ARRIVING AT A STRATEGIC THEORY OF THE FIRM

Steven E. Phelan
and
Peter Lewin

School of Management
University of Texas at Dallas
Box 830688, Richardson, Texas, 75083
(972) 883 2754

sphelan@utdallas.edu

plewin@utdallas.edu

First Draft
January, 1999
Please do not cite or quote without permission.
Comments welcome.
ABSTRACT

Several authors have characterized their work as “moving towards” a strategic theory of the firm (Rumelt, 1984). The authors of this paper argue that it is time to start “arriving” at a strategic theory of the firm because a theory of the firm necessarily underlies every decision to enter a new line of business or outsource a function. Existing strategic theories of the firm are contrasted with economic theories of the firm and are found to be weak in explaining the existence of the firm while strong in their understanding of value creation and the location of firm boundaries. A distinction is made between firms in general and corporations. It is concluded that firms exist for a variety of reasons but that corporations have grown to dominate the business landscape because of their superior ability to create, and protect, value for their owners.
ARRIVING AT A STRATEGIC THEORY OF THE FIRM

INTRODUCTION

As many as half of all transactions in the economy are now conducted within firms. The theory of the firm sets out to explain the nature and limitations (or boundaries) of the firm as an economic institution. The theory of the firm can be located within the broader analysis of economic organization, which attempts to explain the observed diversity of institutional arrangements in the economy. For instance, why do some transactions occur in markets, others in firms, and still others in hybrid structures such as franchises, joint ventures and strategic alliances? The study of economic organization seeks to understand the conditions that create this diversity.

Ronald Coase’s (1937) observation that “…firms exist because there are costs to using the market” has been the touchstone for the study of organizations by economists. Attempts to identify the costs of using the market have led to fruitful developments in economics such as transaction cost economics (Williamson, 1975; Williamson, 1985), incomplete contracts theory (Hart, 1995), property rights and measurement costs (Barzel, 1989; Cheung, 1983) and agency theory (Alchian & Demsetz, 1972).

More recently, there has been an explosion of interest in the theory of the firm among strategy scholars (see Foss, 1998b for a review). Their attention has been caught by the possibility that there may be benefits to using firms rather than costs to using the market (Conner, 1991). It is argued that contemporary theories of strategy (particularly the resource-based view) are well positioned to explain these benefits. Several theories of the firm have been proposed from a strategic perspective, using a variety of (related) theoretical lenses, such as resources (Barney, 1996; Rumelt, 1984), knowledge (Conner
To date, there has been a tentative quality to much of the work in strategy, with several authors characterizing their work as “moving towards” a theory of the firm rather than forming a definitive theoretical statement. We believe that the field has evolved to the point where we are ready to start arriving at a strategic theory of the firm by drawing together the disparate threads of previous discussions into a single coherent theory. This conviction is the motivation for this paper.

DOES STRATEGY NEED A THEORY OF THE FIRM?

Every time a manager engages in a make-or-buy decision, he or she is utilizing (either explicitly or implicitly) a theory of the firm. The use of outsourcing, strategic alliances, joint ventures and franchising has grown markedly during the 1990s and has been accompanied by an increased scrutiny of all make-or-buy decisions within the firm. We have already alluded to the fact that the problem of where to set the boundaries of the firm is one of the central issues in the theory of the firm. A corollary to the question, “Why do firms exist?” is “Why are all transactions not performed by a single firm?” Any advance in our understanding of where to draw firm boundaries must ultimately improve the quality of managerial decision making in this area.

A theory of the firm can also inform the entrepreneurial process because it seeks to provide the conditions (both necessary and sufficient) for the establishment of a firm. The decision, by an entrepreneur, to start an enterprise implicitly reflects the belief that he or she can “beat the market” and produce a product more effectively through a firm than via a set of market contracts. Similar considerations apply when an existing firm
decides to vertically integrate or enter a new line of business. By offering plausible explanations for the existence and boundaries of the firm, a theory of the firm becomes capable of offering prescriptions for both the founding of new enterprises and adjustments to the horizontal and vertical boundaries of existing firms.

Some of the motivation for developing a strategic theory of the firm also emerges from the dissatisfaction felt by management scholars and practitioners towards existing economic theories of the firm. As will be discussed in the next section, existing economic theories of the firm tend to characterize firms as efficient devices for minimizing information and transaction costs under certain conditions. Williamson (1991) goes so far as to say that economizing on transaction costs should be far more important to managers than strategizing.

While we believe that economic theories of the firm have considerable explanatory power, we do not believe that they provide the whole story (or even half the story) for the existence and organization of the firm. A definitive theory of the firm must ultimately blend strategic and economic considerations.

ECONOMIC THEORIES OF THE FIRM

The set of economic theories of the firm is large ranging from neoclassical theory to industrial organization (Seth & Thomas, 1994) and from Adam Smith to Karl Marx. The subset of economic theories we are about to review currently rank among the most influential theories of the firm. While there are clear differences between the various theories within this set, all share a common set of assumptions about: 1) the need for exchange arising from specialization, 2) the need for coordination and cooperation among
economic agents, and 3) the assumption of efficiency in economic organization. Each of these assumptions will be discussed in turn.

Specialization and Exchange

It is well known that specialization through division of labor increases productivity (as Adam Smith so clearly discerned over two hundred years ago). It follows that allowing people to specialize in activities in which they have a comparative advantage increases the wealth of the society as a whole.

For Demsetz (1995), a ‘firm’ is any individual, or group of individuals, that undertakes specialized production, that is, producing for others rather than for their own households. Firms exist because they are able to produce goods more efficiently than households can. The alternative to the firm is not the market but the household. If firms are less productive than households, people will become self-sufficient and produce goods for their own consumption (Demsetz, 1995; Demsetz, 1997).

