Osgoode Hall Law Journal

Volume 33, Number 1 (Spring 1995)

Article 4

The Economic Structure of the Firm

Robert Flannigan

The Economic Structure of the Firm

Abstract
Considerable effort has been channelled into theoretical investigations of the structure of the firm over the past several years. Most of the new work has been produced by economists. Lawyers have been content simply to draw upon the economic arguments, often in an uncritical way. The author examines the various economic models and identifies their shared dependence on the significance of an actors' control over the employment of assets. The control proposition is then further developed in the course of the construction of a general model focussing on the production unit.

THE ECONOMIC STRUCTURE OF THE FIRM©

BY ROBERT FLANNIGAN*

Considerable effort has been channelled into theoretical investigations of the structure of the firm over the past several years. Most of the new work has been produced by economists. Lawyers have been content simply to draw upon the economic arguments, often in an uncritical way. The author examines the various economic models and identifies their shared dependence on the significance of an actors' control over the employment of assets. The control proposition is then further developed in the course of the construction of a general model focussing on the production unit.

Dans les années récentes on a mis en marche beaucoup d'enquêtes théoriques vis-à-vis la structure des firmes. Pour la plupart, ce sont des économistes qui ont fait du travail sur ce sujet et les avocats n'ont pas considéré les arguments économiques trop critiquement. L'auteur examine les différents modèles économiques et identifie, parmi eux, un aspect commun qui met l'accent sur le contrôle des acteurs quant à l'usage des actifs. Cette proposition de base est développée par la construction d'un modèle qui se concentre sur l'unité de production.

---

* B.Sc., LL.B. (Alberta), LL.M., S.J.D. (Toronto), Professor of Law, University of Saskatchewan. I am grateful to those persons who commented on the numerous drafts of the manuscript. I am particularly indebted to Harold Demsetz, Oliver Hart, and Greg Dow.

I. INTRODUCTION

A business organization has both a legal and an economic structure. The legal structure is comprised of the various rules which define the rights and responsibilities of the parties in the different types of business organization. The economic structure is simply the material relations of the particular economic arrangement. It is easy enough to identify and describe these cognate structures in a given case. It is a more difficult task to develop theoretical models to organize the specific legal or economic characteristics of the individual organization forms. For many years, few lawyers or economists engaged in this task. Eventually, however, the attention of economists did turn to the elaboration of a number of theories purporting to explain the physical existence and size of firms. Lawyers, during this time, did not produce their own distinct theories of legal structure. Rather, they invoked those new economic models that served their particular theoretical purposes. These efforts to enlist the support of economics, however, were often rudimentary or largely rhetorical. Consequently, the actual application of economic theories in the legal analysis of the firm has been limited. Part of the difficulty is that economists themselves admit no consensus on the specific elements and mechanisms of firm structure. It may be, however, that there is order underlying the surface disparity of the various economic theories. If that could be established, it would imply, or lay the foundation for, a coincident theoretical model of legal structure.

The primary object of the discussion that follows is to construct a model of the economic structure of the firm. This is an economic analysis of the physical arrangement. It is not an economic analysis of the legal structure (*i.e.*, the legal rules) although, once constructed, the model is intended to inform that analysis. The primary object is pursued initially through a critical review of the various theories economists have offered to explain the structure of the firm. The analysis reveals a common (either overt or derivative) appeal to control as the explanation for firm structure. Generally speaking, the structure exists for, and is limited by, the projection of risk (control). Specifically, the significant benefits and major limitations of the firm are shown to be located in the actor's control over the employment of human and physical capital. Control (fiat), accordingly, is the essential defining attribute of the firm. This fundamental proposition, widely accepted but poorly understood, is then developed in a model in which the production unit, rather than the transaction, is made the focus of analysis. This departure from

transaction cost economics (where the transaction is the focus of analysis) allows for a theory of structure that more easily accommodates or explains the variation in the material relations of the firm arrangement.

As will soon appear, the discussion is exclusively economic in nature and parts of the material deal with what, for some lawyers, are novel concepts. With this in mind, an effort has been made to make the discussion accessible to a general legal audience with at least an interest in economic analysis. Still, in certain instances where clarification or elaboration is required, the reader may find it worthwhile to make reference to the commentary in the multiple sources provided, chosen in part for their own relative accessibility.

II. FIRMS AND MARKETS

Firms are distinguished from markets by their production function. Markets involve only exchanges. A market does not produce goods. Without prior production, or the promise of production, there is nothing to exchange. Markets therefore presume the existence of production units. These production units are generally labelled "firms." The nature of this "firm" or production unit may be investigated initially by analyzing the production of a particular good.

A simple example illustrates the choice an actor will make between exchange (the market) and production (the firm) in order to obtain a particular good. Consider the production of good Y, a process involving two stages. In the first stage, good X is produced. In the second stage, good X is consumed in the production of good Y. The two stages utilize separable serial technologies. The issue for an actor intending to produce good Y is how to obtain good X. The actor can choose to enter the market for good X and simply purchase the required amount. Alternatively, the actor can enter the factor market for good X and purchase the means to produce it internally. Recourse to the market for good X leaves production in the hands of another. Recourse to the factor market brings production within the control of the actor. This indicates that the nature of the firm can be understood in terms of the position of the actor subsequent to the use of the market. At that time, if the actor purchased good X, there is no process remaining to occur in its production. The actor's firm is limited to the production of good Y. The firm of some other actor produces good X. Purchasing the means to produce good X, on the other hand, contemplates action. Further effort on the part of the actor is necessary in order to convert

the factors into product. Specifically, the actor will be required to make a series of choices concerning how the various factors will be coordinated so as to eventually produce good X. The actor, in other words, will control the employment of assets. On this view, where firms are associated with production, the nature of the firm is found in the applied control of the actor. Firms, then, are arrangements for the exercise of control over the employment of assets.

The firm is both a structure and a process. The structure is created by the collection of contracts the actor negotiates to acquire the productive factors. The exchange aspect of these contracts, once they are executed, is exhausted. It is their content that now becomes significant. That content, in the case of each particular contract, is the acquired right to control the associated physical or human capital. Usually, in the case of physical capital, this right is obtained by contracting for ownership. More limited rights to control may instead be purchased through a loan, lease, license, or other contractual form. For human capital, the right to control is acquired through the negotiation of contracts of service. The combination of these rights creates a framework or structure for the realization of the actor's choices.

The firm is a process in both a structural and a production sense. Initially, the actor must exercise control (the process) over existing assets (often borrowed cash) to allocate them in such a way as to acquire the optimal collection of pre-production control rights. That is, the actor must make choices as to the nature of the structure to be created. Thereafter, in the production stage, the actor must again make choices (the process) in the manipulation of the acquired control rights. These structural and productive processes interact with and influence one another. Thus, the actor's acquisition of control rights will be affected by the kind of production contemplated. Similarly, the exercise of control over one factor may be circumscribed by the degree of control exercisable over another factor. These processes, and the structure they create, define the essential character of the firm. Fundamentally, the firm is a control structure—an arrangement for the idiosyncratic taking of risks.

Firms and markets are associated with different processes. Markets involve exchanges, firms involve production. This is not to say, however, that firms and markets are unconnected. Firms acquire intermediate goods in markets and supply intermediate and final goods to other markets. Markets also regulate firms. Product, factor (managers, capital), and control markets discipline firms and determine their viability. Further, in one respect, a firm can replace a market. When a producer of good Y integrates the production of good X into

what is now a firm of two combined but separable technologies, the former exchange to obtain good X is replaced by its internal production. The movement of good X into the good Y production stage becomes exclusively a matter of production (a directed transfer), rather than an initial exchange followed by a transfer. This does not mean, however, that firms can do what markets do. It only means that the market is utilized at a different point in time. The acquisition of good X was achieved when the actor acquired the means to produce good X.

Firms and markets are necessarily connected in one other important general sense. The production boundary of a firm occurs at the interface where the control of one actor is met by the control of another actor. The inability of actors to control beyond their distinct collection of control rights necessarily requires an exchange if there is to be a connection between their production units. The absence of control, in this respect, requires the presence of exchange. The fact of exchange marks or identifies the pre-exchange control boundaries of the interacting actors (firms).

The observation that the scope of a firm corresponds to the ambit of the actor's control applies both to the reach of the productive process (the extent of control) and to the composition of the actor (the source of control). For production, the boundary is found at the point where no further control rights are available to be exercised. Exchanges to acquire additional control rights are required in order to extend internal production and thereby extend the scope of the firm. The boundary of the firm, in terms of the composition of the actor, is similarly premised on control. Normally, for example, the board of directors, executive officers, or a single major shareholder will be the actor in a corporation. However, if the corporation becomes the subsidiary of a controlling parent, the real actor becomes that parent corporation. In such a case, the boundary of the firm moves upstream to reflect the new source of control. From the perspective of the parent, its own boundary moves downstream in the course of what is simply a case of vertical integration. The firm is now the combination of the subsidiary and its parent. The connecting factor is the singular control of the parent corporation over both production units.

The foregoing analysis pictures the firm as a control structure (production unit) bounded at either end by markets and subjected to their continuing regulation. This is a sufficiently detailed description with which to begin an analysis of the historical treatment of the nature of the firm in economic theory.

III. ECONOMIC THEORY

Neoclassical economic theory treated the firm as a production function, or "black box," operating in an environment of markets and state regulation.[1] The internal workings of the firm went unexplored.[2] The firm was simply a contained set of feasible production plans, the object of which was to maximize the welfare of the firm's owners. The theory did not attempt to explain how production within the firm was organized or what determined firm size. At the micro level, it was incomplete.

Although there were a number of earlier important efforts to describe the internal structure and processes of the firm,[3] much of the work (at least in terms of quantity) has been done in the last thirty years. Those efforts are reviewed and assessed here.[4] Specifically, those arguments that have attracted the most attention, or which are relatively new, are examined. They will be discussed in the order of their appearance, more or less, in the literature.

Before proceeding to this task, a number of observations are in order. The first has to do with the fact that a good portion of the literature explores the nature of the firm by examining the reasons for vertical (backward and forward), horizontal, and conglomerate integration. Such an approach is appropriate in this context because the integration or disintegration of separable production technologies determines the size of the firm. Integration occurs, presumably, because of some advantageous quality of the firm. Accordingly, analyzing the reasons for integration can suggest what is important about the firm. Integration, however, cannot be the exclusive analytical subject matter. Since many firms exploit only one technology or never integrate beyond their initial technology, the character of firms must be independent of integration *per se*.

A second and related observation is that much recent work tends to focus on transaction (exchange) cost explanations for the

---

1 See the brief summary of neoclassical theory by L. De Alessi, "Property Rights, Transaction Costs, and X-Efficiency: An Essay in Economic Theory" (1983) 73 Am. Econ. Rev. 64 at 64-65.

2 See the critique by H. Demsetz, "The Theory of the Firm Revisited" (1988) 4 J.L. Econ. & Organ. 141 at 142-44. P.J. McNulty, in "On the Nature and Theory of Economic Organization: The Role of the Firm Reconsidered" (1984) 16 Hist. Pol. Econ. 233, describes the gradual de-emphasis of the role of the firm in economic theory.

3 Adam Smith, Frank Knight, and Ronald Coase were early contributors. Coase assesses his original argument in "The Nature of the Firm: Meaning" (1988) 31 J.L. Econ. & Organ. 19.

