Reintroducing *The Economic Nature of the Firm*

RANDALL S. KROSZNER AND LOUIS PUTTERMAN

Two decades ago, noticing the increasing attention being given to questions of economic organization and the nature of firms, we compiled the initial version of this Reader. At that time, we noted the growing frequency of references to writings such as Ronald Coase's 1937 classic as well as work published in the 1970s by Armen Alchian and Harold Demsetz, Michael Jensen and William Meckling, and Oliver Williamson. The years that followed have seen those and related works serve as the core of a literature that has deepened in institutional detail, branched out into empirical studies, and inspired progress in formal analysis. Coase's pioneering role was recognized with the awarding of a Nobel Memorial Prize in Economic Science in 1991.[1] Williamson's ideas, which build on those of Coase and incorporate a number of new elements, have occupied a prominent place in the study of organizations by economists and students of related disciplines. What can be called the "new institutional economics" has had increasing influence, as shown by its treatment in a growing formal analytical literature by such writers as Oliver Hart, Bengt Holmstrom, and Paul Milgrom,[2] a marked departure from a time when the "mathematical-formalist" and nonmathematical literatures included few citations by the one of the other.[3] With unfolding research still suggesting that our original selection of "classic writings" indeed brought together core sources, with interest in the materials remaining strong, and with the fruits of more recent research providing the basis for further steps forward, we are pleased to offer our second revision of *The Economic Nature of the Firm: A Reader*.

How production and related activities are organized – and the relations among actors involved in these activities – is a subject of interest that crosses several disciplinary boundaries. As before, students of law, organizational behavior, management, finance, and related fields should find this compilation of economics sources on the organization problem useful to their own endeavors.

[1] On the centrality of *The Nature of the Firm* to his overall contribution, see Coase's Nobel Lecture (1992). Also see Williamson & Winter (eds.) (1991), a volume honoring the fiftieth anniversary of that paper's publication; and Williamson (1994).

[2] Some examples are Grossman & Hart (1986), Hart & Moore (1990), and Holmstrom & Milgrom (1991 [and this Reader]). Hart & Holmstrom (1987) and Holmstrom & Tirole (1989) provided surveys of the technical literature developing "new institutional economics" themes. See also the text by Milgrom & Roberts (1992).

[3] See Jensen (1983).

Indeed, some of these fields were concerned with issues raised in our literature at a time when economists seemed resistant to such a discussion. The debates of the 1950s and 1960s, when some neoclassical economists defended "black box" theories of the firm against "behavioralist" and "managerialist" challenges,[4] seem happily behind us, as the study of the contractual and organizational structure of business enterprises matures and incorporates new analytical tools and concepts. As economists make further contributions to the understanding of these matters, controversy about the appropriate scope for inquiry has given way to a more sustained and illuminating discussion about why there are firms, what determines firm–market boundaries, and why firms are organized as they are. The papers collected in this volume include many of the most influential contributions to this evolving body of work.

Although we see the gap between formal theory and discursive approaches as having narrowed in recent years, this volume continues to emphasize the latter. The main reason is that verbal presentations are accessible to a wider audience, including both scholars in some of the disciplines mentioned previously, law and business students, and advanced undergraduate and graduate students of economics and other disciplines. Just as the better ideas in the discursive literature tend eventually to be taken up in formal models, so the better ideas in the formal literature are soon – if not simultaneously – given nonmathematical exposition. We excerpt mostly verbal portions of some technical papers representing such ideas, and we favor discursive treatments where possible. The equations and graphs included are either unusually accessible and illuminating or are retained for a hint of the flavor of the original.[5]

We also concentrate on the theoretical literature rather than empirical work. This is the realm in which new interpretive ideas are born, although the ideas developed herein should ultimately prove their power in analyzing cases and data. Empirical literature related to the ideas represented in this volume has expanded rapidly in recent years, but doing justice to that literature would require another volume, not simply a few additions to the present one.[6]

The papers and edited texts collected herein fall into four overlapping groups. The first concerns the "division of labor" among markets, hierarchies, and other coordination devices. It consists of selections from some of the older, classic writings on our topic, plus some more recent pieces, all helping to introduce the general theme of the economic nature of the firm and its place in the market

[4] See Machlup (1967). Coase (1991b, p. 52) draws attention to related issues.