Specialization, however, is not in itself sufficient to improve economic welfare. Specialist producers must also be able to exchange their output with others in the economy. The emergence of money, for instance, greatly facilitated the process of exchange because it enabled agents with disparate outputs and time preferences to transact.

Because gains from trade can only be realized after exchange has occurred, economists have tended to focus on the transaction – the act of negotiating, executing and enforcing an exchange – as the most appropriate unit of analysis.

---

1 Foss (1993) refers to them as contractual theories of the firm.
This raises the question about what is actually being exchanged in a transaction. Most economists would now agree that it is the *rights* to an asset (broadly defined) that are being exchanged. Common rights that may be transacted include the rights to use, alter, transfer, alienate, or appropriate income from, an asset. An asset in this case includes labor, land, capital, or any combination thereof.

Transactions and contracts are closely related (in fact, Cheung (1969) views *all* transactions as contracts). It is not surprising that developments in the law of contract closely correspond with developments in exchange mechanisms (Masten, 1988). Over the centuries, the common law and statutes have prescribed (and proscribed) the rights that may be transferred in a transaction. For instance, a person cannot alienate (or sell) his or her own human capital nor is it possible to own someone else’s human capital (i.e. slavery is *per se* illegal). Similarly, under common law, the master-servant principle ensures that an employer (who has contracted the right to *use* someone’s human capital in exchange for wages) is also entitled to any gains or discoveries made by the employee while in his or her employ.

One may distinguish between *specific* rights, that are explicit terms in a contract, and *residual* rights, which are not. Common law has generally upheld the right of individuals and firms to make almost any specific provisions they wish in a contract provided certain basic conditions are met\(^2\). The law (i.e. society) also determines the nature of the default or residual rights that remain after a transaction.

\(^2\) Such as both parties being competent when they made the contract, considerations were exchanged, and nothing specifically illegal was contracted.
Ownership can be conceived as the bundle of residual rights that remains after all specific contractual arrangements. The most important of these residual rights are the rights to income and control. The owners of a firm, for example, are entitled to any surplus received after contractual obligations (including wages, rent, and debt payments) have been met. In lieu of any specific provisions, the owners of a firm are also entitled to use, alter, divide or sell the firm’s assets in any way they see fit. As we shall see, these two rights are very important in determining the benefits and costs of organizing exchange within a firm.

**Coordination and Cooperation**

In textbook models of comparative advantage, firms produce one discrete commodity (such as wine) and trade it for another discrete commodity (such as wheat). This greatly oversimplifies the nature of specialization in a sophisticated economy.

Advanced technologies, such as computer systems or modern jet aircraft, require the inputs of a multitude of specialists and specialist services. In the case of human capital, some of these inputs (e.g. a programmer) may be purchased in the market via a service agreement or, alternatively, employed by the firm using a wage contract. In the case of physical capital, such as a photocopier, a firm may choose to either hire the services of a copy bureau, lease a copier machine, purchase the machine outright, or even build its own copier machine.

This leads to the multifaceted problem of *coordination*. In the first instance, a firm (or entrepreneur) must perceive a market demand for a product. Second, a bundle of specialist inputs (or resources) must be identified that is capable of meeting the perceived market demand. Third, a firm must ensure that the process of assembling the inputs is
undertaken and that all parts are capable of productively working together to form the final product. Finally, a firm must determine the best form of economic organization for each input (lease/own, make/buy) to maximize the long-run performance of the company\(^3\).

The *cooperation problem* follows from the last point. The specialists that provide the inputs to the final product are assumed to be maximizing their own gains from exchange. The owner of the firm must therefore organize contractual arrangements in such a way as to ensure the interests of the agents (specialists) are aligned with the interest of the principal (owner). This is achieved mainly through the use of monitoring and incentives.

Problems of coordination and cooperation can be seen as types of information costs. In an economy with perfect information, there would be no coordination or cooperation costs. The correct technology to meet consumer needs would be known, and the effort and level of cooperation of employees could be easily measured. Consumer needs could also be easily determined. The fact that these things are not easily determined makes the issues of coordination and cooperation important. In fact, with no information costs the question of economic organization would be irrelevant. All resources would be accurately priced and could be costlessly exchanged via market transactions (Demsetz, 1988).

\(^3\) The so-called choice between the firm or the market is really a choice using one firm or two as all specialization is by firms and hence all exchange is between firms.
Economic Efficiency

All of the most influential economic theories of the firm subscribe to the hypothesis of economic efficiency which assumes that all observed forms of economic organization are efficient (absent external effects). A corollary of this hypothesis is that all opportunities for gain have been exhausted. If there was a better (i.e. more efficient) way of organizing production then firms in the economy would have implemented it.

Thus, the task of the organizational economist is not to find ‘better’ ways of organizing. Rather, it is to understand why certain forms appear in some contexts and not in others. Workers in the textile industry, for example, are almost always paid on the basis of piecework rates while college professors are hardly ever paid this way. Attempting to explain such differences in economic organization is the primary goal of organizational economics.

The theory of the firm is thus a subset of the overall problem of explaining economic organization and is focused on the tendency for some sections of the economy to coordinate transactions within the boundaries of a firm, while others parts of the economy choose to coordinate their transactions through the market. The goal is to explain this diversity of organization.

Specific Theories

We are now in a position to review some of the more influential economic theories of the firm. This will be no more than a brief overview. The goal is to gain an

\[^{4}\text{in the sense of Pareto efficiency where no one can be made better off without making someone else worse off}\]
appreciation for the reasoning involved. This will act as an aid in understanding both how strategists have departed from economic reasoning in their theories of the firm, and to illustrate the basis of the criticisms that economists have leveled at such theories.

*The Coasian Tradition*

According to Coase (1937), if all transactions are efficient then firms must exist because there are costs to using the market (or price mechanism). Coase identified several of these “transaction costs” including the costs of discovering prices; and searching, negotiating and concluding contracts. Coase believed a firm could avoid these costs by negotiating long term contracts with its employees. The fewer contracts signed over a given period of time the lower the transaction costs.

