Arriving at a strategic theory of the firm

Steven E. Phelan and Peter Lewin

The theory of the firm seeks to explain the existence and boundaries of the firm in relation to the market. Since the pioneering work of Coase (The nature of the firm. *Economica*, 4, 386–405, 1937), economics has developed a whole family of theories that focus on the ability of firms to economize on certain costs of using markets. More recently, researchers in strategic management have published several theories of the firm that have tended to emphasize the benefits of incorporation rather than the costs of using the market. Although researchers in the strategy profession have tentatively labeled their work as ‘moving towards’ a strategic theory of the firm, economists have been very critical of existing approaches. This paper seeks to begin ‘arriving’ at a strategic theory of the firm by addressing these criticisms and offering an integration of the strategic and economic perspectives within an institutional framework. The paper concludes with future directions for research in the theory of the firm.

Introduction

The theory of the firm sets out to explain the nature and limitations (or boundaries) of the firm as an economic institution (Demsetz 1995). 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. 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, 1985), incomplete contracts theory (Hart 1995), property rights and measurement costs (Barzel 1989; Cheung 1983) and agency theory (Alchian and 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. These theories have utilized contemporary theories of strategy (particularly the resource-based view) to highlight the benefits of incorporation rather than the costs of using the market (Conner 1991). 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 and Prahalad 1996; Grant 1996), competencies (Foss and Knudsen 1996; Penrose 1959), etc.
Arriving at a strategic theory of the firm

capabilities (Langlois 1992) and real options (Barney and Lee 1998; Sanchez 1998).

In making their case, strategy theorists have been critical of existing economic theories of the firm, particularly Oliver Williamson’s transaction cost economics (Conner and Prahalad 1996; Ghoshal and Moran 1996). In turn, economists have responded by developing their own set of criticisms about the new strategic theories (Foss 1996a,b; Williamson 1996a). In our view, much of this conflict arises from a misunderstanding of the other side’s terminology, central issues and research agendas rather than a fundamental disagreement about the nature of the phenomenon in question. The aim of this review is to demonstrate that the strategic and economic theories of the firm are complementary, rather than competing, worldviews.

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 (Foss and Knudsen 1996; Grant 1996; Rumelt 1984). It is our contention that, by resolving their differences with economists and integrating perspectives, strategists will finally be able to start arriving at a strategic theory of the firm. The paper concludes with a discussion of future research directions that support this call for integration.

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, accompanied by an increased scrutiny of all make-or-buy decisions within the firm (McFarlan and Nolan 1995; Quinn 1999). The problem of where to set the boundaries of the firm is one of the central issues in the theory of the firm (Coase 1937). A corollary to the question, ‘Why do firms exist?’ is ‘Why are all transactions not performed by a single firm?’ Any advance in understanding 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 integrate vertically 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 (Poppo and Zenger 1998).

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 (Conner and Prahalad 1996; Ghoshal and Moran 1996). 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 complete story for the existence and organization of the firm. A definitive theory of the firm must ultimately blend strategic and economic considerations.

As discussed in the introduction, we believe that much of the conflict between economic and strategic theories of the firm lies in a misunderstanding by one group of the terminology, research agenda and perspective of the
other group. We propose to highlight these misunderstandings by first discussing the assumptions and agenda of the economic approach to the theory of the firm and then doing the same for the strategic approach. The result will be a synthesis of the approaches towards the end of the paper. In discussing each group of theories, we have chosen to move from the general to the specific, that is, we will discuss common assumptions and frameworks shared by all theories in a group (whether economic or strategic) and then move to a discussion of the specific theories that have developed from these perspectives. The discussion of general principles and assumptions is very important because these ideas often go unstated in specific theories and what is unstated, or not clearly stated, is often misunderstood or misrepresented (Williamson 1999). We apologize to the reader if some of the references in the following discussion seem dated, but some of these assumptions are very deeply rooted in the history of each respective discipline. We contend that many of the misunderstandings in the field have arisen precisely because of a lack of historical appreciation.