[5] We select, for example, Hart (1989) rather than Grossman & Hart (1986) or Hart & Moore (1990); we include Bowles & Gintis (1990a) rather than Bowles (1985) or Shapiro & Stiglitz (1984); and we excerpt from Zingales (1998) rather than Rajan & Zingales (2001).

[6] Examples of this burgeoning empirical literature include Baker & Hubbard (2003); Crocker & Masten (1991); Goldberg & Erickson (1987); Hubbard & Weiner (1991); Joskow (1985, 1987, 1988a, 1988b); Kenney & Klein (1983); Klein (1988); Leffler & Rucker (1991); Masten & Crocker (1988); and Mulherin (1986). Klein & Shelanski (1994) provided a survey.

system. The volume begins with Adam Smith's classic discussion of the division of labor in the manufacturing establishment and in society as a whole, and its relationship to the development of markets. Smith tells us how efficiency is served by bringing together many hands and assigning each to specific tasks, but his treatment of the firm remains more technological than institutional or contract-theoretic. Next is Karl Marx's discussion of the productivity effects of "cooperation" in the workshop and of the capitalist character of the factory system. In this reading, differences in character are highlighted between firm and market, production and exchange. Following then are excerpts from Knight, whose ideas on the role of risk bearing remain important to discussions of both employment and financial relationships; from Berle and Means, who set off the modern discussion of financier–manager agency problems;[7] and from Hayek, who provided the classic discussion about the relationship among information, the level of economic decision making, and the price system. Finally, having begun the part with Adam Smith's "invisible hand" view of the market, we conclude it with a selection from a recent discussion of the "visible hand" of corporate governance by Zingales.[8]

Part II of this Reader includes papers dealing with the scope of the firm. Coase's famous paper, which simultaneously presents a theory of the differences between internal and market resource allocation, initiates this literature with the argument that a firm will expand until the cost of undertaking additional transactions through market exchange falls below the cost of undertaking those transactions within a common hierarchical structure. In the pieces by Klein, Crawford, and Alchian and by Williamson, Coase's focus on costs of using the price mechanism and on limited entrepreneurial attention gives way to a theory of integrated ownership in the presence of idiosyncratic investment. Although using much the same explanatory framework as Klein et al. to account for full vertical integration,[9] Williamson's piece adds the dimension of transaction frequency and attempts to account for intermediate modes, which he labels relational (or "bilateral") and neoclassical (or "trilateral") contracting.

Whereas the Klein et al. chapter and the first Williamson chapter offer explanations for vertical integration, they have less to say about what limits the growth of firms. The fourth chapter in this part, from Williamson's 1985 book, attempts to address that issue and go beyond the managerial-attention approach of Coase and the control-loss counterpart to that approach that was used earlier

[7] The enormous impact of Berle & Means' *The Modern Corporation and Private Property* was demonstrated in the special issue of the *Journal of Law and Economics* (June 1983) on the fiftieth anniversary of its publication. See especially the article by Stigler & Friedland (1983). We note later how several papers in Part IV of this Reader can be viewed as responses to Berle & Means.

[8] Chandler (1977) used the phrase "visible hand" to refer to the controlling hand of corporate management.

