Patents and Welfare in an Evolutionary Model

SIDNEY G. WINTER*
(US General Accounting Office, Washington, DC 20548, USA)

1. *Introduction*

This paper addresses two related problems, one at the level of general theory, the other a specific policy issue. The first is that of conceptualizing and modeling innovative opportunity. Any theoretical model of invention or innovation must include a model of inventive/innovative opportunity—though the model may be as simple as the assumption 'there is a single invention that could be made'.[1] The policy issue is the desirability of patents and the appropriate duration of patent protection. The question is whether very short-lived patents, or even no patents at all, might in some contexts yield higher economic welfare than the patent rights conferred under existing institutional arrangements.

The two problems are intimately related. A conceptual framework specific enough to provide a structured approach to answering the patent policy question is necessarily one that involves strong commitments regarding the nature of inventive opportunity. Yet the problem of modeling inventive opportunity in a realistic way—appropriate to the very real policy problem—is extremely challenging. Granted, in specific fields of activity, invention often follows identifiable 'trajectories' or 'paradigms' associated with the use of particular classes of heuristic methods for bringing about improvements (David, 1974; Dosi, 1984; Nelson and Winter, 1982; Pavitt, 1986; Rosenberg, 1969). Understanding of the trajectories being followed at a particular time may yield qualitative predictions about the nature of the improvements that are likely to be forthcoming in the near future. But when it comes to the question of the pace of advance, and particularly to the question of when a trajectory will top out, it becomes problematic to even identify the knowledge base that is relevant.

* This paper is based on a presentation delivered at the Conference on Innovation Diffusion held in Venice, Italy in March 1986. The present written version is a revision of one produced in the summer of 1988. The views expressed herein do not necessarily represent those of the US General Accounting Office, where the author serves as Chief Economist.

[1] This is the model in Arrow's classic paper (1962).

However difficult such assessments may be, they are logically indispensable to the economic case for any definition of property rights in productive knowledge. The case for recognizing and protecting such rights rests on the assumption that resources for advancing knowledge are scarce. The appropriate level of rents for such property rights depends on the scarcity of the resources and the benefits they can yield; we would like to know how much difference it makes, at the margin, if the quantity of resources devoted to advancing knowledge is a little higher or a little lower. This sort of question simply cannot be answered without reference to a model of inventive opportunity.

In section 3 of this paper, I describe a model of inventive opportunity employed in previous simulation work of Nelson and Winter; specifically, the model we referred to as the 'science based case'. Section 4 then describes briefly how the model simulated in the earlier work has been expanded into a model of industry evolution. Section 5 adds the patent policy instrument to the model, and sets forth the results of a simulation experiment comparing industry evolution with and without patent protection. Section 6 offers concluding comments appraising the significance of the results. Before proceeding to the main business of the paper, however, it will be helpful to review some of the theoretical background.

2. *Approaches to Modeling Innovation and Diffusion*

At the risk of reworking familiar ground, let me briefly characterize the models of innovation and diffusion that orthodox economists have explored over the past few decades. In the orthodox view, the production possibilities available to a firm, an industry or an economy at a particular time may be characterized by a production set—S_0, say. All techniques within S_0 are perfectly known. Initially, there are no other techniques. Alternatively, economic actors know that there are techniques not in S_0, and perhaps some of the characteristics of those techniques. They know nothing at all, however, about how to accomplish such results in practice.[2] At some subsequent time, one or more 'innovations' occur; this means that the production set becomes S_1, of which S_0 is a proper subset. The problem of the economics of *innovation* is to describe the role of economic factors in determining why S_1 is what it is, and why the change occurs when it does. The problem of the economics of *diffusion of innovation* is to explain (on economic grounds) why economic actors choose to adopt techniques in S_1–S_0 when they do. In particular, if it is assumed that the new techniques are economically superior at prevailing prices, the diffusion

[2] Rather than elaborate the evolutionary critique of this orthodox construct, I refer the reader to Nelson and Winter (1982, esp. Chapter 3), Nelson (1980), Winter (1982).

problem is to explain why every actor who faces this situation does not adopt such techniques immediately.