Economic Theories of the Firm

The set of economic theories of the firm is large, ranging from neoclassical theory to industrial organization (Seth and Thomas 1994), and from Adam Smith to Karl Marx (Putterman and Kroszner 1996). 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:

- the need for exchange arising from specialization;
- the need for coordination and cooperation among economic agents; and
- the assumption of efficiency in economic organization.

Each of these assumptions will be discussed in turn.

Specialization and Exchange

Since at least the time of Adam Smith (1776), economists have known that specialization via division of labor increases productivity (Kogut and Zander 1996). The wealth of society as a whole is increased by allowing people first to specialize in activities in which they have a comparative advantage and then to trade or exchange their output with other specialized producers.

For Demsetz (1995), a ‘firm’ is any individual, or group of individuals, that undertakes specialized production, that is, producing for exchange with others rather than for their own households. It follows that firms exist because they are able to produce goods more efficiently than households can. Contrary to the view of many strategists (Ghoshal and Moran 1996), 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, 1997).

Gains from trade can only be realized after exchange has occurred. Thus, economists have tended to focus on the transaction — the act of negotiating, executing and enforcing an exchange — as the most appropriate unit of analysis (Williamson 1996b). What is actually exchanged in a transaction? Most economists would now agree that it is a bundle of rights over an asset (broadly defined) that is being exchanged (Barzel 1989). Common rights that may be transacted include the rights to use, alter, transfer, alienate or appropriate income from an asset (Barzel 1989). An asset in this case includes labor, land, physical or financial 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 to developments in exchange mechanisms (Masten 1988). Over the centuries,
Arriving at a strategic theory of the firm

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 (Coase 1937).

One may distinguish between specific rights, which are explicit terms in a contract, and residual rights, which are not (Hart and Moore 1990). 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. The law (i.e. society) also determines the nature of the default or residual rights that remain after a transaction.

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 (Grossman and Hart 1986). 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 (Barzel 1989). As we shall see, these two rights are very important in determining the benefits and costs of organizing exchange within a firm.

Co-ordination and Co-operation

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 (Malmsen 1961).

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 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 co-ordination (Foss 1998). In the first instance, a firm (or entrepreneur) must perceive a market demand for a product. Secondly, a bundle of specialist inputs (or resources) must be identified that is capable of meeting the perceived market demand. Thirdly, 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.

The co-operation problem follows from the last point (Alchian and Demsetz 1972). 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 that 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 (Alchian and Demsetz 1972).

Problems of co-ordination and co-operation can be seen as types of information costs. In an economy with perfect information, there would be no co-ordination or co-operation costs (Demsetz 1997). The correct technology to meet consumer needs would be known, and the effort and level of co-operation 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 co-ordination and co-operation 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).

**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 (Williamson 1999). A corollary of this hypothesis is that all opportunities for gain have been exhausted. In an efficient economy, there is no better (i.e. more efficient) way of organizing production because otherwise firms in the economy would have implemented it (Williamson 1991, 1999). Of course, with hindsight, it may be possible to argue that resources could have been used more efficiently. For this reason, Williamson (1999) argues that efficiency only has meaning as an *ex ante* concept. That is, did the decision-makers make the best choice given the information and transaction costs that they faced? Of course, the question is tautological – if they could have made a better decision then they would have, therefore the actual decision made must have been the best decision.

The goal of organizational economics is not to find ‘better’ ways of organizing. Rather, it is to understand why certain choices (and hence organizational forms) appear (i.e. are efficient) in some contexts and not in others. Workers in the textile industry, for instance, are almost always paid on the basis of piecework rates, while college professors are hardly ever paid this way (Cheung 1969). Attempting to explain such differences in economic organization is the primary goal of organizational economics (Foss 2000). 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 other parts of the economy coordinate their transactions through spot, or market, contracts.

**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 appreciation for the reasoning involved. This will act as an aid to 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.

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
Arriving at a strategic theory of the firm

a power did not exist in a principal–agent agreement.