[9] Earlier thinking along similar lines to that of Klein, Crawford, and Alchian is found in Williamson (1971, 1975).

by Williamson himself. Williamson asks why the head office of a decentralized conglomerate could not give its divisions effective autonomy, thus replicating the "high-powered incentives" of the market, while outdoing the market by the selective imposition of rationalizing and coordinating measures. His answer is that selective intervention is impossible, and that it therefore follows that the incentives facing agents in firms are typically "lower-powered" than those of independent entrepreneurs or self-employed people. Milgrom and Roberts' paper (1990a) offers another reason why hierarchy becomes costly: subordinates can pursue private ends by distorting the information they supply to their superiors, and decision makers are forced to discount the information provided them – and may even find it optimal to restrict information flows. Rather than present a unified alternative theory of firm boundaries, Holmstrom and Roberts' final chapter of the part reviews earlier theories and uses intriguing real-world examples to show how well-crafted contractual arrangements can substitute for both internal organization (i.e., the firm) and reliance on arm's-length transactions (i.e., the market).

Part III addresses internal organization and the human factor, which in most instances contracts with the firm under what some of our authors term "the employment relation." Several of the chapters give centrality to the question of how effort is elicited when the contribution of the individual is costly to observe and, hence, to appropriately reward in a team setting. Alchian and Demsetz's provocative and influential discussion, which explains the hiring of employees by a residual-claiming employer through its analysis of this problem, is the first chapter in Part III. Karl Marx had anticipated the modern agency approach to the employment relationship with his discussion of the "extraction" of "labor" from "labor power" (see Part I), and radical economists have made contributions to the internal-organization literature by expanding on Marx's insights.[10] In their now-classic "efficiency-wage" model, Shapiro and Stiglitz (1984) rediscovered the role of Marx's "reserve army of the unemployed" in parallel with radical writers such as Samuel Bowles (1985). Whereas this approach has a more macroeconomic aim in the hands of Shapiro and Stiglitz, the focus of Bowles and co-author Herbert Gintis's chapter is on other issues, including doubts about whether the capitalist employment relation is, in fact, the most efficient solution to the team-effort problem.[11]

In Chapter 15 by Williamson, Wachter, and Harris, the monitoring problem takes second place to the previously noted problem of idiosyncratic investment.

[10] Their emphasis on the "effort-extraction" problem contrasts with an increasing tendency to ignore Marx's labor-based value (i.e., price) theory or to treat the latter as having only normative or sociological importance. In addition to the works cited herein, see Goldberg (1980b), Stephen Marglin (1974), and Richard Edwards (1979).

[11] For a parallel exposition and interesting comments by Stiglitz & Williamson, see the exchange in the Winter 1993 issue of the *Journal of Economic Perspectives*, pp. 83–114.

These authors argue that firms attempt to prevent haggling over and possible extraction by workers of the rents from firm-specific human capital. They do this by designing internal job ladders that orient employees toward the long run, while permitting some relaxation of current period monitoring, with resulting benefits for "atmosphere" and, hence, employee effort. Competition for promotion nevertheless can be keen, as in the "tournament" models formalized by other authors.[12] We present an excerpt from Holmstrom and Milgrom's recent article, which provides new formal results on the design of incentives by principals, as Chapter 16. The authors corroborate an old intuition that offers an explanation for the "lower-powered incentives" referred to by Williamson. They also provide important insights into specialization, the way tasks are assigned and combined into jobs, and other issues.

With the partial exception of Williamson et al., the chapters considered thus far in Part III adopt a standard motivational model in which agents are strictly rational and concerned only with their private incomes and leisure. Bewley's chapter is inspired by interviews with managers whose way of explaining why wages are not reduced during recessions emphasizes behavioral considerations such as worker loyalty and reciprocity. In the full article from which the chapter is excerpted, Bewley showed how ideas reminiscent of those in George Akerlof's "gift-exchange" model and in a growing body of behavioral and experimental research (Camerer & Fehr, 2004; Fehr & Schmidt, 2003) can be formalized.