There is something inherently awkward about the notion of constructing an economics of innovation within the orthodox framework. If economic factors are relevant to the change of production sets; if resources can be applied with the intention of bringing about such change, what has happened to the notion that the production set itself is a 'given' and comprehensive description of technical possibilities? This difficulty has been dealt with in different ways over the years. In the age of golden ages (that is, when the old growth theory was in vogue), the puzzle was in most cases avoided by eschewing the subject entirely. Technical change of various sorts was usually assumed to fall from the heavens, or from the progress of science. It affected the economic system but was not in turn affected by the system. In other cases, higher level choice possibilities were conceived to provide the basis for an economic rationalization of innovative change: such devices as meta-production functions, innovation possibility frontiers, and the like come to mind. While some models addressed the problem of socially optimal investment in costly research, there was little attention to the question of how self-interested economic actors operating in competitive environments could cover the costs of innovation. This problem has recently received more attention in the 'new growth theory' (e.g. Romer, 1986, 1990).

Reflecting its origins in macroeconomic concerns, the modeling of technical change in the old growth theory at least had the virtue of addressing large-scale questions of considerable social and historical importance, and doing so in a way that established at least a few significant connections between theory and empiricism (Solow, 1970).

Subsequently, a substantial literature grew up that dealt with the theoretical economics of innovation. This literature consisted primarily of various elaborations and extensions of Arrow's seminal paper of 1962. In these models, the set S_1 typically differs from the set S_0 by one possible innovation, and the questions are whether this innovation will be made or not, when it will be made, who will make it, what the welfare consequences will be, and what market structure has to do with all of this. A parallel stream of activity starting from the same simple formulation of technological opportunity explored the problem of optimal patent life (Nordhaus, 1969; Scherer, 1972).

Barzel (1968), with an assist from Hirshleifer (1971), contributed tremendously to the interest of these questions by introducing countervailing considerations to the formerly presumptive inadequacy of incentives for innovation. This set the stage for a variety of theoretical analyses exploring the balance between the forces producing inadequate incentives and the forces providing excessive incentives. In these studies, the absence of first-best optimality was a foregone conclusion and the occurrence of an optimal second-best standoff between the different forces seemingly a low probability event (see Reinganum, 1988, for a survey covering much of this work).

Whatever the complexities they address in other respects, these models typically have an extremely stark and simple characterization of technological opportunity at the center of the story. They are 'one bit models' in the sense that only two possible production sets are envisaged. With efficient coding, it only takes one bit of information to describe which technological state of affairs obtains. This feature, by itself, is a considerable impediment to translation of the lessons of the models into realistic contexts.

The problem of modeling diffusion is, superficially, not as problematic in orthodox economic theory as the problem of modeling innovation itself. Theoretical manipulations can go forward on the assumption that 'the innovation' and 'adoption of the innovation' are well-defined concepts—and one can point to examples in the world where these assumptions seem reasonably justified. The basic puzzle of 'why doesn't everybody adopt immediately' can be solved in various ways that are within the restrictive bounds of the orthodox paradigm of optimization in equilibrium. For example, the answer 'because the potential adopters are in relevant respects not a homogenous population' has a lot of appeal on its face, and models of rational non-simultaneous adoption have been constructed on this foundation.

In the older literature that had its origins in rural sociology, the answer was 'because not everybody hears about the innovation right away from sources they consider reliable'. There is certainly nothing fundamentally irrational about the behavior portrayed in this sociological explanation. Concerns about the credibility of information sources are entirely appropriate in the real world, and economic actors adopt a variety of strategies to economize on the effort devoted to processing unreliable information. Because there are two types of error to be balanced, some valid information is undoubtedly rejected by these strategies. There is little doubt, I think, that both the heterogeneous actors models and the information diffusion models are highly relevant in explaining some real situations, and the fruitful questions to ask relate not to methodological issues but to the identification of the situations in which one or the other or both of the mechanisms are operative.