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 allocate resources effectively, 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 co-ordinate 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, 9)

Cheung suggested that all 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 (1972) observed 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 co-ordination 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. 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 approach. Like Cheung (1983), Williamson (1975, 1985) and Klein et al. (1978) 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 Klein et al. (1978) 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 or future state of the world).

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 hold-up its partner and appropriate the quasi-rents. Of course, if the partner firm could foresee this risk prior to negotiations, the deal would not proceed in the first place. Paradoxically, hold-up never occurs in Williamson's world (unless the parties factor the expected cost of hold-up into their profit calculations and still find the relationship potentially profitable).

The alternative for a firm with substantial quasi-rents is to integrate vertically into the activity of its partner, thus holding all specific assets itself and removing the threat of hold-up. The extent of a firm's integration is limited, in Williamson's theory, at the margin by the trade-off between governance costs and transaction costs.

Property rights approach. The property rights approach (Barzel 1989; Hart 1995; Hart and 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 and Hart 1986; Hart and 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 suggest 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 an asset. Many employees are in their highest valued use (i.e. produce the most quasi-rent) 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.

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 (Hart 1995).

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
Arriving at a strategic theory of the firm

c onsiderably 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.

Discussions on a strategic theory of the firm are relatively recent, dating from Rumelt’s (1984) paper “Towards a strategic theory of the firm” (although an intellectual tradition stretches back at least to the work of Penrose 1959). 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 homogeneous among strategic theories than those in economics are.

Common Principles

Strategists tend to agree on three broad principles:

- the resource-based nature of the firm;
- the determination of firm boundaries; and
- bounded rationality.

Each of these assumptions will be discussed in turn.

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 co-ordination and co-operation between people and other resources (Grant 1991; Schulze 1994). Capabilities include research and 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 and Schoemaker 1993). Strategists seek to create, and protect, rents in order to enhance the value of their firm. Resources are known to be more valuable if they are rare, difficult to imitate and non-substitutable (Barney 1991). However, in a perfectly competitive market, rents can only arise because of luck or differences in expectations concerning the future value of a resource (Barney 1986a). 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 its value (or opportunity cost) in the next best use its owner can find for it. The strategic value of a given resource is thus likely to depend on the way a firm combines, co-ordinates, and deploys that resource with other firm-specific and firm-addressable resources (Sanchez and Heene 1997). It is in the interest of the firm to use a new resource in ways that its owner (or 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 and 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 and 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 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 and 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, 1997). Firms are assumed to have the ‘blueprints’ for all production possibilities (Foss 1998a). 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 (Penrose 1959). 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 pay-offs from adopting various alternatives (Newell and 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 and 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 et al. 1997).

For Demsetz (1997), information costs are 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, 428)

Because they are uncomfortable with the mainstream economic view that resources are deployed in their most efficient use, strategists have tended to embrace alternative economic theories, such as Austrian economics and evolutionary economics, which share an assumption that the future is, to a greater or lesser degree, uncertain and unknowable (Hayek 1945; Nelson and Winter 1982). For these theories, firms are only relatively efficient (or inefficient) in reference to other firms. Firms adapt, but do not optimize or, more accurately, firms never know if they have achieved an optimal allocation of resources (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.

Knowledge. Knowledge theories of the firm (KTF) form the largest subset of the growing literature on strategic theories of the firm (Conner 1991; Conner and Prahalad 1996; Grant 1996; Kogut and 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, 112) goes so far as to say that "... all human productivity is knowledge dependent, and all machines are simply embodiments of knowledge". 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 and 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 do not 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 co-ordination. Firms are therefore more efficient than the market at integrating knowledge.

Conner and Prahalad (1996) focus on the advantages that administrative direction provides over market contracting. They note that a market contract needs to be renegotiated 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 and Robertson 1995, 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 and Robertson 1995, 35)

Transferring knowledge to suppliers also runs the risk that the suppliers might appropriate the gains from innovation (Langlois and Robertson 1995; Teece 1986). Appropriation can occur either in the form of hold-up (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 and Prahalad’s (1996) insight that flexibility is valuable to a firm (Barney and Lee 1998; Foss 1998b; Sanchez 1993, 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 (Brealey and Myers, 2000).