However, as contract periods lengthened so did the likelihood that business conditions would change in unforeseeable ways. To counter this effect, he predicted that employers would want to leave many details of their employment contracts unspecified so they could be free to direct employees to take different actions as conditions dictated. For Coase, this ability for owners to (re)direct the work of their employees was the distinguishing characteristic of the firm\(^5\).

He defended his argument by pointing out that the legal definition of a master-servant relationship (versus a principal-agent agreement) turned on the ability of the employer to direct the work of the employee whereas such a power did not exist in a principal-agent agreement.

\(^5\) i.e. the firm has residual rights of control over its employees.
Coase went on to provide several reasons why all transactions did not occur in one large firm. He argued that the number, dispersion and dissimilarity of transactions would negatively impact the ability of management to effectively allocate resources thus increasing the costs of integration.

**Measurement Cost Approach**

Cheung (1969; 1983) extended Coase’s analysis by discussing several additional costs to using the price system. Coase had always discussed transaction costs between firms, not between firms and consumers. Cheung reasoned that in a world without firms, consumers (or households) would be forced to coordinate the assembly of complex products themselves by buying the individual components from specialists. This raised several problems (i.e. costs) for consumers:

...the determination of prices is costly because of the number of transactions, because consumers lack detailed information on the use of each component or contribution to a commodity, because of the difficulty of measuring varied and changing activities, because of the need to separate contributions (Cheung, 1983, p.9).

Cheung suggested that all of the difficulties were due to *information costs*. In a world of perfect information these difficulties would not arise. Cheung also argued that firms did not supersede markets. Rather, one form of contract was replaced with another.

**Firm as Nexus of Contracts**

Alchian and Demsetz (1972) explicitly rejected the notion that authority provided a justification for the firm. Either side of an employment contract could terminate an
agreement just as either side of a market contract could terminate an agreement. An employee did not have to follow orders he or she did not like, so the firm had no special authority over its employees. Like Cheung, they believed that the term ‘firm’ was simply shorthand for a particularly dense nexus of contracts between economic agents.

Alchian and Demsetz’s started with the observation that much of the production in a firm could be characterized as “team production” in that the services of several specialists were required to produce an end product. While they viewed the coordination of production as relatively unproblematic, they felt it was very difficult to monitor the effort of individual workers in a team. Individual team members had an incentive to shirk by reducing their effort. Shirking reduced the surplus of everyone in the team prompting the hiring of a monitor to measure effort. But this created a new problem – monitoring the monitor.

The solution was to award the residual income of the firm to the monitor (now called the owner). The incentive to earn an uncapped income would motivate the monitor to keep a close check on the team. A firm was therefore defined as team production in the service of a monitor holding residual income rights.

Transaction Cost Approaches

Like Cheung (1983), Williamson (1975; 1985) and Klein, Crawford and Alchian (1978), hereafter KCA, made major contributions to Coase’s transaction cost framework, introducing new elements and significantly expanding the explanatory power of the theory. Whereas Coase had emphasized the ex ante costs of search and contract negotiation, Williamson and KCA focused on ex post transaction costs arising from an inability to enforce contracts.
Williamson (1975) began his argument by assuming that economic agents were boundedly rational. This ensured that any contracts written between agents were incomplete (i.e. agents could not contract for every contingency).

Williamson also introduced the notion of asset specificity. A contract between two parties invariably created assets specific to the relationship that could not be easily deployed to alternative uses (i.e. sunk costs). The difference in value between an initial investment and its salvage value (or its value in its next best use) was termed quasi-rent (Klein et al., 1978).

By acting opportunistically to renegotiate an (incomplete) contract in the light of unforeseen contingencies, an unscrupulous party could potentially holdup its partner and appropriate the quasi-rents. Of course, if the partner firm could foresee this risk, the deal would not proceed in the first place. Paradoxically, unanticipated holdup never occurs in Williamson’s world.

The alternative for a firm with substantial quasi-rents was to vertically integrate into the activity of its partner, thus holding all specific assets itself and removing the threat of holdup. The extent of a firm’s integration was limited at the margin by the tradeoff between governance costs and transaction costs.

*Property Rights Approaches*

The property rights approach (Barzel, 1989; Hart, 1995; Hart & Moore, 1990) has further extended the transaction cost approach. The focus of the work is on the distinction between specific and residual rights (discussed earlier). In a world of incomplete contracts, it is impossible to define specific rights for all contingencies. Firms will thus spend resources trying to acquire valuable unspecified rights (Barzel, 1989).
Ownership of physical assets becomes important because ownership grants residual rights of income and control (Grossman & Hart, 1986; Hart & Moore, 1990). When contracts are incomplete, it is the owner that ultimately has the right to decide on the final disposition of assets. Furthermore, if changing the deployment of an asset increases the value of the firm then the asset owner is also entitled to any surplus received. This has prompted Hart (1995) to recommend that the firm with the greatest potential surplus from an asset should own that asset.

Earlier, we saw that Alchian and Demsetz (1972) claimed that a firm has no right of fiat (or authority) over its employees. Initially, the property rights approach would seem to support this position as human capital cannot be owned (i.e. slavery is illegal). This conclusion has been disputed in the property rights literature. One of the rights of ownership is the right to exclude others from access to your asset. Many employees are in their highest valued use only when working for a particular company – their human capital is relationship specific. Presumably, these employees will be amenable to directions because the firm has the power to a) appropriate their quasi-rents and b) remove access to the corporate assets altogether (i.e. fire them).

For example, a commercial pilot is in his or her highest valued use when working for a commercial airline. During the Australian pilots’ strike of 1989, 85%-90% of scheduled air services were cancelled for several weeks as pilots pushed for a 30% wage rise. Buoyed by Federal government support, the commercial airlines simply fired all of their pilots and hired new pilots from overseas at lower rates than before the strike. Some
of the ringleaders of the strike have never flown in Australia again and have suffered a considerable loss of income and prestige\(^6\).