What is more to the point for my present purposes is the critique of the basic conceptual structure of 'diffusion of innovation' that arose, in large part, from the attempt to subject various models and hypotheses to empirical test. An important part of the critique is that 'adoption' of a pre-existing 'innovation' often seems very much like innovation itself (Downs and Mohr, 1976). The same set of facts can often be plausibly viewed in either way, depending on the extent of the difficulties that must be creatively overcome by an adopter. The greater those difficulties—or the greater the emphasis one chooses to place on them—the more the adopter looks innovative and the more the adoption looks like an innovation. The version of this ambiguity that bedevils the patent office and the courts is the question of whether result of an adopter's efforts to make a patented invention more useful to himself and others is (a) an infringement of the original patent if done without a license, (b) a patentable invention in its own right, (c) both a and b, (d) neither a nor b.

There is, in short, ample reason to think that the subjects of innovation and diffusion are not as separate as they are often made out to be, and that a unified theory is called for. The proper subject matter of this unified theory is the growth and diffusion of productive knowledge. Innovation itself is in part a process of diffusion, since it inevitably draws on pre-existing productive knowledge and diffuses that knowledge in a new form. What is needed is a much richer theory of productive knowledge and its changes.[3] What follows is a small contribution in this direction.

3. *Technological Opportunity and the Latent Productivity Model*

Embedded in the simulations of Schumpeterian competition that Nelson and I reported in our book (1982, Part V) is a simple stochastic, dynamic model of the way in which R&D effort produces innovations by drawing upon exogenously given technological opportunities. In terms of the survey above, this model contains as special cases the 'manna from heaven' models of standard neoclassical growth theory and the 'one bit' models of the later literature on incentives. (I refer here to the way in which technical change *per se* is modeled, not about the other features of these various models.) Particular versions emphasizing the stochastic process features were studied by Horner (1977); Iwai (1984a, b) did interesting work on a deterministic 'large numbers' limiting case of the model, unifying it to some degree with the deterministic model in Chapter 10 of my book with Nelson. Here, I am going to re-expound the model and its rationale, isolated from the complexities introduced by other features of the simulation model.

So far as the rationale is concerned, the first point worthy of emphasis is that innovative effort generally takes place in a social context that both sets goals for the effort and provides solution fragments upon which the innovator may draw in pursuit of those goals. In common with the great bulk of the

[3] My essay on production theory (Winter, 1982) puts forward a more substantial argument to this same effect.

theoretical literature, this model treats the motivational context as that of a for-profit firm facing constant input prices. Following further in the familiar tradition, the model abstracts from the details of elements upon which the innovator draws and of the processes by which a 'new combination' of these elements leads to innovation.

The results of successful innovative effort are represented by their direct economic consequences for the firm; further, these consequences are represented by a single number. In this mode, the single number is output per unit capacity (in a particular time period), and it is assumed that variable input coefficients are affected in the same proportion (Hicks neutrality).

Technological opportunity at a point of time is characterized by a probability distribution of possible alternative values of the productivity variable. Expenditures on innovative R&D efforts yield results in terms of draws from this distribution. More precisely, the probability of a draw occurring in a particular period is proportional to the expenditure level in that period.[4] R&D effort in the model, as in reality, is subject to a double uncertainty: (i) a given level of expenditure maintained for a period of time may or may not yield a technically successful development (a draw); (ii) even if a technically successful development is achieved, it may not be an economic success (the productivity level yielded by the draw may not surpass the firm's prevailing productivity level). Note also that this modeling approach implies in one sense constant returns to R&D effort: the expected number of technical successes achieved (draws) is proportional to R&D expenditure. Yet, in another sense, returns to R&D are diminishing: if the probability distribution of draw outcomes in constant, higher rates of expenditure yield results higher and higher in the probability distribution over time, and it thus becomes more and more improbable that a technical success achieved by future R&D will surpass what has already been achieved.