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 (Brealey and Myers, 2000). 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).

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 (Brealey and Myers 2000). 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 1998d, 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
Arriving at a strategic theory of the firm

the ways in which resources may be combined increases the value of options (Barney and 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 (Barney 1986b).

Criticisms of Existing Theories

Strategists have taken exception to the dismal view of the firm presented by economists; a view in which the firm merely exists as a tool for guarding shareholders from hold-up and opportunism. For Kogut and Zander (1996, 504), self-interested behavior is only one aspect of human motivation. There are emotions, such as those associated with friendship, empathy, and loyalty, and abstract values such as notions of good, beauty, and truth ... self-interest should be taken not as the limiting but the modal case of human motivation.

As we have discussed at length, strategists prefer to focus on beneficial issues of learning, communication, co-operation, and co-ordination – dynamic processes that appear to operate more easily within the firm than across market boundaries. Strategists will continue to find economic theories of the firm inadequate as long as they neglect these apparent benefits of firm organization.

In response, Williamson has argued that “... economizing is more fundamental than strategizing” (1991, 76) and “... the firm is advised to integrate only for ‘compelling reasons’ ... economizing on transaction costs is the ‘main case’” (Williamson 1999, 1090). Williamson (1991, 1999) does not completely dismiss the strategic aspects of the firm. Rather, he chooses to minimize their importance relative to transaction costs (particularly ex post transaction costs). For Williamson (1999, 1099), any attempt to relegate opportunism to a minor role is to engage in 'utopian fantasies'.

Foss (1999) has argued that a strategic theory of the firm can be reduced to just two propositions. First, firms exist because they create specialized assets that markets cannot duplicate; assets such as shared understanding, culture, reputation, identity, learning capabili-
ties and tacit knowledge (Conner and Prahalad 1996; Grant 1996; Kogut and Zander 1992). In a telling criticism, Nicolai Foss (1996a,b, 1997) has argued that “co-specialization [carries] no implications for ownership” (Foss 1996a, 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 1999, 738)

In essence, Foss is arguing that a team of contractors can also duplicate any capability that can be developed by a firm. This view is strongly endorsed by Alchian and Demsetz (1972), who argue that the firm is no more than a particularly dense nexus of contracts.

The second proposition that Foss (1999) believes underlies the strategic theory of the firm is the notion of flexibility. Managers within the firm have the flexibility to deploy resources quickly and decisively without haggling in the market to achieve a desired result (Conner and Prahalad 1996). This is an old argument that dates back to Coase (1937), who was the first to contend that firms exist because they lower the cost of resource allocation and co-ordination. Nevertheless, Foss (1999) is critical of the flexibility thesis because it cannot explain why the ownership of physical assets plays such a central role in defining a firm.

Arriving at a Strategic Theory of the Firm

The central contention of this paper is that we can only start arriving at a strategic theory of the firm after the differences between strategic and economic views of the firm have been reconciled. The need for reconciliation has been
recognized in the literature (Foss 1999; Kogut and Zander 1996; Williamson 1999) but, to date, proposals for reconciliation have been rather tepid; taking the view that economic theory should be 'informed' by strategy or vice versa. In this section, we will attempt a more decisive reconciliation of the two views. First, we recognize that firms and markets (and hybrids, such as joint ventures and strategic alliances) are merely tools for unlocking the perceived value from exchange. Secondly, every form of economic organization must balance the relative costs and benefits of each organizational form. Finally, efficient forms of organization must efficiently create and protect value.

**Firms and Markets as Tools**

In reading the extant literature in the theory of the firm, one could be forgiven for assuming that firms and markets are the only forms of economic organization and that decision-makers must make a binding decision to favor one form over the other. For economists, this typically means favoring the market and, for strategists, the firm. If, however, the problem of economic organization is placed in its wider context then this forced choice between firm and market appears to be something of a false dichotomy.