This power inherent in controlling access to physical assets has led property rights theorists to define the firm in terms of the ownership of physical assets (where ownership denotes residual rights of control and income). Unlike the nexus of contracts view, this definition makes the boundaries of the firm relatively easy to determine (Foss, 1997). If firm X owns, or has residual rights over, asset \(x\) then asset \(x\) lies within the boundaries of firm X.

Mahnke (1997) has argued that the power to control access to corporate assets may also extend to non-physical or intangible assets, such as reputation, information and brand names. This may explain why top accounting and consulting firms are able to hire the best students from prestigious colleges at salaries considerably below market rates. From the students’ perspective, they are enhancing their human capital through exposure to a leading firm’s reputation, techniques and client base. Thus, recognition of the ability to appropriate quasi-rents by controlling access to corporate assets may be important in developing a comprehensive theory of the firm.

**STRATEGIC THEORIES OF THE FIRM**

As a rule, strategic theories of the firm have been developed by scholars in strategic management, published in management journals, and have used concepts and constructs drawn from management and evolutionary economics to support their arguments.

\(^6\) A similar example would be the air traffic controllers strike in the U.S. during the Reagan administration.
Discussions on a strategic theory of the firm are relatively recent, dating from Rumelt’s (1984) paper “Towards a strategic theory of the firm”. Like the economic theories surveyed above, there are several core principles that are common to all strategic theories of the firm. In fact, given the short period of development, these principles are probably more homogenous among strategic theories than those in economics.

**Common Principles**

Strategists tend to agree on three broad principles: 1) the resource-based nature of the firm; 2) the determination of firm boundaries; and 3) bounded rationality.

**Nature of the Firm**

Following Penrose (1959), the firm is seen as a bundle of productive resources under administrative direction. Resources are defined as inputs into the firm's operations that are used to produce products or services (Wernerfelt, 1984). Examples of resources include patents, capital equipment and skilled human resources. A capability (or competency) is the ability to perform a task or activity that involves complex patterns of coordination and cooperation between people and other resources (Grant, 1991; Schulze, 1994). Capabilities include research & development, excellent customer service, and high quality manufacturing.

Strategists are particularly (some would say exclusively) interested in those resources and capabilities (known collectively as strategic assets) that earn rents, where a rent is defined as a surplus of revenue over cost (Amit & Schoemaker, 1993). Strategists seek to create, and protect, rents in order to enhance the value of their firm.\(^7\) Resources

---

\(^7\) The value of a firm is the sum of its rents discounted at the relevant cost of capital.
are known to be more valuable if they are rare, difficult to imitate and non-substitutable (Barney, 1991). However, in a competitive market, rents can only arise because of luck or differences in expectations concerning the value of a resource (Barney, 1986). Otherwise, the supplier of a resource can appropriate its full value.

If the owner of a resource cannot envisage how a firm will use a given resource then the price of the resource will reflect the next best use (or opportunity cost) that its owner can find. The strategic value of a given resource is thus likely to depend on the way a firm combines, coordinates, and deploys that resource with other firm-specific and firm-addressable\(^8\) resources (Sanchez & Heene, 1997). It is in the interest of the firm to use a new resource in ways that its former owner could not envisage.

Of course, not all resources can be traded. Some resources are developed within the firm and cannot be bought or sold in factor markets (Dierickx & Cool, 1989). Examples of these non-tradable assets include reputation, culture, firm-specific know-how, and values. These assets may be particularly difficult for competitors to imitate because of time compression diseconomies, asset mass efficiencies, interconnectedness of asset stocks, asset erosion, and causal ambiguity (Dierickx & Cool, 1989).

**Boundaries**

According to Penrose (1959), the routinization of the firm’s activities frees up management resources to explore new opportunities. The indivisibility of certain assets also creates excess capacity that, although not easily traded (see non-tradable assets above), can be utilized in new activities. Management’s inability to conceive and control

\(^8\) A resource that lies outside the boundaries of the firm but can nevertheless be used by the firm.
new activities places limits on the growth of the firm. This ‘cognitive limits to growth’ perspective has found considerable support in the recent strategy literature (Ginsberg, 1994; Prahalad & Bettis, 1986).

Richardson (1972) views the economy as a set of activities each requiring a set of capabilities (i.e. knowledge, skills, and experience). Similar activities share common capabilities. Firms can realize gains (through economics of scale, scope and experience) by expanding into similar activities. Complementary activities lie in the same value chain but require dissimilar capabilities. The closer the complementarity between stages of production, the greater the degree of cooperation required between firms. This cooperation could include any form of joint planning between firms including strategic alliances, joint ventures or vertical integration.

As rent-seekers, firms are advised to expand only into areas where they have a competitive advantage. Growth will thus be constrained by the fungibility and transferability of the firm’s most valuable resources (which are by definition rare and difficult-to-imitate).

**Bounded Rationality**

In the neoclassical world of perfect information, differences in expectations cannot exist (Demsetz, 1988, Demsetz, 1997). Firms are assumed to have the “blueprints” for all production possibilities (Foss, 1998b). Economic theories of the firm do not move far beyond their neoclassical roots. The notion that production knowledge may be difficult to acquire receives little attention in the literature. It is assumed *a priori* that a potential gain from exchange exists. Consequently, the theories focus on structuring the transaction so that the partners to the deal can agree on the division of the expected gains.
The assumption that all production possibilities (or gains from exchange) can be easily discovered is antithetical to strategists. At a given point in time, a firm will never know whether its particular combination of resources is in the best possible use to maximize rents. Nor will it know whether its choices reflect the best possible deployment of resources to meet future contingencies (i.e. to optimize the value of the firm over its lifetime). Strategic theories problematize the assumption that production possibilities can be easily discovered.