4 See the review of developments by De Alessi, *supra* note 1 at 65-66.

phenomenon of integration. This has the effect of discounting other possible economic determinants of integration. Market power (monopoly and monopsony) considerations, for example, have been implicated in decisions to integrate.[5] While the matter is disputed, recent commentary has suggested that anti-competitive motives can be important.[6] More significantly, there may be production or technological reasons for integration. Thus, integration or disintegration may occur in order to ensure the supply of inputs, exploit comparative advantages, or exploit economies of scale and scope.[7] The reasons commentators offer for giving production motives only perfunctory treatment are diverse. Some writers suspect the independent significance of production determinants in comparison with the transaction cost approach.[8] Other writers accept the importance of production motives,[9] but, for example, claim to have no comparative advantage in their study[10] or assume their inclusion prior to their own analysis.[11] If, however, production-based reasons are important, as they appear to be, and if integration tells us something about the nature of the firm, any general theory of the firm must accommodate them in an explicit way.

The final preliminary observation is to note the importance of opportunism in economic theorizing about the nature of the firm. The literature analyzes the operation of opportunism at two points. The first may be called "exchange" opportunism because it occurs at the time of the exchange event. For example, when considering whether to use the market, an actor will consider the possibility that the other party will

---

5 The monopoly analysis is part of the neoclassical approach. See O.E. Williamson, "Assessing Contract" (1985) 1 J.L. Econ. & Organ. 177 at 188-90.

6 See M.K. Perry, "Vertical Integration: Determinants and Effects" in R. Schmalensee & R.D. Willig, eds., *Handbook of Industrial Organization*, vol. 1 (Amsterdam: Elsevier Science, 1989) 183; R.D. Blair & D.L. Kaserman, *Law and Economics of Vertical Integration and Control* (New York: Academic Press, 1983).

7 See F.M. Scherer & D. Ross, *Industrial Market Structure and Economic Performance*, 3d ed. (Boston: Houghton Mifflin, 1990) c. 4; and R.D. Buzzell, "Is Vertical Integration Profitable?" (1983) 61 Harv. Bus. Rev. 92.

8 O.E. Williamson, *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting* (New York: Free Press, 1985) at 86-87; and B.R. Holmström & J. Tirole, "The Theory of the Firm" in Schmalensee & Willig, *supra* note 6, 61 at 66.

9 M.H. Riordan & O.E. Williamson, "Asset Specificity and Economic Organization" (1985) 3 Int'l. J. Indus. Organ. 365 at 369.

10 S. Davies, "Vertical Integration" in R. Clarke & T. McGuiness, *The Economics of the Firm* (Oxford: Basil Blackwell, 1987) 83 at 86.

11 Perry, *supra* note 6 at 187.

hold out for a greater share of the surplus when the time comes to renew the contract. The second point in time when opportunism might occur is during the course of production and may be referred to as "production" opportunism. Workers (or co-venturers) may shirk, appropriate assets, or otherwise divert value away from their employer. Opportunities to act in this way arise because of the absence or inadequacy of monitoring or because of asymmetric information.

The traditional "agency" or "principal/agent" literature investigates generally the numerous circumstances in which opportunism (moral hazard) can be a problem.[12] It involves attempts to develop economic solutions to the problem, primarily through the design of contractual incentives that will induce the agent to maximize the welfare of the principal.[13] Other parts of the literature tend to focus more specifically on the firm as a response to the problem of either exchange or production opportunism. This work is identified by a number of redundant or overlapping labels. It includes the so-called property rights, incomplete contracts, transaction cost, new institutional economics, and new science of organization literatures.[14] It is to this material we now turn.

---

[12] See, for example, B. Holmström, "Moral Hazard and Observability" (1979) 10 Bell J. Econ. 74, and"Moral Hazard in Teams" (1982) 13 Bell J. Econ. 324. See also K.J. Arrow, "The Economics of Agency" in J.W. Pratt & R.J. Zeckhauser, *Principals and Agents: The Structure of Business* (Boston: Harvard Business School Press, 1985) 37.

[13] See O. Hart, "An Economist's Perspective on the Theory of the Firm" (1989) 89 Colum. L.R. 1757 at 1758-60.

[14] See the various references to, and descriptions of, these various literatures in E.F. Fama, "Agency Problems and the Theory of the Firm" (1980) 88 J. Pol. Econ. 288; J. Tirole, "The Multicontract Organization" (1988) 21 Can. J. Econ. 459; P. Milgrom & J. Roberts, "Economic Theories of the Firm: Past, Present and Future" (1988) 21 Can. J. Econ. 444; R.H. Coase, "The New Institutional Economics" (1984) 140 J. Inst'l. & Theor. Econ. 229; O.E. Williamson, "Reflections on the New Institutional Economics" (1985) 141 J. Inst'l. & Theor. Econ. 187; O.E. Williamson, "Economic Institutions: Spontaneous and Intentional Governance" (1991) 7 J.L. Econ. & Organ. (Special Issue) 159; and Hart, *ibid.*. Oliver E. Williamson describes his view of the ways in which agency theory and transaction cost economics differ in "Corporate Finance and Corporate Governance" (1988) 43 J. Fin. 567.

A. *The Transaction Cost Argument (Coase)*

Ronald Coase did not explain "the nature of the firm" in his 1937 paper;[15] rather, he assumed or asserted it.[16] He took as given the proposition that control is the essential defining characteristic of the firm[17] and then examined the circumstances in which such a structure would arise. Coase argued that "islands of conscious power"[18] (firms) arose in response to the costs of using the price mechanism.[19] He identified the price mechanism (the market) and the firm as "alternative methods of coordinating production."[20] After observing that there were costs associated with the use of each, he concluded that the price mechanism and the firm would replace each other depending on their relative cost. This conclusion, it will be appreciated, purports to tell us when integration will occur, it is an argument concerning what might be the motive for the use of a control structure.

It is possible to be more precise about the nature of the choice that is being made when, for whatever motive, "command" (the firm) replaces "exchange" (the market). Moreover, it is possible to do this without reliance on an anthropomorphic price mechanism to do the coordination work. It is obvious that markets do no actual work themselves. The price mechanism or market terminology is only shorthand for the process of exchange or negotiation where persons do the work of buying and selling (the allocation or coordination of resources).[21] Recall the example of the production of good Y where good X is an input. The actor who chooses not to produce good X will arrange an exchange to acquire it. Thereafter, because the simple acquisition of good X does not determine its subsequent use, the actor will direct good X into the good Y production process (rather than, for example, re-sell it). The actor who instead chooses to produce good X

---

15 R.H. Coase, "The Nature of the Firm" (1937) 4 Economica 386.

16 Coase describes the origins of the paper in "The Nature of the Firm: Origin" (1988) 4 J.L. Econ. & Organ. 3.

17 This is quite clear in the paper where, as one example, he asserts that within the firm "market transactions are eliminated and in place of the complicated market structure with exchange transactions is substituted the entrepreneur-co-ordinator, who directs production": *supra* note 15 at 388.

18 Coase borrowed this phrase from D.H. Robertson: see *ibid.*

19 *Ibid.*

20 *Ibid.*

21 *Ibid.* at 387. See also McNulty, *supra* note 2.

will also utilize the price mechanism when negotiating the purchase of the means to produce good X. Then, once good X is produced, the actor will similarly direct it into the good Y production process. Accordingly, in these respects, the price mechanism and the firm operate together. There is an acquisition followed by a transfer between productive stages in both cases. It will be observed that the only difference between the two procedures is in who controls the actual production of good X.

Whereas Coase saw a choice between command and negotiation, we see here a choice between internal and external control of production. The difference is one of perspective. Coase focussed on what occurs at the interface between the good X and good Y technologies. From that perspective, he saw either a command or a negotiation. He did not present those events as merely the respective manifestations of the choice the actor makes between internal and external production. If, however, the choice is one of internal versus external production (*e.g.*, the integration question), the interface focus is too narrow. The internalization of the exchange by another exchange (to become a directed transfer) is only an incident of, or the means by which to implement, the decision to internalize production. Coase's approach is useful only when the particular motive for integration operates at the point of his focus, the interface between independent production technologies.[22] Transaction cost motives are of this kind. However, since there are other motives for integration, it is necessary to expand the analytical perspective to include the production process.

The foregoing analysis implicates control over production (including transfers between production stages) as the defining characteristic of firms. The choice the actor makes determines who controls production. The actor will choose to produce when it is advantageous, relative to purchase, to do so. In this respect, the actor sees control as a solution to some problem associated with external production or as conferring some advantage over external production. It is a solution, for example, where a benefit is gained by the mere replacement of an exchange or because actual production is more efficiently carried out internally. In either case, it is the actor's control that defines the firm's utility.

---

22 Coase might respond that only that focus is relevant because only transaction costs determine whether firms exist. In fact, he has described his original article in those terms: "Transaction costs were used [in the article] to show that if they are not included in the analysis, the firm has no purpose": "The Nature of the Firm: Influence" (1988) 4 J.L. Econ. & Organ. 33 at 34.

Coase began his analysis by asking why there were firms at all given that markets could perform the coordination task. His answer, it has been noted, was that there were costs involved in the use of the market. The most "obvious" cost, in his view, was the cost of "discovering what the relevant prices are."[23] It is doubtful, however, that this is a cost that matters. The reason is that the identical cost exists for firms. Whether the choice is to purchase or produce, the actor will necessarily have to ascertain prices. The actor must determine either the pricing of good X or of the means to produce good X. Moreover, even the actor who chooses to produce will continue to monitor prices for good X in the future in order to set intra-firm transfer prices or to determine whether cost savings can be achieved by replacing internal with external production.

The costs of "negotiating and concluding a separate contract for each exchange transaction" was the second cost identified by Coase.[24] Again, however, the significance of this cost is unclear. The argument seems to be simply a quantitative one. Thus, instead of negotiating several serial contracts for external production, the actor can negotiate a single contract whereby the other party agrees to accept directions or instructions in the future. A saving would be obtained, ostensibly, because the costs of negotiating only one contract would be incurred. The reason this argument is suspect is that the costs are probably of the same order of magnitude whether or not the firm is used. Where the transactions involved are similar but repeated, the most likely candidates,[25] a standard form contract (*e.g.*, a purchase order and invoice) will be used and there will be no extended negotiations for contracts subsequent to the first. The costs of using standard forms are likely to be comparable to the costs of formulating, recording, and transmitting a command within the firm.

The third cost Coase identified was the cost of writing a long-term contract where the actor is unsure of what will be required of the other party in the future.[26] The problem the actor faces is production uncertainty. The cost of contracting is high, in such cases, because the contract would have to address the numerous contingencies that might arise over the term of its performance. This cost can be avoided,

---

[23] *Supra* note 15 at 390.

[24] *Ibid.* at 390-91.

[25] Dissimilar transactions imply different goods, and rare or occasional transactions do not amount to a significant cost.

[26] *Supra* note 15 at 391-92.

according to Coase, if the contract does not specify the details but simply gives the actor a general right, within certain limits, to direct the worker. The uncertainty initially faced by the actor is ameliorated at lower cost, presumably, than if a fully contingent contract had been negotiated. The problem with this argument is its identification of high transaction costs as the motive for integration. In fact, production uncertainty is the motive.[27] The actor is not prepared to make abstract decisions in a contextual vacuum and is therefore not interested in specifying today a response to every contingency that might arise tomorrow. More positively, the actor intentionally seeks flexibility precisely in order to respond to changes at the time they occur. It is a production motive at work here. The actor acquires control in order to deal with production uncertainty, not to avoid the unknowable cost of a fully contingent contract.