Whereas wage flexibility is rare in the firms studied by Bewley, it is found in some jobs in which profit shares account for a non-negligible part of compensation. Pencavel studies the limiting case in which workers are full residual claimants – worker-owned firms in the U.S. plywood industry.[13] His finding that worker–owners achieve higher productivity than workers in conventional companies seems at odds with Alchian and Demsetz's claim that a central agent must receive the residual so as to be motivated to monitor the production workers. Yet, his observations about why worker-ownership is relatively rare are in the tradition of Knight's views on worker risk-aversion.[14]

The final section, Part IV, concerns questions of firm organization and behavior related to financing and ownership. The first three chapters present different responses to the problem of "separation of ownership and control" in the modern corporation – the problem raised by Berle and Means in Part I. These chapters begin with Manne's frequently cited argument regarding the monitoring function of the market for corporate takeovers. Next is Fama's contribution,

[12] See Lazear and Rosen (1981) and Malcomson (1984).

[13] Less extreme forms of profit-sharing are studied, *inter alia*, in Blinder (ed.) (1990) and Kruse (1993).

[14] For an alternative view of the relative rarity of worker-ownership, see Hansmann (1988); for a survey of such views, see Dow & Putterman (2000).

which develops a second view of markets as monitors of internal discretionary behavior, this one focusing on managerial compensation as an ultimate source of managerial accountability. Chapter 21 is a major portion of the article by Jensen and Meckling that helped spawn a large literature in the field of corporate finance and that – similar to Manne but both more broadly and more subtly – answers questions about the financier–manager relationship with reference to the monitoring properties of capital markets.

The fourth piece of this section, Chapter 22, teams Fama and Jensen. Taking the idea that corporations incur greater agency costs than closely held firms as a given, they ask why the corporate form is ever chosen, and answer with a formal argument about capital requirements. They also provide less formalized discussions about comparative advantage in managing versus financing and about the risk-reducing benefits of limited liability and diversification.

Chapters 23 and 24 reflect the continued vibrancy of debate and contrasting viewpoints on the ability of existing capital-market arrangements to control managerial malfeasance. Holderness, Kroszner, and Sheehan's chapter argues that contrary to the predictions of Berle, Means, and many of their successors, corporate ownership by managers has become less, not more, diffuse over time. Bebchuk and Fried view corporations as rife with malfeasance and offer an agency-theoretic outlook on how managers extract value from shareholders.

The final two chapters reach beyond questions of financing to treat the institution of firm ownership more broadly. In a discursive treatment of ideas developed in more formal, co-authored papers, Hart argues that the essence of ownership is not the right to residual income or the responsibility for monitoring but rather the right to make decisions that otherwise have not been determined in prior negotiations between parties to an enterprise. In arguments reminiscent of but somewhat distinct from the ideas of Klein et al. and Williamson, Hart argues that the efficient assignment of this residual control right is to the owner of critical, often idiosyncratic, inputs and that if there are several such inputs, they should be commonly owned. In Chapter 26, Putterman integrates financial agency and control-right concerns, suggesting that the supply of risk-bearing and financing services, substantially determined by the distribution of wealth, has implications not only for the financial but also for the internal incentive structure of firms. He argues that the factors that lead to the nonidentity of owners and labor suppliers also can be viewed as being responsible for the characteristic incentive problems of the workplace. He notes that the unbundling of the control, revenue, and alienation aspects of ownership – although problematic – shows signs of alleviating the trade-offs between efficient risk bearing and provision of incentives.

We organize our discussion of the literature from which our readings are drawn around a set of themes that feature prominently in it. Viewing firms as distinctive modes of organizing economic activity, the literature attempts to explain two related questions: Why do they exist? What determines their sizes and scopes? However, two more fundamental questions that are bound up with and, in some respects, prior to the first two are: Just what is the thing we call a firm? How distinctive are firms as organizing modes? Or, relatedly: How sharp are the boundaries between them and the environments in which they operate? Because the answers to all of these questions are closely intertwined, the order in which we address them is, to some degree, arbitrary. As a matter of convenience, we choose to begin with a provisional discussion of what firms are, then move to the explanation of their existence and the determinants of their sizes and scopes, and finally address the extensive literature on the sharpness of firm–market distinctions. We inevitability begin to touch on that literature under the earlier topics, which also brings us back full circle to the question of the firm's nature.