Within this general approach to modeling innovative R&D, Nelson and I worked primarily with a model that involves a specific commitment regarding the probability distribution: it is log normal in the productivity level, with a constant standard deviation over time. The central tendency of the distribution may change over time; it is characterized by the mean of the logs, denoted $L(t)$, or by $\exp[L(t)]$, called 'latent productivity'. In general, latent productivity may be presumed to be improving over time, as the fields of knowledge relevant to the industry are advanced by processes other than the

[4] The constraint that the probability of a draw not exceed one is avoided by an appropriate choice of parameter values. In this respect and others, the simulation model actually implemented is best thought of as a discrete time approximation to an underlying continuous time model, with the simulation model's time period chosen short enough so that single period binomial processes can adequately represent the Poisson processes of the underlying model.

industry's own research. A much more specific commitment is typically made for modeling purposes: $L(t)$ is increasing as a linear function of time; productivity growth is exponential.

$$L(t) = L_0 + \gamma(t - t_0) \tag{1}$$

The kinship of this formulation to neoclassical growth theory may be noted. If the standard deviation of the distribution were zero, and if 'draws' were costless, the assumptions made would bring us to the familiar Hicks-neutral, disembodied, exogenously determined, exponentially-rising productivity case of neoclassical growth theory. In this case, there is no economics of R&D to discuss and no interesting role for individual firms.[5] The interesting issues arise precisely because costly and uncertain activities undertaken by individual firms intermediate between exogenously given opportunities and actual realization of the economic benefits latent in those opportunities. The 'science based case' is distinctive in that the opportunities are strictly exogenous for the industry as a whole; the *only* role for private innovative R&D is to realize those opportunities. The case of 'cumulative' R&D explored in the earlier work is, by contrast, one in which there is no exogenous component to technological opportunity at the industry level, and from the individual firm's point of view the only options are direct imitation of other firms and the gradual, cumulative advance of its own competence. There are, obviously, many intermediate cases in which the innovative/imitative success of an individual firm depends in varying degrees on its own efforts, on the accomplishments of rivals, and on the expansion of technological opportunity from sources external to the industry.[6]

4. *Birth and Evolution of an Industry*

This paper extends the model presented in my 1984 paper, which was itself an extension of earlier work done with Nelson. The earlier Nelson–Winter work on Schumpeterian competition was concerned with evolutionary 'contests' among firms with different strategies, and also with the development of concentration in initially unconcentrated industries. For both purposes, it was appropriate to start the model from highly stylized and symmetric initial conditions. The phenomena of industry evolution, on the other hand, are largely a reflection of the fact that the situation of an industry as its founding firm

[5] As the literature amply illustrates, there is an economics of innovation to discuss in the 'one bit' models. These have their counterpart in our model of opportunity when the standard deviation is zero, latent productivity is *constant* at a level above prevailing productivity, and innovative R&D is costly.

[6] See Jaffe (1986) and Levin and Reiss (1988) for empirical explorations of the effects of technological opportunity and spillovers on R&D intensity and patenting.

appears is a very specific sort of situation and one that is typically remote from long run equilibrium. The same fundamental adjustment laws that would be operative in the vicinity of a long run equilibrium presumably drive the process from the start; it is the peculiarity of the initial conditions that fundamentally defines the subject matter.

In general, then, a theory of industry evolution must be based on a set of commitments regarding those special features of the data of the system that distinguish early periods from later periods. More specifically, a theory of industry evolution must include a theory of industry birth.

A plausible structure for such a theory of birth can be sketched by a series of linked observations. First, if industry birth is to be identified empirically with the appearance of the first firm(s) producing the product or products in question, the theory must include an account of this initial entry. If initial entry is not to be treated as a *sui generis* event—a methodological path that would impose severe limitations on the explanatory power of the resulting theory—it must be an instance reflecting a more general set of theoretical commitments regarding entry. Such is the case with the model here; the first building block to be described is the general model of entry.