After Coase concluded that the introduction of a firm was due to the costs of using the market,[28] he asked why there were market transactions at all.[29] If firms had cost advantages, why was all production not carried on by one very large firm? His answer was the same as for markets: there were costs associated with the use of firms. He pointed to the rising costs "of organizing additional transactions within the firm," the entrepreneur's failure "to make the best use of the factors of production," and the increase in the price of supplies because "the 'other advantages' of a small firm were greater than those of a large firm."[30] Unfortunately, there is little elaboration of these costs by Coase.[31] He did indicate, however, that "the first two reasons given most probably correspond to the economists' phrase of 'diminishing returns to management.'"[32] Presumably, with respect to the first two reasons, Coase had in mind specific problems such as control loss in hierarchies and physical limitations on the ability of individuals to process

---

27 The actor experiences a condition of uncertainty and seeks to reduce that uncertainty by some appropriate measure. One possibility is to write a comprehensive contract; another is to purchase the right to control. There may be other solutions, each with its own particular costs. The point is that the motivation is production uncertainty, and comprehensive contracts and firms are simply alternative possible solutions.

28 Coase also mentions the market costs of differential government regulation, *e.g.*, taxation and quotas, but dismisses their significance in bringing firms into existence: *supra* note 15 at 393.

29 *Ibid.* at 394.

30 *Ibid.* at 395.

31 Consider the "individualistic spirit" argument cited with reference to the third cost, *ibid.* at 395 note 1, and Coase, *supra* note 3 at 32.

32 *Supra* note 15 at 395.

information.[33] If that is correct, it becomes apparent that the size of the firm is constrained by the limitations of its defining characteristic. That is, when control breaks down, the firm becomes costly, in relative terms, and will have no advantage over the market. This is an expected result if the firm is regarded as essentially a control structure. If that control fails because it is not transmitted (either up or down) through more than a few levels of hierarchy,[34] or because the actor is burdened with too little or too much information,[35] the potential benefits of exercising control cannot be realized. Firm size is thereby limited unless structural changes[36] or technological innovations[37] can partially relieve the control breakdown in some way.

Coase assumed that the firm is characterized by the control of an actor. In the course of the above analysis an attempt has been made to develop this proposition beyond its stark assertion. It has been shown how the control element generally determines whether the structure can be advantageous in given circumstances. Thus, acquiring control allows the actor to deal with production uncertainty in a flexible manner. Where it was not clearly advantageous, as in the cases of the first two market costs identified by Coase, there was reason to doubt that control was a solution or that there was even a solvable problem or reducible cost. As well, the optimum size of the firm was seen to depend in part on the failure of control processes. The analysis, in this way, amounts to a rehabilitation of Coase's view of what motivates the use of the firm, as well as an elaboration of his control assumption.

---

33 Coase alludes to this in his reference to increases in costs when the spatial distribution and heterogeneity of transactions increase: *ibid.* at 397.

34 See G.A. Calvo & S. Wellisz, "Supervision, Loss of Control, and the Optimum Size of the Firm" (1978) 86 J. Pol. Econ. 943; W.G. Ouchi, "The Transmission of Control Through Organizational Hierarchy" (1978) 21 Acad. Mgmt. J. 173; and O.E. Williamson, "Hierarchical Control and Optimum Firm Size" (1967) 75 J. Pol. Econ. 123. Note that Williamson later changes his views: *supra* note 8 at 134-35.

35 See the discussion of information considerations by Demsetz, *supra* note 2.

36 An example would be a change from a unitary to a multidivisional form of oganization. See O.E. Williamson, "Corporate Governance" (1984) 93 Yale L.J. 1197 at 1222-26 and *infra* note 56 at 85.

37 Examples of technological innovations in the communications area include the telegraph, telephone, and computer.

B. *The Monitor Argument*

Alchian and Demsetz denied that the firm could properly be characterized by the power to resolve disputes by fiat.[38] They believed this idea to be a "delusion" in relation to the employment of human capital.[39] They argued that an employer has no more control over employees than over independent contractors. The employer can direct, but the employees can refuse to accept that direction. The employer must therefore obtain their agreement to any proposal. The interaction of the parties is just another contract, like any other market contract (the "nexus of contracts" idea).[40] Notionally, the labour of employees is being purchased, as in the case of independent contractors, in a spot market. Alchian and Demsetz concluded that employment contracts (which Coase associated with the firm) were "not the essence of the organization we call the firm."[41]

As Alchian and Demsetz saw it, the internal organization of the firm is a response to the problem of shirking (production opportunism). Where joint production by a team is undertaken, the individual effort of each team member is difficult to evaluate solely on the basis of total output. This creates an incentive for individual members to shirk. The members recognize this, however, and therefore arrange *ex ante* for someone "to specialize as a monitor to check the input performance of team members."[42] To ensure that the monitoring function will be effective, they will give the monitor the right to instruct members and to alter the composition of the team (hire and fire). Finally, as an incentive to perform the monitoring function efficiently, the monitor will be assigned the residual claim to the team's profits. These and other rights[43] together create the internal contractual structure of the firm. According

---

38 A.A. Alchian & H. Demsetz, "Production, Information Costs, and Economic Organization" (1972) 62 Am. Econ. Rev. 777.

39 *Ibid.*

40 According to Alchian and Demsetz, "[t]o speak of managing, directing, or assigning workers to various tasks is a deceptive way of noting that the employer continually is involved in renegotiation of contracts on terms that must be acceptable to both parties": *ibid.* at 777.

41 *Ibid.*

42 *Ibid.* at 781.

43 Summarized *ibid.* at 783 and 794.

to Alchian and Demsetz, "the arrangement is simply a contractual structure subject to continuous renegotiation with the [monitor]."[44]

The limitations of the Alchian and Demsetz analysis have been described elsewhere.[45] Generally, it is thought to be constrained by its own terms. Apart from that, however, and notwithstanding their initial objection to such a characterization, Alchian and Demsetz seem only to reproduce the control conception of the firm. The utility they attribute to their monitor, or "central agent," seems very much to depend on the right to give instructions on "what to do and how to do it."[46] The function of their monitor, ultimately, is to direct production. The control of production opportunism is only one aspect of that function.[47]

It would appear, in this regard, that Alchian and Demsetz missed the important analytical fact when dismissing the view that the firm is characterized by the power to direct human capital. An initial observation is that employees do not refuse to accept direction. They do what is asked of them, and they do so for very good reasons.[48] More fundamentally, it is quite irrelevant that employees can, and sometimes do, refuse to obey. So long as they do follow instructions, the control structure will exist. While they remain employed, the firm will operate as an arrangement through which control is exercised over both physical and human capital. This will be the case even in the Alchian and Demsetz model. It is of no concern, at this point, that the structure may conceivably collapse in whole or in part.[49] Recognizing this indicates that the Alchian and Demsetz firm is more than a set of ordinary market contracts. It is a set of contracts that create a structure by subjecting

---

[44] *Ibid.* at 794. Demsetz subsequently notes that "[a]bating the cost of shirking helps explain the firm's inner organization but provides no rationale for the firm's existence": *supra* note 2 at 152.

[45] See Holmström and Tirole, *supra* note 8 at 66-74; Jensen and Meckling, *infra* note 50 at 310; and Demsetz, *ibid.*.

[46] *Supra* note 38 at 782.

[47] Identification of the control of production opportunism as the reason for the existence of firms is a significant difference between the Alchian and Demsetz model and those of Williamson and Hart who point to exchange opportunism.

[48] First, there are substantial costs involved in obtaining new employment, which employees will not incur simply to demonstrate their ultimate personal autonomy. Second, the value of their labour capital is often greatest to their current employer. Third, they make personal investments in co-workers and neighbours. For these and other reasons, workers tend to perform their employment contracts. Note, in this regard, that Demsetz subsequently appears to have accepted the durability of employment relationships: *supra* note 2 at 150.

[49] The Alchian and Demsetz criticism is not entirely without significance. It indicates another way in which control can fail. Workers may refuse to obey, for example, because of the personality or management style of their employer. If they do refuse, production will be disrupted.

assets to the control of a given actor. As we shall see, at least Demsetz appears to have moved towards this latter view.

C. *The Nexus Argument*

Jensen and Meckling, in a paper examining the agency cost tradeoffs between debt and equity capitalization, agreed with the Alchian and Demsetz objection to the view that firms are characterized by authority, and accepted their insistence on contracting as the proper emphasis.[50] They pursued this emphasis in their general abstraction of the "nexus" view of the firm:

> It is important to recognize that most organizations are simply *legal fictions which serve as a nexus for a set of contracting relationships among individuals. ...*
>
> Viewed this way, it makes little or no sense to try to distinguish those things which are "inside" the firm (or any other organization) from those things that are "outside" of it. There is in a very real sense only a multitude of complex relationships (i.e., contracts) between the legal fiction (the firm) and the owners of labour, material and capital inputs and the consumers of output.[51]

The nexus of contracts idea, in part for its supposed ideological content, has been elevated to the status of a theory of the firm in recent years. This has occurred in spite of the fact that it amounts to no more than an unhelpful preliminary analysis in the Alchian and Demsetz paper, and an undeveloped assertion in the Jensen and Meckling paper. Jensen and Meckling themselves admit that the idea "has little substantive content,"[52] and they essentially ignored the construct in their subsequent analysis.[53]

The difficulty with this idea appears in its assertion. It may be conceded that the firm is a "nexus of contracts," as it obviously is in the sense contemplated by Jensen and Meckling. However, this only returns us to a sort of "black box" notion of the firm where nothing "inside" the firm is seen to be of special significance. The main feature of the nexus idea is that the "firm" has no boundaries, that contractual ubiquity or

---

[50] M.C. Jensen & W.H. Meckling, "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure" (1976) 3 J.Fin. Econ. 305 at 310.

[51] *Ibid.* at 310-11.

[52] *Ibid.* at 311.

[53] The appeal of the construct to Jensen and Meckling may be its presumed congeniality to their subsequent analysis, where debt and equity are seen to be substitutable regimes for the regulation of agency costs. It allows them to formulate their approach without reference to the traditional view that debt is "outside" the firm while equity is "within."

indifference precludes theoretical fencing or partitioning. This notion is obviously at odds with the usual physical perception of the firm as a bounded structure. The difficulty again appears to be one of focus. Like Coase before them, Jensen and Meckling focused on exchanges at the interface of production technologies. As a result, when they perceive that exchanges are negotiated whether the actor chooses to produce or to purchase, their conclusion is exchange or contractual continuity. But this picture of contractual continuity obscures the real boundaries defined by the distinct production control of different actors. When that deeper control investigation is made, the institutional boundedness of the firm we perceive can be explained. Absent some formal or symbolic boundary, we demarcate production according to the hand which directs it. That is, the scope of the firm is congruent with the control a particular actor has acquired through contracts for production control.[54] In focusing solely on the exchange interface, Jensen and Meckling have failed to account for the institutions or structures that exchange can create. Ironically, while according primacy to contractual exchange, they failed to recognize the full extent of its power.