After considering the questions just raised, the remainder of this introduction (paralleling the structure of the book) moves into a more detailed discussion of how firms are organized, focusing on two types of relationships: those involving ownership of assets and those involving contracting with labor providers. We note how contributors to the literature can be partially classified into those who view the employment contract as the essence of the firm and those who view asset ownership as its essence, with a third group seeing both sets of relations as equally important or as manifestations of a common principle. We then survey the parallel literatures on agency and incentive issues in the relations between managers and workers, on the one hand, and between owners and managers, on the other. Finally, we discuss the question of why firms might be organized as they are (e.g., with respect to the distributions of control rights and risk bearing and the concentration of ownership) and how this might matter to our view of their economic nature or essence.

What is a firm?

In standard microeconomics, as mentioned previously, firms are the economy's basic units of production, just as households are its units of consumption. Firms purchase inputs such as labor services and materials from households and from other firms, transform these inputs into goods and services, and sell the latter to households and firms with the objective of maximizing the difference between their revenue and their outlay. As between the two entities, firms differ from households in that they maximize profit rather than utility and in the different activities (i.e., consumption versus production) in which each engages.

In the literature exemplified by this Reader, however, most authors have something more in mind when referring to a firm than simply a profit-seeking transformer of inputs into outputs. A set of factor owners could hypothetically agree today to contribute their inputs at the going rate so as to produce something for which an unmet demand is perceived, divide any surplus earnings, and go their separate ways tomorrow according to newly observed opportunities. Producers' coalitions of this type would differ, however, from the literature's *firms* because the latter usually entail (1) long-term contracts between at least some input providers; and (2) an assignment of control rights in which some agents hire others and direct them in the activities of production. Rather than being momentary assemblages of cooperating factor suppliers, then, firms are ongoing organizations that manage and coordinate the activities of participating actors.

Asking where the boundaries of the firm lie raises some difficult theoretical issues. Most authors, however, have relatively simple notions of control and ownership in mind when deciding where one firm ends and another begins. Few if any question that a company having multiple plants producing the same or even different products should be treated as a single firm as long as the rights to the residual earnings and to hire and fire the managers of subunits are ultimately in the hands of the same owner or owners. We return to the centrality of ownership to both the nature and the definition of the firm later in this introduction.

Why are there firms?

That the previous discussion is but a first pass at the question of what firms are becomes clearer when we encounter the literature on *why* there are firms. If by *firm* one meant simply an entity that engages in the transformation of inputs into outputs, then the "Why are there firms?" question could be answered trivially by noting that there is a demand for the products in question at prices at least equal to the combined cost of the required inputs. However, although product demand may explain production, it hardly explains the need for an organization of production with such features as long-term contracts, risk bearing by owners, and hierarchical or centralized coordination of those engaged in the production process. "Why are there firms?" is a question to Coase, Williamson, and others because with traditional microeconomic theory assuming that the *price system* coordinates the allocation of resources to their most valued uses, such organizations might appear to have no economic *raison d'etre*.

If standard theory offered any explanation for the existence of firms, it would seem to have been a technological one. Factors of production are needed in certain combinations, and some scales of production are more efficient than others. Once a certain scale is reached, the need for coordination seems obvious. The assembly of automobiles in large factory buildings in which hundreds of workers are arrayed along assembly lines must be coordinated by some type of organizing entity. Having independently contracting workers each owning subsets of the tools and machines comprising the factory and contracting with one another without the coordination of an employer is out of the question, right? However, on further inspection, this obviousness gives way to a mystery. For, as Williamson in particular has forcefully argued, there is no *technological* reason why the participants in this production process could not each operate their own firms, buying and selling their semi-assembled components as they move along the production line. Even an accepted "technological" requirement of coordination or management fails to explain the firm as organization because the contributors of varying specialized inputs could simply contract for managerial services from other specialized agents, as the musicians in an orchestra might hire a conductor or as farmers contract for storage and other services from a cooperative or other company. It is the puzzle of which *organizational* factors give rise to the existence of an overarching framework of ownership, management, and coordination that motivates the prominent hypotheses of Coase, Williamson, Alchian and Demsetz, and others.