Suppose a potential entrant has in hand a technique with productivity level A_E. He or she will choose to create a firm and go into production if it is anticipated that the product can be sold at a price P that covers production cost with something left over. This criterion is modeled as

$$P\,A_E\,u > c + r_E\text{, or} \tag{2}$$

$$\log A_E + \log u > \log\,(c + r_E) - \log P \tag{2a}$$

Here, u is a random variable reflecting the fact that it is not possible to assess the productivity of the technique completely accurately without creating the firm and trying it. This variable is assumed to have mean log zero (and specifically, in the simulations, to be log normal). The role of r_E in the model is to provide for entry decisions being influenced by some crude anticipation of the industry's future, as well as by the prevailing price—so that such decisions need not be thought of as totally myopic. This limited foresight might be derived, for example, from observation of the historical development of other new industries. What crucially distinguishes an evolutionary from an orthodox approach here is that r_E is not conceived as the result of an informed rational expectations calculation *specific to this industry*. Hence, nothing in the model will guarantee that the early development of the industry does not turn out to be a fiasco regretted by all entrants or, alternatively, a brilliant success that makes others say 'why didn't *we* do that?'

An important feature of this model of entry is that the entrant is assumed to confront the situation with a *technique* in hand. An alternative assumption would be that all the entrant has in hand is a way of acquiring a technique—an R&D strategy. Such a strategy might consist of an approach to acquiring a single technique that would make profitable entry possible, or it might involve an approach to generating a continuing sequence of innovations—constituting, in effect, a long-term plan for competing in the industry. While the 'technique in hand' approach seems more generally descriptive of entry processes, the area of biotechnology provides many contemporary examples of firms created primarily on the basis of the promise of their R&D resources rather than on the demonstrable profitability of innovations in hand.

Where does the potential entrant's technique in hand come from? Like the technical advances achieved by existing firms, it may be the fruit of innovative R&D or it may be the result of direct imitation of an existing firm. Unlike existing firms, however, potential entrants do not apply resources in the hope of discovering techniques (the alternative assumption would lead to the 'R&D strategy' entry model). Rather, the techniques are conceived to arise from a pool of continuing activity, some innovative and some imitative, that is financed and carried on without regard to the prospects for commercial success in the industry in question. This activity is called 'background R&D', and its mechanisms in the model are directly analogous to the R&D of firms in being, except that the activity level is constant over time, determined by model parameters conceived as equivalent expenditure levels. Innovative background R&D may be thought of as the fruits of the activity in, for example, university research settings, the workshops of individual inventors, or perhaps the laboratories of firms in other industries. Imitative background R&D may be thought of as the sort of diffuse search for revealed profit opportunities that implicitly underlies the usual idea of free entry, although in the present model such activity is conceived as costly and hence limited in amount in each time period. Only when background R&D generates a technique is the possibility of creating a firm considered, and only when a firm is created and entry occurs does the bookkeeping begin and the long run breakeven constraint come into play.

The founding firm must be an innovative entrant, there being no existing firm to imitate. Aside from that, the only feature that distinguishes the founder is the fact that it cannot look to a functioning market for the value of output price to use in the entry test. In many cases, uncertainty about the initial price might contribute significantly to errors in decision making, occasionally giving rise to 'premature birth' of an industry and perhaps to the early failure of the founder. Here it is assumed, however, that the novelty of the industry derives entirely from the technological basis of its production techniques, and not at all on the function performed by the industry's product. Specifically, it is assumed that there exists a perfect substitute for the