D. *The Transaction Cost Argument (Williamson)*

At the same time the nexus argument was being developed, Oliver Williamson was working to "operationalize" the Coasean view that transaction cost differences determined the choice between market and firm.[55] According to Williamson, "this entailed (1) identifying the microanalytic factors that are responsible for transaction cost differences among transactions, (2) aligning transactions with governance structures in a discriminating way, and (3) discovering and respecting the crucial intertemporal process features that predictably attend economic organization."[56] Williamson has in fact pursued this agenda with

---

[54] All contracts are exchanges and are fungible in that sense. However, the content of each contract must be examined. Some will involve purchases of finished goods, some will involve purchases of factors to be used in the production of other goods. The latter type of contract brings production within the firm when the actor proceeds to direct the combination of factors.

[55] See O.E. Williamson, "The Vertical Integration of Production: Market Failure Considerations" (1971) 61 Am. Econ. Rev. 112; and O.E. Williamson, "Markets and Hierarchies: Some Elementary Considerations" (1973) 63 Am. Econ. Rev. 316.

[56] O.E. Williamson, "The Logic of Economic Organization" (1988) 4 J.L. Econ. & Organ. 65 at 66.

broader strokes than Coase,[57] and his transaction cost economics is now a capacious analytical framework purporting to have application to a variety of diverse phenomena.[58] The approach, although not uncontested,[59] has been influential in the literature.

Williamson's approach begins with two behavioural assumptions. The first is that of bounded rationality, the notion that individuals are "intendedly rational, but only limitedly so."[60] The major consequence of this assumption is that complete contracts cannot be written. His second assumption is that, where credible commitments are lacking,[61] individuals are prone to opportunism, a condition "of self-interest seeking that contemplates guile."[62] Given these two assumptions, Williamson sees the issue as one of determining how to "organize economic activity so as to economize on bounded rationality while simultaneously safeguarding the transactions in question against the hazards of opportunism."[63]

Williamson operationalizes the transaction cost approach in the following way. A transaction can be carried out in the market, within the firm, or through hybrid structures. These "governance structures"[64] differ in their costs and adaptive capacities. As well, the transactions themselves have different dimensions. They may differ in the frequency with which they occur, the degree of uncertainty to which they are subject, and the degree of asset specificity attaching to them. Recognizing this, Williamson's approach is to "align transactions (which differ in their attributes) with governance structures (the costs and

---

57 Williamson, *supra* note 8. Williamson's breadth begins with his wide definition of transaction costs. See the objection by Demsetz, *supra* note 2 at 144-45.

58 These phenomena range from vertical integration to career marriages. See O.E. Williamson, "Transaction Cost Economics" in Schmalensee & Willig, *supra* note 6, 135.

59 Critical commentary includes Demsetz, *supra* note 2; Milgrom & Roberts, *infra* note 111; and G.K. Dow, "The Function of Authority in Transaction Cost Economics" (1987) 8 J. Econ. Behavior & Organ. 13, along with the response by O.E. Williamson in "Transaction Cost Economics: The Comparative Contracting Perspective" (1987) 8 J. Econ. Behavior & Organ. 617.

60 The phrase is Herbert A. Simon's in *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*, 2d ed. (New York: Free Press, 1961) xxiv.

61 See Williamson, *supra* note 8 at c. 7; and O.E. Williamson "Credible Commitments" (1983) 73 Am. Econ. Rev. 519.

62 Williamson, *supra* note 56 at 68. Williamson notes that "this self-interest seeking attribute is variously described as opportunism, moral hazard and agency": *supra* note 58 at 139.

63 Williamson, *supra* note 56 at 68.

64 The notion that serial contracting is a "governance structure" is perhaps counterintuitive. Williamson means governance in the wide sense of a scheme of interaction.

competencies of which differ) in a discriminating (mainly, transaction cost economizing) way."[65]

Williamson's first application of this approach was to vertical integration.[66] In his view, it is the paradigm case[67] and it is the application of particular interest here. Because bounded rationality makes it impossible to write a complete contract, some other device must be employed to allow for adaptation to changing conditions over time. The devices that might be used are the short-term contract or vertical integration.[68] Usually, because its high-powered incentives[69] constrain bureaucratic costs better, the short-term contract will be the efficient device for the acquisition of general-purpose assets.[70] However, as asset specificity[71] increases, a "fundamental transformation" occurs in which *ex ante* large numbers bidding is transformed *ex post* into bilateral bargaining.[72] Exchange opportunism becomes a problem and this limits the serviceability of the short term contract.[73] Internalization of the transaction (vertical integration) then becomes the more efficient choice.

The question at this point is how internalizing the transaction avoids or reduces the prospect of opportunism. It has been pointed out elsewhere that Williamson has not clearly explained the mechanism

---

[65] *Supra* note 56 at 73.

[66] Williamson, *supra* note 8 at c. 4-6; and *supra* note 58 at 150ff.

[67] Williamson, *supra* note 56 at 73.

[68] Williamson, *supra* note 58 at 150.

[69] According to Williamson, "by high-powered incentives, I have reference to residual claimant status whereby an agent, either by agreement or under the prevailing definition of property rights, appropriates a net revenue stream, the gross receipts and/or cost of which stream are influenced by the efforts expended by the economic agent": *supra* note 8 at 132.

[70] Williamson, *supra* note 58 at 151.

[71] Asset specificity is a key element in Williamson's approach and is explained in most of his publications. Generally, an asset is more "specific" to a use the less its value in alternative uses. He has identified several kinds of asset specificity: see O.E. Williamson, "Comparative Economic Organization: The Analysis of Discrete Structural Alternatives" (1991) 36 Admin. Sci. Q. 269 at 281-82.

[72] Williamson, *supra* note 8 at 61-63.

[73] See O.E. Williamson, "Transaction Cost Economics: The Governance of Contractual Relations" (1979) 22 J. Law & Econ. 233 at 241-42, 251; Williamson, *supra* note 8 at 76; and Williamson, *supra* note 58 at 151. See also B. Klein, R.G. Crawford & A.A. Alchian, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process" (1978) 21 J. Law & Econ. 297.

involved.[74] Williamson offers only vague references to internal adaptations being effected by fiat,[75] or carried out in a sequential way without the need to revise inter-firm contracts.[76] Nevertheless, a mechanism can be identified. It is implicit in the idea of a fundamental transformation. The exchange opportunism problem arises for short-term contracts because a highly specific investment creates a bilateral or small-numbers bargaining situation at the contract renewal stage.[77] Vertical integration can avoid this problem by internalizing the transaction and thereby moving the contracting interface backward (or forward) one or more stages where the new transaction does not involve a high degree of asset specificity. This returns the parties to a large-numbers bidding position where opportunism is attenuated or eradicated. Williamson's argument works, accordingly, because of the control nature of the firm. The control device provides a solution to the problem of exchange opportunism by eliminating the vulnerable exchange.[78]

Williamson's analysis of the limits on firm size is also broadly consistent with a control conception of the firm. Vertical integration does not proceed indefinitely, according to Williamson, because there are costs associated with increasing size.[79] He first describes the arguments that firm size is limited by uncertainty, control losses, growth, organizational capital, and the deadlines and delays of hierarchies.[80] He discounts these arguments on the ground that they do not take into account the possibility of selective intervention in the integrated unit.[81] He goes on to conclude, however, that selective intervention is not feasible. Interestingly, although he reaches this conclusion, he does not concede the significance of the arguments ostensibly challenged by the

---

74 See O. Hart, "An Economist's Perspective on the Theory of the Firm" (1989) 89 Colum. L.R. 1757 at 1763 and *infra* note 85 at 692-93.

75 Williamson, *supra* note 8 at 76.

76 See Williamson, *supra* note 73 at 253.

77 Exchange opportunism by a supplier is likely to occur only where the actor is, for some reason, unable to negotiate a contract identical to the original with some other supplier.

78 Williamson's view is that integration will tend to be the efficient solution only in cases where asset specificity is very high—where goods "become very close to unique": *supra* note 8 at 92.

79 *Ibid.* at 131-62.

80 *Ibid.* at 133-35.

81 *Ibid.*

possibility of selective intervention.[82] Instead, Williamson identifies certain other bureaucratic costs as constraining firm size.[83] He argues that, "[a]s compared with market organization, internal organization displays a differential propensity to manage complexity, to forgive error, and to engage in logrolling."[84] These costs, although again Williamson is not very clear about them, are at least partly due to the failure of control processes (*e.g.*, over-managing, rent-seeking). Where they appear to have no association with control failure, their significance remains unclear. Thus, to the extent Williamson has been able to demonstrate a limiting constraint, his analysis tends to support the idea of the firm as a control structure.

E. *The Residual Control Argument*

In an initial article co-authored with Sanford Grossman[85] and in several subsequent articles,[86] Oliver Hart argues that the firm is best understood in terms of residual control rights.[87] Hart defines a firm "to consist of those [physical] assets that it owns or over which it has control."[88] He equates ownership with control and asserts a division of contractual rights into "specific" control rights and "residual" control rights.[89] According to Hart, ownership of an asset is the condition of possessing its associated residual control rights. An owner, except to the extent that specific control rights are contracted away, has exclusive authority to determine the use and disposition of the asset. He defines vertical integration as "the purchase of the assets of a supplier (or of a

---

82 Presumably, if selective intervention is not feasible then such factors as control loss will operate.

83 Williamson states that he would be surprised "if the principal limits to vertical integration turn out to have nonbureaucratic origins": *supra* note 8 at 153.

84 *Ibid.* at 149.

85 S.J. Grossman & O. Hart, "The Costs and Benefits of Ownership: A Theory of Vertical Integration" (1986) 29 J. Law & Econ. 691.

86 See O. Hart & J. Moore, "Property Rights and the Nature of the Firm" (1990) 98 J. Pol. Econ. 1119.

87 Hart's earlier work in the area was in the agency tradition. See S.J. Grossman & O.D. Hart, "An Analysis of the Principal-Agent Problem" (1983) 51 Econometrica 7; and O.D. Hart, "The Market as an Incentive Scheme" (1983) 14 Bell J. Econ. 366.

88 Grossman & Hart, *supra* note 85 at 693.

89 *Ibid.* at 692.

purchaser) for the purpose of acquiring the residual rights of control."[90] Positioning these definitions in a world of incomplete contracting and opportunism, Hart's thesis is that residual control matters because it affects what happens in events not covered by the contract. Residual control rights, because they determine the usage of assets, "affect ex post bargaining power and the division of ex post surplus in a relationship" and therefore *ex ante* investment and effort decisions.[91]

Hart identifies opportunism as a primary influence on the decision to integrate.[92] The scope of conduct he includes in that notion, however, is wider than that included by Williamson. Hart does not require guile. His idea of opportunism is self-interested conduct. This is illustrated in his main example dealing with whether an insurance company or insurance agency will own the client list, or, in different terms, whether the list will be integrated by the company.[93] Where the company owns the list, in his analysis, the agency may suffer if the company decides that "it does not want to insure automobiles in a particular region" and so raises its prices, lowers the quality of its services, or changes the type and quality of its advertising in that region.[94] Alternatively, where the agency owns the list, it might encourage its clients "to switch to other companies if this seems advantageous," presumably because it will be more profitable to itself or its clients.[95] In neither case, it will be appreciated, is guile operating. The actor is simply doing what is expected given the profit-maximizing imperative. These are regular business decisions. Hart's assumption is that the parties will take the prospect of such conduct into account in their *ex ante* decision whether to integrate. In other words, they will seek protection from harmful business judgments of the other party, as well as from opportunism.