Why is discovery difficult? Strategists (either implicitly or explicitly) assume that resource holders are boundedly rational. Managers do not have perfect knowledge of future states of the world, of alternative actions that may be taken should such states arise, nor of the payoffs from adopting various alternatives (Newell & Simon, 1972). Moreover, the way a manager chooses to allocate resources will be a function of that manager’s past experience, skills, values, biases, and personality. Socio-cognitive processes, leadership in creating shared visions, imagination, and creativity thus become important factors in explaining resource allocation decisions (Ginsberg, 1994; Prahalad & Bettis, 1986).

Accordingly, even if two managers were given identical bundles of resources, they would tend to use them in different ways. The result is that, over time, a firm’s stock of tradable and non-tradable assets will diverge from its competitors, as will its performance. Managers in competing firms do not face the same set of choices. Rather, they have different menus with different choices (Teece, Pisano, & Shuen, 1997).

9 Resources are usually assumed to be heterogeneous.
For Demsetz (1997), information costs refer to the difference between the real world and the “perfect” world of neoclassical economics:

This creates a productive role for management where none exists in neoclassical theory. Imperfect information... makes the judgment of managers and owners a source of productivity enhancement. The main source of management's productivity in contemporary theory has been in its response to agency problems. Shirking, opportunism, and reputation are brought to the fore. This effort has led to the neglect of information problems that do not involve agency relationships. These are associated with planning in a world in which the future is highly uncertain, and they include problems of product choice, investment and marketing policies, and scope of operations. Neglect of this class of problems is unfortunate (Demsetz, 1997, p.428).

Strategists are thus uncomfortable with the view that all outcomes are globally efficient. Global efficiency implies that the future can be discerned, or a superior production technique discovered, by simply investing in more “information”\(^{10}\).

An alternative is to embrace the assumptions of Austrian and evolutionary economic theory and assume that the future is, to a greater or lesser degree, uncertain and unknowable (Hayek, 1945; Nelson & Winter, 1982). Thus, no amount of expenditure on information gathering will resolve what Knight (1921) calls ‘fundamental uncertainty’. Firms are only relatively efficient (or inefficient) in reference to other firms. Firms adapt
but they do not optimize (or, more correctly, they never know if they have reached an optimal position).

In the neoclassical approach, inefficient firms do not survive. They are assumed to fail very rapidly leaving only efficient firms (Friedman, 1951). Diversity in firm performance can only be explained using *ad hoc* assumptions such as information costs or short run adjustment constraints. An evolutionary approach offers a more parsimonious explanation for the survival of under-performing firms. First, there is always the chance that a firm will improve its performance by altering its resource allocations. Liquidating a firm prematurely destroys, or damages, the firm’s stock of non-tradable assets. Second, the firm may not, in fact, be under-performing. It may be implementing a strategy with a payoff over an extended time period. In either case, it may be rational to ‘wait and see’ if performance improves. Consequently, firms strive to improve performance but survival is also important (Alchian, 1950).

**Specific theories**

This section briefly discusses several strategic theories of the firm. The goal is to capture some of the essence of the thinking around these issues. We begin with the knowledge-based theory of the firm, move on to options-based theory and conclude with the dynamic transaction costs view.

---

10 In neoclassical economics, efficient behavior is sometimes synonymous with purposeful (rational) behavior. This usage reduces “efficiency” to a behavioral postulate with no normative power to adjudicate outcomes.
Knowledge theories of the firm (KTF) form the largest subset of the growing literature on strategic theories of the firm (Conner, 1991; Conner & Prahalad, 1996; Grant, 1996; Kogut & Zander, 1992; Liebeskind, 1996). All KTF invariably start from the premise that knowledge is the most important strategic asset that a firm possesses. Grant (1996) goes so far as to say that “…all human productivity is knowledge dependent, and all machines are simply embodiments of knowledge” (p. 112). The literature, however, makes a strong distinction between explicit knowledge in the public domain and tacit knowledge.

Tacit knowledge cannot be easily codified and can only be learned through observation and practice (Kogut & Zander, 1992; Grant, 1996). Riding a bicycle is an oft-quoted example of tacit knowledge. Tacit knowledge is a valuable resource because it cannot be directly appropriated and attempts at imitation will be costly in terms of time (even assuming an instructor can be found). Explicit knowledge, on the other hand, is extremely easy to transfer and contracts protecting knowledge are difficult to enforce (Liebeskind, 1996).

Several explanations for the existence of the firm have been offered in the KTF literature. Liebeskind (1996) argues that the firm exists because it is more efficient at protecting (explicit) knowledge. For example, employment contracts hinder knowledge transfer by requiring exclusivity, confidentiality, and restrictions on competitive behavior. Similarly, the threat of dismissal may also deter attempts to transfer knowledge. Job design and division of labor within a firm ensure that employees don’t see the ‘whole
picture’. Firms can also impose exit costs on employees by deferring compensation in the form of pension plans, stock options or promotion.

By definition, tacit knowledge resides in individuals but must be integrated to become useful. Grant (1996) argues that the continuity of association found in a firm enables the development of common knowledge, a shared language and meaning among the firm’s specialists. The organization’s rules and routines also assist in facilitating coordination. Firms are therefore more efficient at integrating knowledge than the market.

Conner and Prahalad (1996), focus on the advantages that administrative direction provides over market contracting. They note that a market contract needs to be re-negotiated if requirements change whereas employees can be more easily directed to new tasks. As Coase (1937) also noted, this gives the firm flexibility and lowers transaction costs. Similarly, market contractors often need to be convinced of the merits of a project before they commit to production. Integrating the activity into the firm removes the need to educate the supplier or distributor on the merits of the project. Conner and Prahalad (1996) refer to this as the knowledge- substitution effect.