Coase initiated the contemporary literature by pointing out that when economic agents interact with each other, they incur "transaction costs" that vary with the mode of interaction. In the market, he argued, agents interact by negotiating exchanges, with existing prices serving as signals of the opportunities facing each supplier of a service or demander of a product. In a firm, conversely, a central coordinator or entrepreneur manages the allocation of resources such as the machinery owned by him and the manpower that he employs. Were the production process to be undertaken through market interactions alone, Coase wrote, the interacting parties would have to negotiate new actions and new terms of exchange each time a change in market or technological conditions made profitable a change in the activity being undertaken. Owners of the substations of the hypothetical production line would have to determine the prices at which to exchange their semifinished products – a difficult task in the absence of external markets for such goods – and these prices would be subject to change in response to changes in factor supply, technical conditions, and so forth. Such trivial matters as a breakdown in a piece of equipment, a stoppage of electricity, or a small change in the most profitable product mix would give rise to costly renegotiation of the full set of bilateral exchanges. The costs of such renegotiation, and of what Coase referred to as the "price discovery process," could be economized by agreement to a framework of ongoing relations under which a central coordinator would be granted the authority to reassign tasks and to offer altered payments as new contingencies arose.

Whereas Coase's ideas attracted enough attention to be cited and even reprinted occasionally in the early postwar decades following their initial publication, the "transaction cost" approach was largely ignored by the modernizers of economic theory who gained increasing dominance in those years. Parsimony and mathematical rigor were the order of the day, and theories that appeared capable of predicting economic phenomena using simple axioms held sway. In such an environment, the concept of "transaction costs" tended to be viewed as a vague and plastic one, inviting suspicions that it might be casually molded to explain whatever anomalies the analyst felt unable to explain by more "rigorous" methods.

By the 1970s, however, an increasing number of economists had come to view existing theory as failing to explain the organizational dimensions of the economy at the micro level.[15] At about the time that economic theorists were beginning to attack this and related flaws by treating problems of information and agency with formal tools, Williamson developed a variant of transaction-cost economics that would dominate the organizational literature, quite broadly, and inspire significant subsequent formal analysis as well. In Williamson's approach, as in Coase's, the market and firms were treated as discrete modes of organizing economic activity, and the choice among these modes was hypothesized to be dictated by economizing with respect to the cost of transactions among the different parties involved. Whereas Coase emphasized the costs of price discovery and negotiation, however, Williamson put his emphasis on the problem of investment in assets specific to a given venture. In an argument paralleling that of by Klein, Crawford, and Alchian, he suggested that when agents contemplated investing in assets that would have far lower returns outside of the activities to which they were initially dedicated, they would become subject to the possibility that owners of complementary assets would attempt to opportunistically extract their *quasi-rents* (i.e., the difference in returns between the dedicated and next-best uses) by threatening to withdraw their inputs. The threat of quasi-rent appropriation (or, in Williamson's terminology, *hold up*) would prevent such investments from being made without organizational safeguards, of which common ownership or vertical integration was the simplest example. Thus, the stations on the hypothetical assembly line would be owned and managed by an integrated entity rather than by a multitude of separate firms transacting across market interfaces, for the organizational rather than technological reason that separate firms would risk hold up, whence any who dared own one would face unacceptable danger.

It is worth noting the link between the Williamson and Klein–Crawford–Alchian idea of asset specificity and the idea of relational exchange or contracting introduced in the law and economics literature by MacNeil (1978) and in economics literature by Goldberg (1980a). As Goldberg stated, firms or individuals having assets uniquely suited to each other's needs can engage in a "relational exchange" that differs from the "arm's-length" trade in markets for generic goods and services in that the identity of the trading partners, and their ongoing cooperation, are matters of interest to each. This may not have been the

[15] One economist, Harvey Leibenstein, even argued that "a branch of economics is missing – micro micro theory." See Leibenstein (1979); see also Leibenstein (1966, 1982).