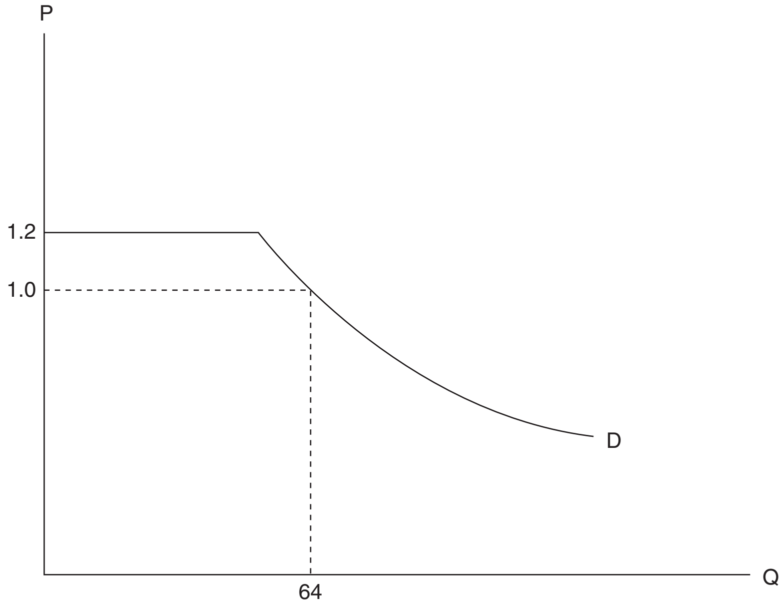

FIGURE 1. The specific demand curve underlying the simulations reported.

industry's product, and that the given price of this substitute translates into a critical price of P_0 per unit for the new product. Above P_0 there is zero demand for the new product, at P_0 precisely the market may be shared between the new product and the old; below P_0 the new product takes the entire market. The result is a demand curve of the general character shown in Figure 1.

In fact, Figure 1 portrays the actual demand curve employed in the simulation reported below. Units choices aside, its significant feature is that it is of unitary elasticity below P_0. Thus, industry sales will be constant once the price has fallen below P_0 (=1.2 in Figure 1). Continuing technological progress, although it reduces the prices of the product, neither increases nor decreases the aggregate size of the market. Informational economics of scale, which play a fundamental role in Schumpeterian competition and industry evolution, do not rise or decline in importance over time, once the industry has 'grown up' and price is below P_0. For this and other reasons, the special demand condition assumed is one in which the potential exists for the industry to move toward a stochastic growth equilibrium path with output, price and productivity changing at the rate of latent productivity growth. This property is interesting theoretically and a convenient simplification of the experimental context explored below, but its importance in the theory of industry evolution is that it is a special case marking the boundary between quite different evolutionary paths.[7]

The model of industry birth that emerges from the foregoing may be summarized as follows. Taking entry condition (2a), substituting $L(t) + \log v$ for $\log A_E$, taking $L(t)$ from (1), and using P_0 for P, we have the following condition:

$$L_0 + \gamma(t - t_0) + \log v > \log (c + r_E) - \log P_0 - \log u, \text{ or} \tag{3}$$

$$\log u + \log v > \log (c + r_E) - \log P_0 - L_0 - \gamma(t - t_0) \tag{3a}$$

The industry is founded at the first time when background R&D yields a technique 'ahead of the state of the art' ($\log v$ large) and this technique is sufficiently optimistically appraised ($\log u$ large) so that the combination leads to a decision to enter. Both $\log u$ and $\log v$ are assumed to be normally distributed, mean zero, and they are independent. There is a time t^* defined by

$$t^* = t_0 + (1/\gamma)\,[\log (c + r_E) - \log P_0 - L_0] \tag{4}$$

Time t^* may be interpreted as the time when it is an even money bet that a single draw will lead to a decision to enter, or, alternatively, as the time when the mean log revenue expected from the technique yielded by an entry draw just covers log costs per unit capital ($\log (c + r_E)$).

Figure 2 illustrates the determinations of the time of industry birth. The normal densities relevant to three times log t^1, t^* and t^2 are shown [to be interpreted as rising above the (t, $\log A$) plane]. The shaded areas show the probability that a single draw at the time in question will lead to the founding of the industry, if it has not been founded earlier. At t^1 the probability is small, at t^* it is 0.5, and at t^2 it is large. Of course, the expected date at which the industry is founded depends not only on the single draw probability but also on the frequency of draws—that is, on the level of background R&D. If, for example, background R&D is high, then the industry may come into being well before t^*, at a time when the probability that a single draw would lead to entry might be quite low.