Hart and Williamson both argue that fear of holdouts (exchange opportunism) can be a motivation for integration. However, when Hart goes beyond that to include the possibility of adverse business judgments, he identifies a distinctly different motivation. Self-interest with guile is different from mere self-interest. In the former case, the question is whether the behaviour will occur. In the latter case, the

---

[90] *Ibid.* at 716.

[91] Hart, *supra* note 13 at 1766.

[92] Hart & Moore, *supra* note 86 at 1120.

[93] *Supra* note 85.

[94] *Ibid.* at 712.

[95] *Ibid.* at 713.

behaviour is expected. The probability of guile, moreover, is something which can be evaluated in the present as a function of character and reputation. The probability of adverse self-interest, on the other hand, depends on changes in the market in the future that alter the self-interested party's original incentives to contract. Finally, mere self-interest involves a *bona fide* refusal to contract on the former terms. Distinguishing these motivations, it should be apparent, has the effect of discounting an exclusive reliance on opportunism to explain integration. Although Hart might define the second type of behavior as opportunism,[96] he has actually generalized his approach. The *ex ante* allocation of residual control now depends on two factors: the possibility of opportunism (as traditionally understood), and the possibility of mere adverse self-interest. This implies that his approach can be further generalized, assuming it is a valid approach to begin with, to other motivations that might operate.

Hart's analysis, generally, was prompted by what he regarded as the failure of the existing literature to explain the mechanism through which vertical integration provides relief from the costs of opportunism.[97] He deals specifically with Williamson's proposition that the advantage of the firm is its resolution of disputes by fiat.[98] According to Hart, "Williamson does not spell out in precise terms the mechanism by which this reduction in opportunism occurs."[99] Hart's object is to suggest what that mechanism is. His argument is that the allocation of residual control determines relative bargaining power (which he links with the ability to act opportunistically) and therefore the division of surplus at the recontracting stage. The parties involved recognize this. Accordingly, where an asset is specific to one of them, the prospect of opportunism (by the other party) will motivate that party

---

[96] At that point, opportunism would take on the wide meaning of simply identifying and pursuing opportunities. Guile would be analytically irrelevant.

[97] *Supra* note 85 at 692-93; and *supra* note 13 at 1763.

[98] Williamson sees a number of benefits to integration and Hart acknowledges this elsewhere: see Hart, "Incomplete Contracts and the Theory of the Firm" (1988) 4 J.L. Econ. & Organ. 119 at 135.

[99] Hart, *supra* note 13 at 1763.

to integrate the asset.[100] The process here is the incentive effect *ex ante* of the *ex post* use of residual control rights to act opportunistically.[101]

Hart's analysis describes the operation of the incentive effect, it does not describe the mechanism through which control solves the problem. As observed earlier in the discussion of Williamson's approach, the mechanism involved is the avoidance of the opportunism problem by moving the contracting interface to a competitive (large-numbers) market.[102] This is the advantage of integration in relation to the problem of exchange opportunism in markets. Hart alludes to this view of the mechanism in one of his later papers but regards it, ambiguously, as [merely?] "consistent with the broad perspective provided by the notion of residual rights of control."[103] In terms of the foregoing analysis, the "consistency" is to be found in the serial connection between the mechanism and its incentive effect. Knowing that moving the contracting interface to a competitive market can reduce the prospect of exchange opportunism, the actor has an incentive to integrate.

Hart distinguishes his approach from others by its appeal to residual control rights over physical assets. This properly tends to discount the analytical significance of such things as ownership *per se* and the method of payment.[104] At the same time, however, it fails to account for the control of human capital except incidentally or in a derivative way.[105] This is problematic since casual observation suggests that control of workers is a significant aspect of many firms. Hart's analysis fails in this respect because of its very reliance on "residual" control.

Hart made a distinction between specific and residual control, and then chose the wrong one. "Residual" control, in an important

---

[100] According to Hart, "integration shifts the incentives for opportunistic and distortionary behaviour, but it does not remove these incentives": *supra* note 85 at 716. But this would not be the case if integration involves a shift to a competitive interface.

[101] Both the problem and its solution are conceived in terms of the incentive effects of control. An allocation of residual control rights that would create opportunities for holdouts would provide the wrong incentives to the other party. The solution is to make an allocation that avoids or restricts such incentives.

[102] See, above, text accompanying notes 77-78.

[103] Hart, *supra* note 98 at 136.

[104] Grossman & Hart, *supra* note 85 at 694-95.

[105] Hart continuously returns to this aspect of his approach: see *ibid.* at 717; *supra* note 13 at 1770-71; and *supra* note 86 at 1121 and 1150.

sense, is meaningless.[106] By definition, residual control is just the specific control one has not contracted away. Consider, in this regard, that one person's residual control will be the "specific" control purchased from a prior holder of residual control who retained some control rights. In such cases, in Hart's dichotomy, the same set of rights is contemporaneously "specific" and "residual" control. It would seem that "residual" control only has meaning if one person has rights that we might usually associate with "ownership;" that is, if we intellectually privilege one person's rights over those of another. This would be an implausible economic construction given that some "specific" rights (*e.g.*, the right to use a truck for five years) can have far greater economic significance than the "residual rights" (*e.g.*, the right to dispose of the truck at the end of the period.

An examination of the integration process also illustrates this. If exchange opportunism is potentially a problem, an actor can choose to produce the specific asset, for example, a unique tractor-trailer pressurized refrigeration unit to deliver a rare gas produced by the actor. When acquiring the means to produce this good, however, not all of the factors need be purchased. The tractor, the flatbed trailer on which the refrigeration tank is mounted, and the refrigerator itself may all be leased in competitive markets. The special technology required to regulate the pressure and temperature of the gas (a computer program) may be licensed from a software company which retains the right to sell it to others. In this stylized example, virtually the whole finished asset is created out of a combination of "specific" control rights. Specific control, in the case of each of these factors, is all the actor requires in order to move the contractual interface to competitive markets and thereby avoid opportunism. The actor has solved the problem through the *ex ante* allocation of "specific" control rights. Even if these factors had been purchased outright, the problem would still have been solved by specific rights since only the right to use the factor was actually employed.

The proposition that "specific" control is the relevant analytical concept, unlike Hart's approach, accommodates human capital. Workers, like lessors and licensors, sell only some of their specific rights. They grant their employer the right, within defined limits, to direct their labour. To the extent they do this, their labour capital is integrated. They retain their "residual" (other) control rights to employ their labour

---

106 Hart's use of the "residual" label may be distinguished from the the common understanding of a "residual claim." In the latter case, the "residual" label refers to the variability of the claim. The control characterized as residual control is definite, not variable.

capital as they see fit (*e.g.*, to read, garden, stay out late). Their employer does not require these other specific control rights and does not integrate them into the production unit.

There is a consideration involved in the integration of human capital, however, that does not arise in the case of physical capital. It is the fact that employers cannot hold on to the labour capital of workers.[107] An actor cannot enforce a specific right to determine when a worker's employment will end. This constraint, it will be appreciated, is not inherent in the economic relationship given that, at one time, an actor could enforce a specific right to insist on the continuing labour of the worker (*i.e.*, slavery). Rather, the constraint is legally imposed as a matter of public policy. This, it must be emphasized, does not discount the utility of a control analysis in relation to human capital. Instead, it confirms it. If workers cannot be compelled to work, the control analysis would predict that there will be holdout problems with labour generally. This is what appears to happen in reality. Thus, the analysis indicates that control (the firm) does not solve the problem of human capital exchange opportunism (it can not, as a matter of law), and this corresponds with the fact that it remains a problem within firms. Devices other than control (*e.g.*, compensation design) must be employed.[108]

The value of a control analysis is demonstrated another way in the context of human capital. Both Williamson and Hart associate the prospect of exchange opportunism with a condition of high asset specificity.[109] However, the great majority of workers do not exhibit high asset specificity and yet are controlled by their employers. This tends to suggest that the significance of asset specificity is overstated. If that is so, exchange opportunism based on high asset specificity is brought into question as a motivation for integration.[110] A general analysis of control rights, however, allows for this. Control is a device which can be used advantageously for a variety of purposes. If exchange opportunism is suspect as a primary motivation, there may be some other benefit to integrating human capital. One possibility is that control over workers is required in order to transmit the value of an actor's comparative advantage to actual production. The actor knows the most efficient way

---

[107] This is Hart's reason for not including control of human capital directly within his analysis: see references *supra* note 105.

[108] Other devices will also be ineffective to control exchange opportunism. Monitoring (in combination with the dismissal sanction), for example, is primarily a device to control production opportunism.

[109] Williamson explicitly develops the notion of human asset specificity: *supra* note 71.

[110] Consider the views of Klein *et al.*, *supra* note 73 at 313-16.

to carry on production (what, where, when, and how) and maintains this efficiency by directing the workers accordingly. There may be other control motivations. The point is that the notion of the firm as a polyfunctional control structure expands the analytical framework in a useful way. There can be situational explanations for the integration of human capital.

F. *The Bargaining and Influence Cost Argument*

Milgrom and Roberts accept the transaction cost approach to the study of economic organization.[111] They believe, however, that it suffers from two conceptual problems. The first is its determination of the total costs of a firm by the simplistic summation of production and transaction costs. Milgrom and Roberts point out that, for example, increased production costs may sometimes be incurred in order to reduce transaction costs. They therefore insist that production and transaction costs be considered together.[112] The second conceptual problem, in their view, is that "the general theory is too vague to be useful."[113] In particular, the predictive power of transaction cost economics is limited because there has been little elaboration of what the relevant transaction costs are.[114]

Milgrom and Roberts purport to enrich transaction cost analysis in two ways.[115] They first consider the costs of using the market to carry out transactions. An initial formal analysis leads them to conclude that the "efficiency of market arrangements is limited only by the costs of negotiating efficient short-term contracts."[116] This conclusion, they state, "points to the central importance of bargaining costs in determining the efficiency of market transactions."[117] Asserting the

---

111 P. Milgrom & J. Roberts, "Bargaining Costs, Influence Costs and the Organization of Economic Activity" in J.E. Alt & K.A. Shepsle, eds., *Perspectives on Positive Political Economy* (Cambridge: Canbridge University Press, 1990) 57.

112 *Ibid.* at 57-58.

113 *Ibid.* at 58.

114 *Ibid.* at 70.

115 Their analysis, it would appear, is directed solely towards the second conceptual problem they identified.

116 Milgrom & Roberts, *supra* note 111 at 69.

117 *Ibid.* According to Milgrom and Roberts, this "accentuation of short-term bargaining costs contrasts with received theory (as presented by Williamson), which emphasizes asset specificity, uncertainty, and frequency of dealings as the key factors": *ibid.* at 58.

centrality of bargaining costs, however, only begins their analysis. To be able to make specific predictions with the transaction cost approach, it is necessary to identify the various types of bargaining costs and how they vary with different circumstances.[118] While Milgrom and Roberts define bargaining costs expansively,[119] they restrict their own analysis to the costs of delays and failures to reach agreement.[120] They identify three costs: coordination failures, measurement activity, and undisclosed preferences. Coordination failures occur when the complexity of a market transaction prevents easy agreement between the parties. The absence of an obvious "focal point"[121] makes the coordination of their demands difficult and leads to inefficient haggling or a failure to agree.[122] Measurement costs are the costs of measuring or evaluating a good (*e.g.*, its quality) prior to its purchase in the market.[123] The third bargaining cost arises where parties strategically misrepresent their actual preference for or valuation of a good.[124] Milgrom and Roberts argue that these costs are wasteful in many cases and, consequently, a short-term contract might not be the most efficient way to arrange the transaction. Other arrangements, such as vertical integration, may economize on these costs.[125]

Milgrom and Roberts argue that "the crucial distinguishing characteristic of a firm is not the pattern of asset ownership but the substitution of centralized authority for the relatively unfettered negotiations that characterize market transactions."[126] That is, they understand the firm to be a device which solves problems through the

---

[118] *Ibid.* at 61.

