong efficiency case for such laws – there may be for some and not for others (Issacharoff and Scharff, Chapter 13 in this volume) – they may have strong welfare justifications beyond the scope of economic analysis. The second opening for reform is at the bottom of the labor market. In the nonunion sector, employers are encouraged to play by the rules because of reputational effects and the threat of unionization. But some employers are not deterred by these forces. Small employers and those who are on a path to go out of business are not as likely to be deterred, nor are employers who hire undocumented workers. For these workers, there is no back-stop even to prevent such core breakdowns as the nonpayment of wages for work performed.

A more fundamental point involves the distinction between normative and positive analysis noted above. Much of traditional labor reform is normative and articulates societal values or what commentators believe the values ought to be. Normative analysis is not antithetical to law and economics, nor does normative analysis provide a critique of the positive theory of the neoclassical model. The fundamental theorems of neoclassical economics, as discussed in the first section, recognize the existence of a social welfare function, which incorporates societal values. Economists, as economists, have little to say directly about the contours of the social welfare function. For example, many labor law scholars favor collective bargaining as a goal in itself. Expressed as a societal value, that is, as an input into the social welfare function, these positions are not open to the positive critique of the neoclassical model.

At issue, however, is whether pro-union normative preferences are captured by the political economy that society has chosen. As stressed in Wachter (Chapter 15 in this volume), society’s choice of a political economy is more a political choice than an economic one. The neoclassical model takes the political economy or the social welfare function as a given. Its analysis of the positive or negative effects of union pay premiums or other outcomes takes place in the context of the accepted political economy. Today, the political economy of the United States is highly competitive. Hence, it is at this level that the normative argument in support of unions’ traditional goals loses support. That unions cannot take wages out of competition is a consequence of society’s choice of a political economy that stresses free competition.

NOTES

1. The author gratefully acknowledges the contribution of the criticisms and suggestions made by Cynthia Estlund, Barry Hirsch, Bruce Kaufman, and Howard Lesnick. The author also thanks Natalie Ditomasso, Sarah Edelson and Marisa Kirio for research assistance.


4. This follows the introductory treatment of labor markets. See, for example, Parkin (2012). For an intermediate-level treatment, see Varian (2010).

5. The neoclassical model was developed from earlier classical models that came to similar conclusions, but in less complete or rigorous ways. Adam Smith’s (1776) was the most well known. Thomas Malthus (1803) had a theory of the labor market which predicted that wages would always be pushed to the subsistence level because of the inability of individuals to control their fertility. More central to this chapter, Karl Marx (1867) had a theory that all value was created by labor and thus could be returned to workers, presumably in the form of higher wages. His theory was contested by Böhm-Bawerk (1889), who believed, correctly as it turned out, that savings was a form of postponed consumption and required a return (interest) in order to occur. If there were no return for savings, there would be little savings and little capital. The value of capital was thus not embodied labor as Marx had theorized.

6. See Katz and Rosen (1994), stating the first fundamental theorem: “As long as producers and consumers act as price takers and there is a market for every commodity, the equilibrium allocation of resources is Pareto efficient. That is, the economy operates at some point on the utility possibilities frontier.”

7. See Katz and Rosen (1994), stating the second fundamental theorem: “[P]rovided that all indifference curves and isoquants are convex at the origin, for each Pareto efficient allocation of resources there is a set of prices that can attain that allocation as a general competitive equilibrium.”

8. Labor economists are sometimes labeled as conservative because they disfavor direct government intervention in specific markets. This is a misguided labeling from my perspective. Such individuals may well be very liberal overall but prefer that redistribution be achieved through tax and expenditure policy.

9. In a recent study Ashenfelter, Farber, and Ransom (2010) conclude that firms have some wage-setting or monopsony power due to imperfect information and job differentiation. Although this line of research is just beginning, the prevailing assumption remains that, except in the very short run, the amount of monopsony power available to firms is small and therefore unlikely to affect the conclusions reached in this chapter.

10. For a standard microeconomic treatment of the union sector see Parkin (2012) for introductory treatment. See Varian (2010) for an intermediate-level treatment of fixing wages or prices above competitive levels. Union pay or total compensation consists of both the wage and nonwage benefits. For ease of exposition I use the term wage in place of total compensation.

11. Cartelization, although descriptive in economics, can be interpreted as pejorative in law. Cartelization is unlawful as a violation of the antitrust laws. However, the Clayton Act specifically provides a safe harbor for union activity. Consequently, as a semantic matter, if cartelization indicates unlawful antitrust behavior, then unions do not cartelize the labor market. To use cartelization pejoratively in the labor context thus requires a separate justification for why cartelization is harmful.


13. The decline in union employment was, of course, due to a number of factors. For an accounting, see Farber and Western (2002).

14. The version of Principles of Economics used here was published in 1920; however the original version of the book was first published in 1890.

15. The version of The Theory of Wages used here was published in 1963; however it contains an original reprint of the first edition of the book published in 1932.

16. Indeed, although the economy has still not fully recovered from the financial crisis of 2008 and the resulting steep recession, arguably the worst since the Great Depression, the evidence to date is that wages have not only not declined, rather they have also continued to grow (Daly, Hobijn and Lucking, 2012).

17. This is from the preamble to the Wagner Act: “The inequality of bargaining power between employees who do not possess full freedom of association or actual liberty of contract, and employers who are organized in the corporate or other forms of ownership association substantially burdens and affects the flow of commerce, and tends to aggravate recurrent business depressions, by depressing wage rates and the purchasing power of wage earners in industry and by preventing the stabilization of competitive wage rates and working conditions within and between industries.”

18. Moreover, between the passage of the Wagner Act and the Taft-Hartley Act, influential commentators such as Simons (1944) and Hayek (1944) had begun to question whether concentrated economic power, even in the hands of labor unions, was socially beneficial.


21. This section draws heavily from Wachter and Wright (1990).
REFERENCES


Cases and statutes cited