After the industry is founded, additional innovative or imitative entry occurs to populate the industry with firms. The relative rates of the two forms of entry depend on, among other things, the levels of the two forms of background R&D and the existence or non-existence of patent protection.

An actual entrant must have not just a technique in hand, but also a capital stock and an R&D policy. In the model, new firms' capital stocks are drawn from a truncated normal distribution specified by model parameters. The assumptions with respect to R&D policies are more complex. The policies of the founder are

[7] The general condition characterizing this boundary situation is that the rate of latent productivity advance (γ), the elasticity of demand (η) and the growth rate of demand (α) combine in such a way that there is no trend in the amount of capital employed in the industry: $\alpha + \gamma(\eta - 1) = 0$. (Nelson and Winter, 1978, note 3, with a change of notation.)

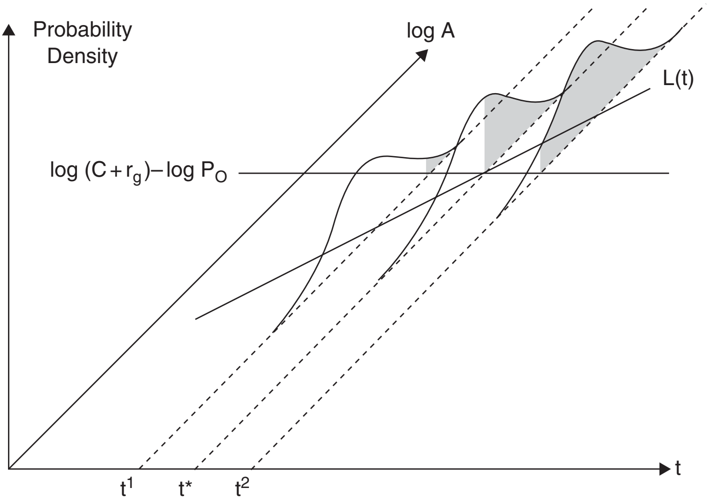

FIGURE 2. The determinants of the time of industry birth.

specified as model parameters. Subsequent entrants determine their policies by adding a random deviation to the capital-weighted average of the policies of existing firms. Firms in being that are persistently below average in profitability also change R&D policies; they add a random deviation to a convex combination of their previous policy and the industry average. Thus, through a combination of search and selection effects, industry average R&D policies respond to the industry environment and to the range of behaviors actually tried.

The position and size of that range, however, is part of the exogenous 'sociology' of the model rather than the economics. For example, the R&D policy specified for the founder will have a strong effect on the early development of the industry. If that policy is radically out of line with the realities of the environment, the industry's evolutionary path may be interpretable as a gradual repudiation of that policy. Such a case provides a good illustration of the contrast between an evolutionary and an orthodox approach. A decision made under uncertainty may be a mistake, by perfect information standards. Mistakes can have important lasting consequences. It is at least as interesting to explore how mistakes are slowly corrected by realistic selection and adaptation mechanisms as it is to circumvent the whole issue by some strong foresight assumption.

5. *Simulation of the Effects of Patents*

In the particular simulation runs described here, the effect of patents on welfare in a 'science based' technological regime is investigated. As patent protection is represented in the model, a patented production technique cannot be imitated by other firms during the life of the parent, here chosen as 3 years.[8] Nothing, however, prevents a firm from generating through its own innovative R&D a technique that yields a productivity level equal, approximately or exactly, to that of a patented technique. The patent system characterized in the model is thus a very strong one in the sense that it totally prevents other firms from benefiting from an innovator's R&D during the life of the patent, and it does so at no cost. On the other hand, the standards of 'Novelty' and 'utility' implicitly applied do not reach the question of whether an invention actually represents an economic advance over the prior art. Given the fact that techniques are described in the model only by the productivity levels to which they give rise, this approach seems appropriate and virtually inevitable.