[119] They write, *ibid.* at 65:

> We interpret 'bargaining costs' expansively, just as we did the term 'transaction costs,' to include all the costs associated with multilateral bargaining, competitive bidding, and other voluntary mechanisms for determining a mutually acceptable agreement. Bargaining costs include not only the wages paid to the bargainers or the opportunity costs of their time, but also the costs of monitoring and enforcing the agreement and any losses from failure to reach the most efficient agreement possible in the most efficient fashion.

[120] *Ibid.* at 72.

[121] Kreps, *infra* note 134 at 121, describes a focal point as "some principle or rule individuals use naturally to select a mode of behavior in a situation with many possible equilibrium behaviors."

[122] Milgrom & Roberts, *supra* note 111 at 72-75.

[123] *Ibid.* at 75-77.

[124] *Ibid.* at 77.

[125] *Ibid.*

[126] *Ibid.* at 72. See also 79.

application of the control of the actor. It does this, in the case of coordination failures, by eliminating the source of the problem. It is simply no longer necessary for two parties to reach an agreement (coordinate their demands). The actor alone deploys the assets.[127] The problems of measurement and undisclosed preferences are also ameliorated or eliminated because a single actor now acts on both sides of the transaction. The buyer and seller of the market are replaced by the single actor of the firm. The bargaining costs of the market are avoided by internalizing the transaction, making it subject to the control of the actor.

Milgrom and Roberts next turn their attention to the costs of the firm. Having analyzed how the costs of bargaining make markets the less efficient choice in some cases, they then consider how the costs of centralized authority can make the firm an unattractive alternative. They seek to explain, at this point, what limits the size of the firm. Milgrom and Roberts identify two kinds of costs that accompany increases in discretionary authority. The first kind "arises because those with discretionary authority may misuse it."[128] In fact, there are actually two kinds of costs contemplated in this statement: that of over-managing, and the quite distinct cost of production opportunism.[129] Milgrom and Roberts attribute these costs to flaws in the incentives, intelligence, or character of those possessing discretionary authority.[130] The second type of cost Milgrom and Roberts associate with centralized authority arises "even when the central authority is both incorruptible and intelligent enough not to interfere in operations without good reason."[131] These are "influence costs,"[132] and they "arise first because individuals and groups within the organization expend time, effort, and ingenuity in attempting to affect others' decisions to their benefit and secondly because inefficient decisions result either directly from these influence activities or, less directly, from attempts to prevent or control

---

127 *Ibid.* at 75.

128 *Ibid.* at 79.

129 *Ibid.* at 79-80.

130 *Ibid.* at 80.

131 *Ibid.*

132 See P. Milgrom & J. Roberts, "An Economic Approach to Influence Activities in Organizations" (1988) 94 Am. J. Soc. (Supp.) S 154; and P. Milgrom, "Employment Contracts, Influence Activities, and Efficient Organization Design" (1988) 96 J. Pol. Econ. 42.

them."[133] All of these costs, it should be apparent, represent failures in the control process. Control is inefficient if it is excessive. It is also inefficient if applied opportunistically, by those who have it, for personal gain. Finally, it is inefficient where it is misdirected as a result of the influence activity of interested parties. The costs of centralized authority, accordingly, are the costs of control failure.

The rehabilitation of transaction cost economics proposed by Milgrom and Roberts buttresses the view that both the advantages and disadvantages of a firm are directly attributable to the operation of control. A control structure can be the efficient arrangement where the costs of using the market are high because it reduces or avoids these costs. Conversely, control itself has limitations which restrict its utilization and, hence, the size of the firm.

G. *The Reputation/Corporate Culture Argument*

Corporate culture, according to Kreps, can usefully be understood in economic terms.[134] Kreps turns initially to non-cooperative game theory to develop the notion of the firm as a reputation bearer.[135] A firm has an interest in maintaining an open and unambiguous reputation about the way in which it exercises authority. This is because reputation is important for transactions that will encounter unforeseen contingencies. Where these contingencies can occur, a potential contracting party will want to have some idea *ex ante* of how the firm will respond. An agreement will be made with the firm that exhibits a reputation the content of which corresponds with the needs of the contracting party.

Acquiring a reputation about how the firm will respond to unforeseen contingencies involves choosing a decision rule or principle and then following it. Kreps doubts that truly unambiguous and universal rules exist. However, the literature on focal points indicates to him that sufficiently unambiguous rules are possible.[136] What is

---

[133] Milgrom & Roberts, *supra* note 111 at 80. Attempts to influence decisions can, of course, be productive. Milgrom and Roberts have in mind influence activities that are directed towards the realization of personal benefit without regard to firm benefit.

[134] D.M. Kreps, "Corporate Culture and Economic Theory" in Alt & Shepsle, *supra* note 111, 90.

[135] For other discussions of the role of reputation, see Holmström & Tirole, *supra* note 8 at 76-78; and Milgrom & Roberts, *supra* note 14 at 453.

[136] Kreps, *supra* note 134 at 120-123 and 125.

needed is a principle that "permits relatively efficient transactions to take place and on which a viable reputation can be based."[137] Having selected such a principle, the task of the firm is to communicate that principle to potential contracting parties. This is the role of corporate culture. Kreps identifies corporate culture "with the principle and with the means by which it is communicated."[138] Corporate culture gives all parties "an idea ex ante how the organization will react to circumstances as they arise; in a strong sense, it gives identity to the organization."[139]

The functions performed by corporate culture appear to be concerned with delineating and maintaining the control processes of the firm. As we have seen, disclosure to potential contracting parties of the basis on which decisions will be made in the future is a primary function of corporate culture. Kreps also gives it a role in conveying the principle to those (managers and workers) who undertake its actual application.[140] This leads to consistency in decisions and avoids deterioration of the principle itself. As well, by identifying proper performance where there is an unforeseen contingency, corporate culture is a standard by which the performance of those who exercise authority can be measured.[141] Accordingly, in these respects, the corporate culture approach is necessarily concerned with the question of control.

H. *The Information Cost Argument*

Although the Alchian and Demsetz "nexus" idea is often referred to, its analytical utility has never been demonstrated in any convincing way.[142] Demsetz himself attributes this to the fact that "[t]he defining content of the nexus of contracts [argument] remains rather vague in literature on the theory of the firm."[143] Recognizing this, but finding the transaction cost and agency approaches inadequate in certain respects,[144] Demsetz has recently sought to rehabilitate the content of

---

137 *Ibid.* at 125.

138 *Ibid.* at 126.

139 *Ibid.*

140 *Ibid.*

141 *Ibid.*

142 See, text accompanying notes 38-54.

143 *Supra* note 2 at 154-55.

144 *Ibid.* at 144-54.

the nexus idea in terms of information costs.[145] His goal is to demonstrate "that information cost has relevance that extends beyond its significance in transaction cost and moral hazard problems."[146]

Demsetz asks the question, "When is a nexus of contracts more *firm-like*?"[147] In his view, "specialization, continuity of association, and reliance on direction are characteristics of firm-like coordination."[148] Specialization arises as a way to economize on the costs of producing, maintaining, and using information. The benefit of specialization is realized when others use it without incurring the costs of educating themselves as to the information on which it is based. This occurs when the specialist communicates by giving directions. Those who are to make productive use of the information, but who are not specialists, simply follow the directions.[149] As Demsetz puts it, "direction substitutes for education."[150] Demsetz also explains vertical integration in terms of information costs. A firm will cease to vertically integrate at that point where the increasing costs of information acquisition and maintenance make the exploitation of information (the giving of directions) infeasible.

This analysis puts considerable distance between Demsetz and the original nexus idea. In asking when a nexus is more "firm-like," he is contemplating, or moving very close to, the notion of the firm as a bounded structure. Moreover, the arrangement he describes is explicitly a control structure. Specialization *per se* is not a necessary attribute of the arrangement. Continuity of association is also unnecessary, given that directed persons can be employed in either long-term or short-term relationships. This leaves only the giving of directions on an ongoing basis as the essential feature of arrangements that are more "firm-like." In the particular context of information costs, this control feature proves to be advantageous. The mechanism of its operation is the substitution

---

145 Other works on the economics of information include J.E. Stiglitz, "Incentive, Risk, and Information: Notes Towards a Theory of Hierarchy" (1975) 6 Bell J. Econ. 552; and K.J. Arrow, "Informational Structure of the Firm" (1985) 75 Am. Econ. Rev. 303.

146 *Supra* note 2 at 154.

147 *Ibid.* at 155 [emphasis in original].

148 *Ibid.* at 156.

149 Consider, in this regard, the skilled worker (specialist) who is given directions. The Demsetz argument works here as well. The worker may be viewed as a specialized product that can be manipulated or instructed by the employer without the employer first acquiring the knowledge possessed by the worker. The worker is a product that requires less information to use than to produce. See the discussion of products by Demsetz: *ibid.* at 158.

150 *Ibid.* at 157.

of direction for education so as to economize on the acquisition and maintenance of information.

IV. TRANSACTION COST CONSIDERATIONS

The foregoing arguments fall loosely into two general categories. The monitor, residual control, reputation, and information cost arguments are primarily grounded in production considerations. The analyses of Coase, Williamson, and Milgrom and Roberts, on the other hand, initially focus on transaction or bargaining considerations. Thus, the two groups of analysts locate elementary significance in different processes, in either production or exchange. No theoretical contest, however, is necessarily implied by this dichotomy. The firm may be perceived, and then defined, either in terms of its internal structure or by reference to its boundaries. In the latter case, firm boundaries may be defined either by the furthermost projection of the actor's control or, alternatively, by the presence of exchange (negotiation). The two phenomena operate on opposite sides of the boundary line and either could therefore be regarded as definitive of that boundary. Recognizing this makes it possible to understand how some economists have come to regard exchange as characteristic of firms (*e.g.*, as determinative of firm size). At the same time, it should be apparent that an exchange is not part of the firm but, rather, is an event found in the space between firms. The presence of exchange is but a proxy for the absence of an actor's control over production. As such, it is limited to derivative definition of the firm. The nature of the firm itself is found in its internal structure, in the application of control to the employment of assets.

While production and exchange are distinct processes, they are also connected processes. Factors of production are acquired, and product is distributed, through exchange. In addition, the size and type of production may be altered by exchange. A full comprehension of the overall economic structure of the firm therefore involves linking these two processes together in a model in which firms are understood as control structures separated from other control structures (other firms) by the need to negotiate (exchange). Before proceeding to construct this model, however, it is necessary to briefly consider the plausible scope of the "transaction cost" approach.

There is a tendency among some economists to regard transaction cost economics as exclusively important in the analysis of the nature of the firm. This is seen in the occasional assertion that firms would not exist but for transaction costs. The assertion is obviously

wrong if it is suggesting that markets, even perfect markets, can accomplish production on their own. Markets are unable to achieve, but necessarily require, the production of goods. Markets would only function without production if a finite set of pre-existing goods were constantly circulated (bought and sold) without in any way being altered. That is not what happens.