Both the firm and the market are *institutions* that have evolved over many centuries (North 1991). Both institutions have served as effective mechanisms for the movement of resources into more valuable (and valued) uses (Moran and Ghoshal 1999; North 1991). In this context, institutions are:

> ... the humanly devised constraints that structure political, economic, and social interaction. They consist of both informal constraints (sanctions, taboos, customs, traditions, and codes of conduct), and formal rules (constitutions, laws, property rights). [They] create order and reduce uncertainty in exchange and ... determine transaction and production costs. (North 1991, 97)

Firms and markets may be dominant forms of economic organization in the twentieth century but other choices have certainly existed in the past (including barter systems, fairs and trading caravans) and will continue to evolve in the future (virtual organizations, joint ventures, strategic alliances, franchises, etc.).

In the broader context, the problem of economic organization seeks to discover better ways of creating, allocating and distributing resources. By focusing on this larger question, theorists may avoid the false dichotomy between markets and firms and the overwhelming, if unproductive, impulse to take sides.

**Costs and Benefits**

Currently, invoking the institution of the firm guarantees certain rights and obligations to its owners. Chief among these are limited liability for shareholders, an indefinite legal life for the entity of the firm, the rights to residual income from the firm's activities, and the right to direct the firm's resources to alternative uses (Casson 1996; Hart 1995). As we have seen, these formal legal functions of the firm give rise to certain benefits such as enhanced co-ordination (Coase 1937; Conner and Prahalad 1996), the creation of strategic options (Sanchez 1998), and greater knowledge sharing (Grant 1996; Liebeskind 1996). The social nature of the firm also gives rise to several informal benefits not recognized in law, including the ability to establish a shared identity and culture (Kogut and Zander 1992), create social and psychological contracts with members (Rousseau 1995), and build a reputation and other non-tradable assets (Dierickx and Cool 1989).

There are also costs of using a firm, including the increasing costs of co-ordinating resources as the scale, scope and geography range of operations increases (Coase 1937; Richardson 1972). The cognitive limitations of senior managers also make the operation of large, diverse firms more costly (Prahalad and Bettis 1986) and the weak incentives within the firm make shirking likely, thereby
Arriving at a strategic theory of the firm

incurs monitoring and agency costs (Alchian and Demsetz 1972).

Like firms, markets also confer formal and informal benefits on their participants. Contract law creates the context in which exchange takes place in markets. The law defines what constitutes a fair exchange and punishes those who choose to operate outside its boundaries. Markets also have certain informal benefits. Market prices serve as signals of an imbalance in supply and demand. The presence of market prices for factors and products also facilitates economic calculation, making it easier to determine the value of new resource combinations (Lewin 1998). Finally, markets provide participants the freedom to interact with others in a relatively impersonal way, making it easier to enter and exit transactions.

Markets can also be costly. Ex ante transaction costs refer to the costs of searching, writing and enforcing contracts (Coase 1937), while ex post transaction costs refer to the threat of hold-up by exchange partners (Williamson 1991). Finally, dynamic transaction costs arise when potential exchange partners must be convinced that a potential exchange will be beneficial (Langlois and Robertson 1995). Market transactions cannot occur without mutual consent (Moran and Ghoshal 1999).

It is not necessarily the case that a benefit to the firm will automatically generate a corresponding cost in the marketplace. There may be benefits of firms that are incapable of being generated in the market at any cost and vice versa. The creation and sharing of knowledge through processes of shared meaning and identity may be a unique function of markets (Kogut and Zander 1996). The ability of the market price to provide incentives for market participants to ameliorate imbalances in supply and demand may be a unique function of markets (Hayek 1945).

Consequently, the choice of economic organization will depend on an estimate or calculation of the relative costs and benefits of each available form (see Figure 1). Firm organization will be favored if the benefits of organizing as a firm and the costs of transacting on the market are both high. Conversely, markets will be favored when market benefits are high and market costs are low. This addresses the strategists' concern that economic theories ignore the benefits of firm organization. The framework in Figure 1 also highlights the need for strategists to address the benefits of the marketplace and the relative costs of both firms and markets.