**Dynamic Transaction Costs**

The knowledge-substitution effect bears a strong resemblance to the concept of dynamic transaction costs discussed by Langlois (1991) and Langlois and Robertson (1995). For these authors:

The boundaries of the firm are determined entirely by the capabilities of the firm relative to the capability of the market (Langlois & Robertson, 1995, p.33).
Dynamic transaction costs are defined as:

…the costs of persuading, negotiating, coordinating and teaching outside suppliers…[about your capabilities and strategic architecture]…it is the cost of not having the capabilities you need when you need them (Langlois & Robertson, 1995, p.35).

Transferring knowledge to suppliers also runs the risk that the suppliers might appropriate the gains from innovation (Langlois & Robertson, 1995; Teece, 1986). Appropriation can occur either in the form of holdup (i.e. higher input prices) or forward integration. In either case, the firm has an incentive to transcend the market through integration.

*Options*

The options-based theory of the firm formalizes Conner & Prahalad’s (1996) insight that *flexibility* is valuable to a firm (Barney & Lee, 1998; Foss, 1998b; Sanchez, 1993; Sanchez, 1998). In finance, an option is the right, but not the obligation, to trade a security at a fixed price (the exercise price) over a given period of time\(^\text{11}\).

The right is valuable because the market price of the security may rise or fall over the life of the option contract. The holder of a call (put) option will exercise the option if the market price rises above (falls below) the exercise price. If the market price moves contrary to the option holder’s expectations then the loss will only be limited to the premium (i.e. the initial price of the option contract).

\(^{11}\) The right to buy a security is called a call options, the right to sell a security is known as a put option.
The probability that the market price will differ from the exercise price is, among other things, a function of the volatility of the underlying security and the length of the contract. Option pricing will reflect the market’s assessment of the likely gain from these potential price movements.

As we have argued earlier, the firm is a bundle of resources over which its owners/managers have residual rights of control. The flexibility to change the allocation of resources in the firm allows management to design new products, alter the timing of investment projects, and vary relationships with suppliers and distributors (Phelan, 1997; Sanchez, 1993). Options theory makes explicit the link between flexibility and value. Firms provide their owners more flexibility and are thus more valuable (i.e. they are more effective than markets at providing strategic options).

Options-based theory is also vocal on the issue of determining the boundaries of the firm:

Firms should internalize only a few inputs that 1) are exceptionally difficult to obtain through markets and 2) are capable of generating superior options values for the firm (Foss, 1998, p.7).

This statement hints at the fact that not all resources are capable of generating options. The decision to acquire a resource must ultimately depend on a) its premium (or acquisition price), b) its current value to the firm, and c) its ability to be recombined with other resources to create value – the expected value of future gains represents the resource’s option value. A commitment to learning about the ways in which resources may be combined increases the value of options (Barney & Lee, 1998). However, unless a difference of opinion exists between the seller and acquirer of the resource, in a
competitive market the acquisition price will tend to reflect the current and future value of the asset to the acquirer.

**Criticisms of Strategic Theories**

Williamson ((1991), p.76) has argued that “…economizing is more fundamental than strategizing”. However, a close reading of Williamson reveals that he views strategy as synonymous with market power and economizing as efficiency. Williamson identifies the resource-based and dynamic capabilities views as the leading efficiency approaches to business strategy but criticizes their lack of structure and focus. While Williamson fiercely defends the *vital* role of transaction cost economics in explaining governance choices he leaves the door open for resource-based explanations.

Nicolai Foss (1996a; 1996b; 1997) has made some thoughtful criticisms of strategic approaches to the theory of the firm. According to Foss (1997), the resource-based theory of the firm cannot easily explain:

- The existence of the firm;
- Why firms are defined by their ownership of [physical] assets;
- When capabilities are important for economic organization and when they are not; and
- Future firm boundaries or causes of success.

Most importantly, he argues that “co-specialization [carries] no implications for ownership” (Foss, 1996a, p. 473).

agents…could simply meet under the same factory roof, own their own pieces of capital equipment or rent it to each other, and develop value-enhancing…principles among themselves. (Foss, 1996a, p.473)
We agree that strategists have often not discerned the need to justify the existence of the firm in these terms. The next section addresses Foss’s central concerns.

ARRIVING AT A STRATEGIC THEORY OF THE FIRM

The Existence of the Firm

As Demsetz (1995, 1997) has noted, the alternative to the firm is not the market but the household. Firms exist because they are able to assemble valued goods and services more efficiently than households can. The source of this efficiency is the productivity enhancing effects of specialization; comparative advantage; division of labor; and associated economies of scale, scope and experience.

However, Demsetz’s use of the word ‘firm’ in this context is somewhat unusual. In everyday speech a ‘firm’ refers to a corporation or cooperative, comprised of a number of people and other assets, which may have legal status as a company or partnership. Demsetz’s ‘firm’ could be a sole trader, a specialist who produces for the consumption of others beyond the needs of his or her own household. The modern theory of the firm is really seeking to explain the existence of the corporation (in the above sense).

Coordination

As we have already discussed, advanced economies are characterized by a high degree of specialization. The creation of any consumer good requires the involvement and coordination of countless specialists each with their own tacit knowledge domain. On this score alone, the firm can transcend the market by reducing the costs of coordination by, for example, lowering search costs for consumers and producers (Coase, 1937; Barzel, 1989) or by providing a common language and set of routines (Grant, 1996).
Moreover, coordination is difficult. Resources cannot be assumed to be in their value-maximizing use at any given point in time. The continuous stream of new developments and innovations emerging into the public domain further complicates the problem of finding appropriate resource combinations (Hayek, 1945; Nelson & Winter, 1982; Schumpeter, 1934; Teece, 1986; Witt, 1998). The problems of coordination ensure that opportunities for gain will always exist.