The difficulty with the assertion may be traced to the man to whom it is usually attributed. Coase regarded the firm and the market as alternative ways of allocating resources. He asked why an actor would ever choose not to use the market. His reason was that there were costs associated with the use of markets. A firm would therefore "arise," as Coase put it, when its costs were less than the costs of the market. This is the source of the subsequent confusion. It will be appreciated that the firm only "arises" in relation to this particular actor. Prior to its internalization by the actor, production had been carried on by some other actor. That is, a firm existed, albeit not the firm of the particular actor. In effect, there has been a shift of production from one actor to another. Thus, although, in one sense a firm "arises," it is clearly not correct to assert that there would be no firms in the absence of transaction costs.

The assertion might nevertheless be thought to apply in the integration context. The assumption involved would be that firms are only *combinations* of technologies. Single technologies controlled by single actors would not be firms. Instead, they would be mere factors of production. The argument would be that integration beyond a single technology would never occur if transaction costs were zero. It would always be inefficient to combine technologies in a "firm." Hence, the firm would never arise. The difficulty with this argument is that it depends on the population of actors having constant abilities. It does not account for the fact that different actors have different abilities. Some actors will possess a comparative advantage over others and this advantage may extend to a number of production technologies. For example, an actor may have a comparative advantage in the production of both good X and good Y. Because of that advantage, the actor will choose to produce both goods. Thus, the combination of technologies will come about (a "firm" will exist) even in the absence of transaction costs.

There is another difficulty in regarding transaction cost economics as a comprehensive or exclusive analytical framework. There are two specific questions that most commentators address when investigating the economic nature of the firm. They are: 1) why do firms exist, and 2) what determines the size of the firm? Transaction cost

proponents have a formally different answer for each question. Their first answer is that firms exist because of transaction costs. Their explanation for the size of firms, however, does not initially involve transaction costs. Instead, they rely on what are primarily production considerations. Coase identified diminishing returns to management, Williamson identified bureaucratic costs, and Milgrom and Roberts identified the costs of discretionary authority. Incurred in the course of production, these costs limit the size of the firm when they exceed the net transaction costs that otherwise would be expended. In this respect, because it must rely on a production analysis, transaction cost economics cannot claim exclusivity for its approach.

Perspective is also a problem for the transaction cost argument. Making the "transaction" the basic unit of analysis concentrates attention at the negotiation interface. The focus is on interactions *between* firms and there is little consideration of actual production *within* firms. This again is a black-box conception of the firm. Production processes are treated as insignificant, uninteresting, or secondary considerations and they tend to remain undeveloped within the framework of the approach. In effect, transaction cost analysis stops short of the firm. The major consequence of this is to continue to obscure the mechanisms through which the firm proves to be an advantageous device in a variety of circumstances.

One other observation may be made in connection with transaction cost theorizing. Throughout the foregoing discussion, there has been an insistence on a distinction between production and exchange. Others might view this distinction as a semantic one, believing that both exchange and production costs are contemplated in the idea of transaction costs. The issue is indeed partly semantic, but not in the trivial sense assumed. Language can clarify or obscure its subject-matter. Where there is no material difference between two phenomena, it serves no purpose to continue to identify them individually. On the other hand, where there is a substantive difference, their linguistic conflation obscures or submerges that difference, hindering its comprehension, examination, and communication to others. The difference between production and exchange is substantive. It is the difference between directing and agreeing—two fundamentally distinct (but connected) processes. The claim that these are theoretically, and therefore linguistically, equivalent "transactions" (the nexus of contracts idea) only submerges what, in the present context, is analytically significant. A firm begins and ends with the control a given actor is entitled to exercise. An exchange is an interstitial event, it is not *within* either of the exchanging firms. While each actor will direct assets in the

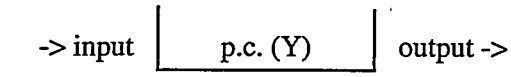
process of exchanging, there can be no exchange without the agreement of the other actor. Accordingly, the analytical distinction, and its linguistic recognition, are required.

V. THE PRODUCTION UNIT MODEL

Transaction cost considerations do not adequately explain the firm. It is necessary to include production considerations in any model purporting to have general theoretical application. Selecting the production unit, rather than the transaction, as the basic analytical construct allows for the accommodation of both production and transaction costs and clarifies the operation of the different types of integration. The factor that links these various elements is the control actors exercise over the employment of assets.

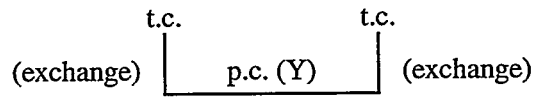
The simplest production unit is the single individual who expends effort in producing a good (whether a service or a tangible or intangible asset). Each such production unit is composed of two operating processes. The first is the actual production process; for example, the manufacture of good Y. The production unit and its associated production costs (p.c. (Y)) can be represented in the following schematic way.

Figure 1



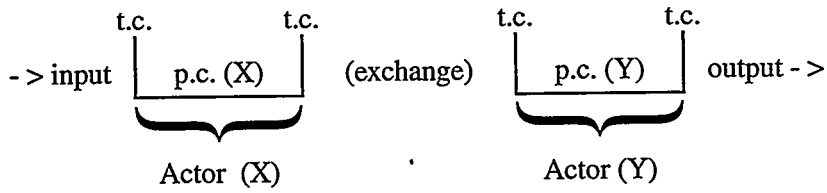
The second process associated with every production unit is the direction of exchange. An actor will employ assets to prepare for and execute exchanges with other actors. At the anterior end of the production process, the actor will incur transaction costs to acquire inputs (the procurement function). Then, once production is complete, the actor will incur transaction costs in the course of selling the output (the marketing function).

Figure 2



In a completely reduced (fully atomistic) economy, every firm would be of this type (a sole proprietor or independent contractor). Each firm would employ a single *production* technology and a single (or dual) *exchange* technology in the course of interacting with other firms or ultimate consumers. The exchange interaction between two firms would be represented by distinct production units connected by an exchange. An example would be the procurement of good X. The actor who produces good X will incur a separate set of production and transaction costs.

Figure 3



A given production unit or firm will succeed or fail depending on the efficiency of the control of the particular actor relative to all other actors producing the same good. That is, the firm will succeed when the actor has a "comparative advantage," or no comparative disadvantage, relative to others. This advantage may be physical (*e.g.*, a location advantage) but more commonly will be found in the particular attributes of each actor. Thus, the actor (*e.g.*, an individual or a board of directors) may have greater dexterity or a greater physical capacity for work. Or the actor may have a comparative advantage in the cognition, interpretation, manipulation, or application of information. These and other attributes are applied in the course of the actor's effort or control to affect both production costs and the two sets of transaction (procurement and marketing) costs. The actor may be a better producer, as well as negotiator, or the advantage in one capacity may exceed the disadvantage in the other.

There are time and information constraints on what an individual actor can achieve. An actor may therefore employ human capital (along with the requisite physical capital) to overcome these constraints. The benefit of employing others is found in the direction or control which guides their work. There may be a benefit, for example, in a certain breadth of control physically or temporally unobtainable by the actor alone. Economies of scale will be realized when the actor horizontally expands control (through employment contracts) to a production capacity that achieves the lowest average cost. These economies may be obtainable, or approachable, in the case of both production and transaction costs.

Even where economies of scale are not available (*e.g.*, where returns are constant) an actor can benefit from the control of others. According to Demsetz, direction reduces information costs through its substitution for education. This is an observation about how control may be valuable when extended beyond the physical capacity of the actor. It is a specific manifestation of the fact that control is a means to exploit an actor's comparative advantage by its transmission to production through human capital. Another possibility is that control allows an actor to exploit the comparative advantage of others, usually workers with skills or expertise not possessed by the actor. Again, these possibilities would potentially apply to both production costs and the two sets of transaction costs. For example, constant returns to scale may be realized in the transmission of an actor's superior information-processing skills to production through several workers (educated or not) employed to assist in the production of good Y. Additional value may be extracted by employing a worker with a comparative advantage (relative to the actor) in the marketing of good Y. Where the employment of human capital is the norm, a further benefit may be realized if the actor has a comparative advantage, relative to other actors, in managing workers. The actor may possess a superior ability to instruct, schedule, motivate, understand, or communicate with workers. Again, this advantage would potentially extend to both production and transaction costs.

The constraints on an individual actor could also be overcome by combination with other actors. That is, two or more individuals may act together to control a particular production unit (*e.g.*, a partnership). In terms of the present analysis, the joint control of these individuals identifies them as a single actor. Economies of scale and comparative advantages can be exploited in this way, but the submission of each individual to collective control creates for each a worker capacity distinct from their controller capacity. These capacities are properly treated as analytically separate. There is an actor (the group) that controls

production through human capital (the individuals). Variations in the control rights of particular individuals (*e.g.*, where one partner has exclusive control on a specific issue) can be treated in the same way.

The employment of additional human and physical capital (the extension of the actor's control) will not proceed indefinitely. This is because control itself has limitations. There have been several suggestions as to why control fails or becomes inefficient as the scale of production increases or becomes more complex. Among these are control loss, managerial diseconomies, bureaucratic costs, influence costs, and production opportunism. There have also been a number of suggestions as to how control failure has been or could be ameliorated to some extent. Structural changes, technological innovations, or compensation design may expand the useful scope of control. From a different perspective, Kreps sees "corporate culture" as a means to reduce control failure by ensuring that the exercise of control is uniform throughout an organization. These and other factors operate to determine the upper limit of the useful application of control.

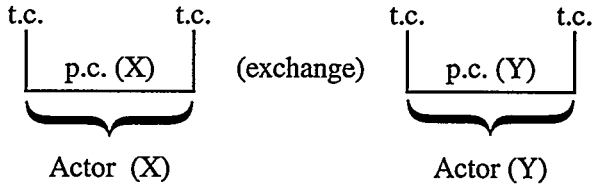
A number of general observations may be added at this juncture to clarify this production unit picture of the firm. The first is that, except in the integration sense, markets and firms are not *alternative* governance regimes. Rather, they are complementary or dependent elements in a regime of decentralized production. Exchange is necessarily interposed between the production of different actors, and each production unit has both production and transaction capacities. The second observation is that the status of a given production unit depends on whether it is subject to external control. A single individual employing a single technology will be a "firm" when that individual alone controls the technology. That same individual/technology will cease to be a separate "firm" when it becomes subject to the direction of another actor. It will then only amount to an additional factor of production in the firm of the controlling actor. The final observation is one of fundamental significance. It is that a firm can exist with or without hierarchy. A single individual or group of co-equals may control assets without the assistance of additional workers. The control that defines the firm is therefore not synonymous or congruent with the notion of hierarchy. The employment of hierarchical relations is a common, but not immanent, feature of the firm. The firm derives its essential character from the asset control of the actor, not the enlistment or subjugation of human capital.

To this point in the discussion we have considered only the case of a single production technology (along with its horizontal expansion). We proceed now to extend the analysis to incorporate those instances

where two or more production technologies are combined under the control of a particular actor. These are the cases of vertical and conglomerate integration.

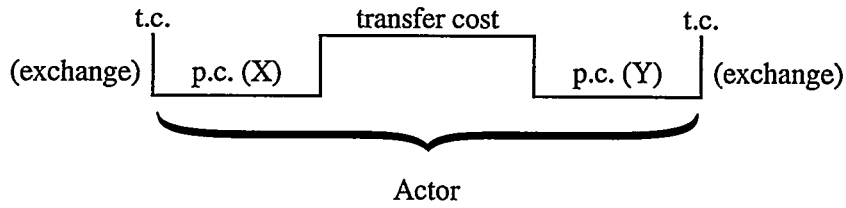
Where an actor chooses to produce only good Y, it will be necessary to purchase good X (the input) in the market. The actor will therefore engage in an exchange with a firm that produces good X. Each production unit will be controlled by a different actor.