<table>
<thead>
<tr>
<th>Firm Benefits</th>
<th>Costs of Using a Market</th>
<th>Market Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Facilitates knowledge sharing</td>
<td>2. Ex ante transaction costs + 3. Ex post transaction costs</td>
<td>1. Price acts as a signal of imbalance in demand/supply +</td>
</tr>
<tr>
<td>2. Facilitates social controls</td>
<td>4. Dynamic transaction costs +</td>
<td>2. Price allows superior economic calculation +</td>
</tr>
<tr>
<td>3. Provides flexibility in resource allocation</td>
<td></td>
<td>3. Freedom to transact with any agent in the economy +</td>
</tr>
<tr>
<td>5. Builds intangible assets</td>
<td></td>
<td>= MARKET</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Costs of Using a Firm</th>
<th>= FIRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Co-ordination costs of scale, scope, and geography +</td>
<td></td>
</tr>
<tr>
<td>2. Agency costs</td>
<td></td>
</tr>
<tr>
<td>3. Cognitive limitations on information processing</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Integrating strategic and economic approaches
Value Creation and Protection

As Demsetz (1995, 1997) has noted, the alternative to the firm is not the market but the household. Firms exist whenever they can 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. Thus, opportunism is not strictly necessary to explain the existence of the firm.

However, the most efficient economic organization will be the one that best creates and protects value (Moran and Ghoshal 1999). Firms and markets are institutions that have evolved over a long period. Presumably, they are both quite efficient at creating and protecting value or else alternative mechanisms would have been developed. Moreover, key benefits of these institutions, such as property rights and contract law, are clearly devices that act to protect value.

Unfortunately, much of the strategy literature has chosen to highlight the virtues of the firm by neglecting or even denying the need to protect the firm from opportunism (Conner and Prahalad 1996). This position is surprising, since strategists have long appreciated the link between shareholder value and appropriability (Moran and Ghoshal 1999). More recently, resource-based theorists have been exhorted to pay more attention to protecting the value created by owning, or controlling, scarce resources (Coff 1997, 1999).

For their part, economists have tended to place great weight on the need to protect gains from specialization and have given relatively little attention to the process that generated those gains in the first place. Coase (1988), the original progenitor of transaction costs theory back in the 1930s, has decried the overwhelming emphasis on transaction costs in the economics literature:

"... the key idea in "The Nature of the Firm" [is] the comparison of the costs of coordinating the activities of factors of production within the firm with the costs of bringing about the same result by market transactions or by means of operations taken within some other firm. (p. 38)

Both sides could benefit from moving towards an accommodation of their opponent's views. Strategists should stop trying to develop special cases where opportunism does not apply and accept that dealing with opportunism (or the threat of opportunism) is important in most situations. For their part, economists should acknowledge that value creation is also a necessary part of the economic process that is worthy of extensive study in its own right.

Moran and Ghoshal (1999) have raised the interesting possibility that mechanisms for value protection may interfere with value creation (and vice versa). Studying the nature and effects of these trade-offs also represents a fruitful research direction. Further research directions are discussed in the next section.

Directions for Future Research

There is very little empirical research to support the various claims of competing theories of the firm. A recent study by Poppo and Zenger (1998) examined transaction cost, measurement cost and knowledge-based explanations for make-and-buy decisions. Although finding conclusive evidence for transaction-cost explanations, the authors wrote that the results "... suggest that a theory of the firm and a theory of boundary choice is likely to be complex, requiring integration of transaction cost, knowledge-based, and measurement reasoning" (p. 853).

The framework presented in Figure 1 represents a useful agenda for future research. The most pressing task is to estimate empirically the relative contribution of each factor in the model. Economists claim that economic organization is largely determined by market costs, whereas strategists claim that it is the benefits of firm organization that...
Arriving at a strategic theory of the firm

determine what exchanges will take place in firms. This, of course, is an empirical question and deserves to be investigated more fully.