*Behavioral Assumptions*

We assume that firms are not formed with the intention of creating social welfare (or public efficiency). Specialists trade for their own gain or profit (private efficiency). So, too, do corporations. Firms are profit-seeking entities. Thus, it is reasonable to assume that all parties to a transaction will act to maximize private gain.

Unlike other strategists (Kogut & Zander, 1992; Conner & Prahalad, 1996) we do not believe that it is necessary to reject the assumption of opportunism to develop a strategic theory of the firm. This is not to say that altruism does not occur in corporations. Rather, we believe that ‘planning for the worst while hoping for the best’ is the path of least regret. Also, as Axelrod (1984) has demonstrated, it may be in one’s self-interest to act cooperatively. Cooperation is not beyond our definition of self-interest (Langlois & Robertson, 1995).

We also assume that managers possess bounded rationality. However, our conception of bounded rationality goes beyond the lack of ability to detect shirking or to write complete contracts. In a *strategic* theory of the firm, bounded rationality is extended to include differences in expectations among economic agents concerning the value of production possibilities. These differences in expectation arise because the
coordination of production is a highly dynamic and complex task. An entrepreneur is someone who is able to perceive that a gain can be made by recombining resources in a novel way and seeks to profit from that perception (Schumpeter, 1931).

*The Special Status of the Corporation*

The property rights of a corporation are protected by common law, social convention and the corporation’s own actions (Barzel, 1989). They include the rights to:

- Receive any surplus income generated from its assets;
- Direct, control or transfer non-human assets;
- Alienate (or sell) physical assets;
- Lawfully direct the work of employees; and
- Appropriate any proceeds from the employment relationship.

*Discussion*

Let us start with an entrepreneur and an existing stock of rare and difficult-to-imitate resources (land, labor and capital) that are held by other agents in the economy. Each agent possesses bounded rationality so there is no reason to presume that the resources are deployed in their most efficient use in some global sense. At some stage, let us assume that the entrepreneur “discovers” some undervalued resource or resource combination (or asset co-specialization) that was hitherto unperceived by the other agents in the economy. For the purposes of this analysis it does not matter how the insight arises – luck, creativity, imagination, superior information – are all possible explanations.

To make the example more concrete, assume that the resources in their new combination are estimated to yield a capitalized present value of $100 million. The marginal contributions from each resource being: land ($20m), labor ($40m), and capital
($40m). The resources in their current use are earning a total of $50 million – land ($10m), labor ($20m) and capital ($20m). [All values are prospective estimates.]

Once a profit opportunity has been discovered, the question then arises as to which organizational arrangement is best suited to its realization. Is the entrepreneur indifferent to hiring the inputs through market contracts or forming a firm (which would entail purchasing at least some of the physical assets and placing at least some of the labor on employment contracts)?

The first point to note is that any contract will be incomplete due to bounded rationality. In effect, this means that any input owner under contract (including labor) will have the opportunity to periodically re-negotiate the terms of their contract. Thus, while the landlord might initially rent the land at a capitalized value slightly above $10m, the rent is likely to subsequently rise towards $20m (and beyond).

The transaction cost story (Williamson, 1985) is that relationship specific assets (i.e. those that are generating quasi-rents or returns above their next best use) are always subject to such holdup or *ex post* appropriation\(^\text{12}\). The entrepreneur is strongly advised to buy the assets outright to protect the gain. The property rights implicit in ownership enable the entrepreneur to alienate the residual income between the asset’s new use and its sale price.

How much will the entrepreneur actually pay for the land? This is entirely negotiable and will lie somewhere between $10m (its value in its current use) and $60m

\(^{12}\) Similarly, the dynamic transaction cost story (Langlois & Robertson, 1995) is that the entrepreneur might not be able to convince the input holders of the value of the new idea and thus cannot convince them to shift their resources to a new use.
(assuming the entrepreneur is able to obtain the other inputs at their current use value). However, given that the landlord is ignorant of the entrepreneur’s future plans, the price could reasonably be expected to lie at the lower end of the range (although we know that acquiring firms often pay significant premiums to acquired firms).

Once the entrepreneur purchases the specialized land and capital, he or she also acquires two additional items of value. The first is a strategic option to recombine the assets in new ways. As we described earlier, even if the input owners appropriate the full $100m value of the initial plan, the deal would still be valuable because of the underlying flexibility to redirect resources (itself a function of the property rights inherent in ownership). The firm will also start to build non-tradable assets like reputation, goodwill, and a corporate culture. While these assets cannot be acquired on the market they are, nevertheless, valuable (Dierickx & Cool, 1989).

Both of these facts complement, rather than supplant, transaction cost logic (Barney & Lee, 1998; Foss & Foss, 1998). The value of the strategic options and non-tradable assets simply increases the quasi-rent that may potentially be appropriated and further biases the entrepreneur towards creating a firm to protect the (increasingly) co-specialized resources.

From the entrepreneur’s viewpoint, the weak link in the arrangement is labor. At some stage, labor contracts (whether market-based or employment-based) will need to be re-negotiated. In our previous example, wages are likely to commence at a capitalized value of $20m. If land and capital were acquired at their opportunity cost of $10m and $20m respectively then labor has the potential to appropriate the remaining $70m in future negotiations (and any gains from strategic options and non-tradable assets that may
also emerge). Is the firm a better vehicle for protecting the entrepreneur’s gains than the market?

The answer is no. Both market and employment contracts need to be renegotiated. Instead, it is the ownership of physical assets that gives the firm bargaining power in both cases. Ownership carries with it the right to restrict or deny access to a resource (Hart, 1990). If labor becomes too strident (as in the Australian airline example), the owners can exclude their labor force and train a new group. Labor with firm-specific skills (often built over a lifetime) is more vulnerable to holdup in these circumstances. The ability to defer compensation also suppresses bargaining power (Liebeskind, 1996). However, even if labor were to appropriate all gains from the relationship, there would still be a period of adjustment between the old and new rates from which an entrepreneur could profit (Rumelt, 1984). Finally, as with the example of professional firms, exposure to the specialized resources of a reputable firm early in one’s career can enhance lifetime earnings (by increasing one’s human capital). This can lead employees into a “rat-race” where current wages are lowered for a winner-take-all prize (Rajan & Zingales, 1998).