Figure 4



If the actor instead chooses to produce both good X and good Y (*i.e.*, vertically integrate), a good X production unit will be acquired by merger or internal expansion and combined with the good Y production unit. The costs of making an exchange at the interface between good X and good Y will be eliminated, leaving instead the costs the actor incurs in transferring good X into the good Y production process.

Figure 5

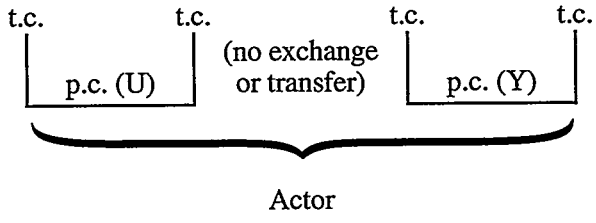


It will be observed that the vertical integration of good X involves the creation of a control bridge between two technologically separable production units. That is, when two or more successive technologies are combined, the control of their common actor represents a bridge between them.

An actor may alternatively choose to produce an unrelated good (good U) in addition to good Y. This is the case of conglomerate integration. Here there is neither an exchange nor a transfer relationship between the combined production units. The actor's

control, however, again creates a bridge, this time between technologically unrelated production units.

Figure 6



We will examine the motives for the various types of integration below. Initially, however, it is worth focussing on the fact that conglomerate integration does not involve the replacement of an exchange. The production units are not technologically related in a successive way. There is no exchange or transfer interface between them. Accordingly, transaction cost economics cannot explain this type of integration in the Coasean sense of fiat replacing the price mechanism where it is less costly. The explanation for conglomerate integration resides elsewhere. It becomes conspicuous, as do the explanations for other types of integration, when integration is conceived generally in terms of the combination of production units.

Consider first the control bridge established upon a vertical integration. There are several explanations for why an actor would choose to create this bridge. Some of these explanations implicate the bridge itself as the source of the benefit to the actor. Other explanations locate the source of the benefit in the manipulation of the production and exchange technologies of the acquired production unit rather than the bridge *per se* and, for that reason, provide general explanations for all types of integration. A pure transaction cost explanation is an example of the former variety. The control bridge removes transaction costs by positioning the actor at both ends of the former exchange. Haggling costs or unrecognized focal points are no longer a concern. Here the bridge itself is the source of the benefit. Another explanation that points to the bridge as the source of the benefit is the desire to reduce production uncertainty by ensuring the supply of an input. The actor relies on the control bridge to protect the good Y production unit from foreseen and unforeseen supply shortages of good X or terminations of supply relationships for other reasons. The exchange opportunism explanation (a variation of the supply argument) is another example of this type. The bridge is itself beneficial because it moves the

contractual interface to a competitive market and thereby reduces the opportunism threat. Nothing about the technology of the acquired production unit, in each of these cases, is motivating the actor to integrate.

We have reviewed some of the explanations that are specific to successive production. Other explanations are general. They flow from the fact that the control bridge puts the actor in a position to affect the production and exchange technologies of the acquired production unit. The nature of the effect will depend on the physical and information attributes of the actor and, where human capital is employed, the success with which the actor handles workers, transmits a comparative advantage, or exploits the advantages of others. Thus, an actor may integrate backward, forward, or horizontally to realize on the value of the slack thought to be present in the target production unit. This action does not find its impetus in the bridge *per se*. The actor is instead relying on a perceived superior ability to restrain production opportunism. This is also the case where the actor is relying on a comparative advantage as the motive for integration. It is the more efficient manipulation of the technology of the acquired production unit that is the source of the benefit. The bridge itself does not motivate the actor. It serves only as the means by which the advantage in the acquired technology can be exploited.

These general explanations would also apply in the conglomerate integration context. An actor may integrate unrelated technologies in order to extract the value previously lost to production opportunism or to exploit a comparative advantage. The bridge is not the source of the benefit. The actor benefits from the direct application of control to the acquired production and exchange technologies, and not from the simple control combination of the production units. There may be an explanation for conglomeration, however, that does implicate the bridge. The explanation is that conglomerate firms arise from a desire to diversify risk. The objection to this argument is that diversification could be arranged through appropriate personal investments in the market. It may be, however, that the actor prefers to diversify into investments that can be controlled. Where that is the case, the primary source of the benefit is the insurance function of the bridge rather than the manipulation of the technology of the acquired firm.

Horizontal, vertical, or conglomerate integration may, alternatively, be explained by the exploitation of economies of scope. Here the relevant production relationship is secondary. The integrated technologies may or may not be related in a primary way. Thus, for example, an actor may horizontally integrate in order to acquire a local

brand name to use to market the comparable products of the original firm. Or vertical integration may occur so as to make use of the heat produced by the neighboring technology. Finally, an actor may acquire an unrelated technology (conglomerate integration) in order to use its office, marketing, or distribution network. An insurance company, for example, might acquire a credit card company or department store chain through which it can extend the promotion of its products. The control bridge, in all of these circumstances, is the source of a secondary production or exchange technology benefit.

Disintegration is explained in the same terms. All of the various explanations for integration are in fact arguments about what motivates the actor. The actor, however, may have misjudged the benefit of acquiring control. The comparative advantage may not exist or the prospect of exchange opportunism may have actually increased. The actor may not be able to reduce information costs or to obtain economies of scale or scope. Alternatively, the size or complexity of the firm may lead to control failure in any of a number of ways. In all of these circumstances, there is an incentive for the actor to reduce or dispose of one or more technologies.

It is perhaps worth observing at this point that it is the net costs of control that are determinative for all types of integration and disintegration. Even in those cases where the control bridge itself is a source of benefit, integration may still not occur. This is because the benefit of the bridge will be assessed against the potential cost (or benefit) associated with the fact that the actor now controls the production and transaction costs of the acquired production unit. If the actor's control increases these costs beyond the benefit of the bridge, integration will be inefficient. For example, where a bridge is established, the costs of transfer between production units may be less than the costs of the former exchange. An additional benefit may be realized if the bridge also reduces the cost of exchange opportunism. These benefits, however, may not exceed the increase in the production and transaction costs, relative to the independent state, resulting from the actor's comparative disadvantage in relation to the production and exchange technologies of the acquired production unit. Thus, it will always be necessary to examine the total net costs of the independent and integrated states of the production units.

It might also be appreciated at this point that the production unit analysis identifies the boundaries of a firm (*i.e.*, its size) as congruent with the scope of the control of a particular actor. This can be understood in terms of the collection of production units controlled by the actor or, in more abstract terms, the collection of individual control

contracts negotiated by or for the actor which supply the operative physical and human capital. The absence of exchange or transfer relationships between commonly controlled production units (*i.e.*, conglomerate firms) does not alter these boundaries. On the other hand, it is possible for formal boundaries to be erected. The standard formal boundary is created by incorporation. Thus, an actor can isolate production units from each other by assigning them to separate corporate forms. This is an important effect and is the subject of much legal analysis. Interestingly, in that regard, the control principle also applies at the level of the formal construction to determine the boundaries of the corporate firm. Moreover, in some cases of external control the control principle overrides the formal boundary. In any event, for present purposes, it is sufficient to confirm the consistent appeal to control to establish the boundaries of all firms.

The foregoing analysis indicates that the firm is usefully understood in terms of the costs and benefits of an actor's control over the employment of assets. This conclusion is not undermined by the claim that all of the benefits of the firm could be secured through market purchases if transaction costs were zero. The reason this is so, quite apart from the plausibility of the claim, is that the analysis is intended to provide a framework that is comprehensive. It seeks to accommodate, in a single model, the various factors that apply at the production stage as well as at the exchange interface. Focussing on the production unit, and on exchanges and transfers between production units, achieves this. Doing so, however, does not preclude a more restricted enquiry. The model, in fact, anticipates this. The identification of a number of relevant variables contemplates their individual manipulation while holding the remaining variables constant. Thus, for example, the model allows for an investigation of the effect of holding production or transaction costs constant across production units. Similarly, in the immediate context, it allows for an investigation of the effect of reducing transaction costs to zero. In this way, the proposition that integration would not occur where transaction costs are zero is simply an assertion about what the model (*i.e.*, the world) would look like when the transaction cost variable is held constant at zero. This is how all multi-variable models are utilized. Carrying out the exercise under this model, it will be observed, does not result in the conclusion that firms do not exist where transaction costs are zero. The conclusion is instead that the economy becomes fully atomistic (or fully disintegrated), where each technology is represented by a single firm or set of firms. Each production unit is a firm, but no production unit is integrated, whether vertically or in a conglomerate firm. Apart from the exercise itself,

however, the point is that the model is a construct with which to organize the variables that may have application in the context of the firm. The construct would fail to be analytically useful only if it were unable to accommodate all of the relevant variables or credibly identify the source and mechanism of their individual or combined application.

VI. CONCLUSION

The firm is a structure through which production is made to occur. The structure is created in markets but subsequently exists, in a substantive way, apart from them. Some economists will dispute this. They find their explanation for the firm primarily in market considerations or in terms of the exchange interface. This fails, however, to provide a comprehensive picture of the firm. Production is what firms do. A complete understanding of the firm must necessarily proceed from that observable fact. Production is the process of exercising control over the employment of assets. It includes the creation of the control structure and its engagement in both exchange and production. Appreciating this involves discovering and assessing the benefits and structural limitations of control, including the associated technological and bridge-based motives for combined production. The task involves selecting the production unit as the basic analytical construct and then exploring the sources and mechanisms of the various factors said to affect the operation or combination of production units. The result is a model that accommodates human capital, single technology and integrated firms, production and transaction cost considerations, and the different types of integration, all within the single analytical rubric of production unit control.

This model has immediate consequences for lawyers engaged in the analysis of the legal structure of business organization forms. First, it will no longer be possible to invoke the "nexus of contracts" conception in the belief that this conveys meaning of any structural consequence. It will now also be more difficult to rely on propositions formulated solely on the basis of transaction cost considerations. More significantly, this model possesses normative implications that might reposition or redirect the legal regulation of economic organization. Thus, for example, it may be worth examining why exchange opportunism is not legally constrained in the same way as production opportunism. Or it may be important to investigate how the law affects the manipulation of symbols implicated in the creation of corporate culture. And it may be useful to determine, in view of the control

motives for integration, the current suitability of particular competition legislation provisions. The potential in the analysis of these kinds of questions is for a more efficient jurisprudence of business organization structure.

One final observation remains to be made. The production unit model is a representation of the economic structure of the firm. One of its functions, as noted, would be to inform assessments of the economic efficiency of the firm's legal structure. The nature of the legal structure itself, however, is a different matter, assuming that legal rules have primary sources other than efficiency. If the legal rules applicable to the firm arise initially out of public policy concerns with responsibility or with relationship integrity, for example, a model is required to explain the different legal characteristics of the alternative business organization forms. Such a model would necessarily have to proceed on assumptions about the material relations (economic structure) of the firm. Thus, in order to provide a foundation for those assumptions, the modelling of the economic structure should precede the modelling of legal structure. It is that economic modelling exercise that has been undertaken here. It is another task to construct the model of legal structure.