A useful second step is to investigate how the factors governing economic organization change under different contingencies. For example, transaction cost economists have argued that the degree of asset specificity, uncertainty and frequency of transactions are important contingencies in determining integration (Williamson 1985). Similarly, strategists have argued that certain types of knowledge can be shared more easily in firms. One could also question whether the evolution of shared identity and culture develop more readily in some environments rather than others. Testing the extent and limitations of the claims that economic organization will change under different contingencies is a much needed area of research.

A third area of future research could focus on the process of adjustment between different institutional forms. Economists have assumed that the observed forms of organization are efficient, that is, existing arrangements are the best possible outcome given the information and constraints faced by agents in the economy. Accordingly, economists have tended to neglect the process of adjustment whereby one institutional arrangement is supplanted by another. Economists tend to engage in comparative industry studies; assuming that strong selection pressures will cause firms in a given industry to follow very similar institutional arrangements. On the other hand, strategists, including Poppo and Zenger (1998), have assumed that weak, rather than strong, selection mechanisms are at work. Consequently, there may be a range of institutional arrangements in a given industry and firms that perform well can be assumed to have superior forms of economic organization ceteris paribus. Given the large number of factors presented in Figure 1 that affect the choice of exchange mechanism, there could be quite a large variation in organizational practices.

The process of adjustment could be studied at several different levels. For instance, at the level of the economy, how does the introduction of a new technology, such as the Internet, affect the relative costs and benefits of various exchange mechanisms? Arguably, the Internet has lowered transaction costs and favored a shift to market forms of organizing. At the next level down, research could focus on how a given industry has responded to changes in the costs and benefits of using various exchange mechanisms over time. Finally, adjustments in economic organization could also be studied at the level of the individual organization. At all levels, researchers could divide their attention between the actual process of adjustment as well as the substantive outcomes of decisions.

Finally, there is a need for a fourth area of future research focused on hybrid and emerging forms of economic organization. A preoccupation with firms and markets has led to a paucity of explanations for institutional arrangements such as joint ventures, strategic alliances and virtual organizations. Earlier on, we used the term ‘hybrid’ organization to describe these institutions. This presupposes that these new forms of organization share characteristics of both firms and markets; however, this may be overlooking important new characteristics of these mechanisms or even masking the fact that these forms of organization may represent a third category of institutional arrangement that is separate from either firms or markets.

As the above discussion indicates, there is no classification or taxonomy of exchange mechanisms in the existing literature. sole traders, partnerships and corporations all share some of the characteristics of what we have been describing as firms. Is ‘the firm’, then, three related institutions or one unitary institution? Similarly, labor markets, stock markets and food markets share some or all of the characteristics of ‘markets’. A clear definition of terms would thus make identifying and classifying new forms of exchange mechanisms a much easier task.
Conclusion

According to Kirsten Foss (1998), 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 exist because they are efficient institutions for protecting these gains after they have been 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 among opportunistic agents. It is also more than just a superior way of co-ordinating knowledge and resources among co-operative agents.

Future theories of the firm need to acknowledge explicitly the costs and benefits of various exchange mechanisms in value creation and protection. Economists and strategists need to view their work as complementary rather than competing models. Only then will we start arriving at a comprehensive theory of the firm.

Notes

1 Address for correspondence: Steven E. Phelan, PO Box 830688, JO.51, Richardson, TX 75083, USA. Phone: (972) 883 2754; Fax: (972) 883 2709; e-mail: sphelan@utdallas.edu

2 Foss (1993) refers to them as contractual theories of the firm.

3 Such as both parties being competent when they made the contract, considerations were exchanged and nothing specifically illegal was contracted.

4 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.

5 In the sense of Pareto efficiency, where no one can be made better off without making someone else worse off.

6 That is, the firm has residual rights of control over its employees.

7 The value of a firm is the sum of its rents discounted at the firm's cost of capital.

8 A resource that lies outside the boundaries of the firm but can nevertheless be used by the firm.

9 Resources are usually assumed to be heterogeneous.

References


Arriving at a strategic theory of the firm