There is also another set of reasons the firm may prefer employment contracts. First, as Coase (1937) first noted, employment contracts provide the flexibility to direct staff into new assignments. Firms also have the legal right to appropriate inventions and other assets arising from employment (a residual right they don’t have with market contracting). In both cases, the option value (and hence quasi-rent) of the firm is increased. Second, as Grant (1996) points out, an on-going association between employees creates shared knowledge and a common language, both non-tradable assets that increase option value.
Implications

Two views on the nature of transaction costs have emerged from our discussion. The first view, attributable to Coase (1937), emphasizes the *ex ante* costs (in terms of time, money, and effort) of identifying and negotiating for the exchange of valued inputs using the price system. Strategists have identified strongly with this theme, making several contributions concerning the benefits of the firm in coordinating knowledge (Grant, 1996) and the role of resources in determining boundaries (Penrose, 1959).

The second view of transaction costs, attributable to Williamson (1975; 1985), emphasizes the *ex post* costs of holdup when suppliers and distributors appropriate the gains from asset specificity. Awareness of these *ex post* costs influences *ex ante* behavior. The firm, with its common law property rights, provides a convenient mechanism for avoiding these costs (or alternatively protecting any gains from specialization). The key message is that the value realized from entrepreneurship, organizational learning, saving on *ex ante* transaction costs, or strategic options is *inconsequential* if the public gain can not be translated into a private gain.

The antagonism between the strategic and economic theories of the firm (Williamson, 1991; Ghoshal & Moran, 1994; Williamson, 1994; Kogut & Zander, 1996) is thus difficult to explain. Views of the firm that focus on value creation are clearly *complementary* to Williamson’s transaction cost view. Perhaps it is because Williamson has used the term ‘transaction cost’ when talking about ‘appropriability’ that this misunderstanding has occurred since strategists have long appreciated the link between shareholder value and appropriability (Porter, 1980; Teece, 1982).
However, we disagree with the view that opportunism is either necessary of sufficient to explain the existence of the firm (Foss, 1996a, 1996b). Coase’s ex ante transaction costs arise solely as a consequence of bounded rationality. There is no holdup in Coase’s world, simply a problem of finding information that will enable resources to be allocated to their best use. Thus, whenever the costs of managed coordination are lower than the costs of price coordination, firms will subsume markets (Demsetz, 1995).

The assumption of opportunism is needed to explain the overwhelming prevalence of capital-managed firms (KMFs) relative to labor-managed firms (LMFs). Dow (1993) argues that capital providers would never lease their resources to LMFs because of their inability to protect fixed assets (i.e. highly specific assets) from appropriation. LMFs only arise when capital plays a small part in value creation (e.g. in professional firms). He goes on to argue that the firm should be viewed as a nexus of [negotiation] strategies rather than a nexus of contracts as the division of income between labor and capital is highly negotiable. We would add that negotiating to acquire physical assets also plays a key role in protecting entrepreneurial value.

The threat of appropriation of rents by buyers and suppliers has always played a large role in the theories of Porter (1980) and Teece (1986). Curiously, the resource-based literature has always been more concerned with imitation of key resources by competitors than appropriation of quasi-rents by resource owners. However, increasing a firm’s value depends as much on obtaining favorable terms from resource providers as it does on protecting resources from imitation by competitors. Opportunism is not a necessary condition for the existence of the firm but it would be incorrect to suggest that it plays no role at all.
The Boundaries of the Firm

Strategists have generally had more success in explaining the boundaries of the firm than they have had in explaining its existence. It was Coase (1937) who first posited that firms exist because there are costs to using the price system. Similarly, the reason that all production is not organized in one large firm is because the cost of planned production will, at the margin, exceed the cost of using the price system.

For Penrose (1959), the number of resources under the firm’s control contributed to the development of new resources and new opportunities for gain. This was offset by the inability of management to envision new ways of combining the increasing number of assets under their control (Coase, 1937; Prahalad & Bettis, 1986). For Coase (1937):

…the cost of losses through mistakes will increase with an increase in the spatial distribution of transactions organized, in the dissimilarity of transactions, and in the probability of changes in the relevant prices (25).

Strategists are generally well-informed about the cognitive, social and behavioral factors that hinder resource allocation in large firms (Bower, 1986; Hoskisson & Hitt, 1994). A discussion of these factors is beyond the scope of this paper but once a limit to the size (or resources) of a firm has been accepted, the question naturally arises about which resources should be included within the firm.

The answer to the above question is that the firm should acquire the bundle of resources that are expected to produce the most value for the firm. Resource-based theorists (Amit & Schoemaker, 1993; Barney, 1991; Peteraf, 1993) have catalogued several characteristics of valuable assets including:

- scarcity
• inimitability,
• inappropriability,
• durability, and
• flexibility.

Maximizing value entails balancing the current and future (option) value of the firm. Resources that can be leveraged into new applications or used to develop new resources are clearly more valuable from an options-based perspective (Sanchez, 1993).

CONCLUSION

According to Kirsten Foss (1998a), economists have tended to work from the assumption that potential gains from exchange are known to exist and agents simply need an incentive to realize them (typically an assurance that they will receive some or all of the potential gains). Strategic management enriches this perspective by pointing out that opportunities for gain are not readily known and value must be discovered.

Firms (as corporations) exist because they are better at protecting these gains after they are discovered. Firms are also incubators (or engines) for future value creation. The property rights associated with ownership play an essential role in creating these benefits of incorporation. The firm is thus more than a nexus of contracts. It is also more than just a superior way of coordinating knowledge and resources among cooperative agents.
REFERENCES