More realistically, a patented invention might be thought of as occupying a region in a space of technological attributes; a firm seeking to invent around the patent must avoid encroaching upon that region regardless of whether the knowledge it draws upon is acquired by imitation or not, and regardless of whether its alternative is economically inferior or superior to the patented one. In general, the need to avoid the technological region occupied by the patent might be expected to raise the cost of making an economically comparable invention; on the other hand, the availability of some of the information disclosed in the patent might lower that cost (even given the need to avoid infringement). The net effect might, in reality, be approximately zero. The model treats it as precisely zero.[9]

After a patent has expired, the technique involved does not become instantly available to all other firms, but rather is subject to being acquired by them through costly imitative effort. As noted below, the need to incur these costs should be considered reflective of the innovator's continuing antipathy to being imitated and, relatedly, of the fact that the application for the now-expired patent involved less than perfect disclosure of the technique actually employed.

The welfare consequences of 3-year patents (versus no patents) are described here in terms of the percentage of 'reference surplus' achieved. Reference surplus is the total discounted present value of consumer's and producers' surplus, as measured from the ordinary supply and demand curves, that would be achieved under first-best socially optimal conditions. The calculation of this social optimum assumes that the only barriers to diffusion of technology are ones that are deliberately created to defend against the imitative efforts of

[8] The reason for the short patent life will be discussed subsequently.

[9] As noted in Levin *et al*. (1987, pp. 810–11), there is evidence that in some cases, at least, patenting reduces costs of duplication of a competitor's advance.

other firms; thus imitation costs are not a factor in the calculation of the social optimum. It assumes also that it is unfeasible to vary R&D intensity from time period to period according to the prevailing relationship between actual and latent productivity; industry R&D intensity is constrained to be constant over time. Reference surplus may thus be thought of as the surplus generated when there is a single firm, constrained to meet the demand at marginal production cost in every period, and operating at a constant R&D intensity calculated to maximize the present value of the total surplus.[10] The maximized total surplus involves a negative producer's surplus because pricing at marginal production cost implies that the firm does not recover its R&D costs. Finally, in both reference surplus and actual surplus calculations, costs associated with the advance of latent productivity and the existence of background R&D are ignored, and the calculation of reference R&D expenditure takes into account the fact that some R&D—background R&D—is available at no cost to the industry. This treatment is consistent with the interpretation that the technological opportunity that gives rise to the industry is a byproduct or 'spinoff' of other social activities pursued for other reasons, and similarly that background R&D is not financed for reasons involving specific anticipations of profitability in this industry.

The numerous parameter values that characterize these simulation runs are similar to those explored in Winter (1984); readers interested in the details of the runs should consult that paper and its appendix for further explanation of the model and interpretation of the parameters.

Among the particularly significant features of the numerical setting are the following. First, there is a relatively high rate of the latent productivity growth, 6% per year, and the standard deviation of the draws distribution corresponds to 3 years of latent productivity growth. Thus, a single draw has only about a 5% probability of yielding a technique that is more than 6 years ahead of the 'state of the art' as represented by the log latent productivity. (Note, however, that the industry may actually be following a track well above or below the latent productivity track). Second, the interest rate used in discounting surplus is 6%, which is also the private rate of return on capital. Discounting is done to the same absolute date regardless of when the industry is actually born; random differences in industry birth date are thus among the many sources of random variation in outcomes among runs.

[10] Actually, the R&D intensity used in the calculation of reference surplus in the runs reported is an approximation to the true optimum, and the approximation method used is biased toward understating the productivity of innovative R&D expenditure. It probably underestimates, therefore, both the level of optimized surplus and the optimal level of R&D. Since the runs compared in the present paper are all conducted under identical technological opportunity conditions (and hence have the same reference surplus value), the fact that the reference surplus value is itself approximate merely attaches a minor caveat to assessments of how big the run-to-run differences are.

Results

Table 1 summarizes the results of five runs under each parameter setting. The 3-year patents have a strong negative effect on total surplus, amounting to about 10% of reference surplus. This change is the net result of a decline of 25% of reference surplus obtained by consumers and an increase of 15% of reference surplus obtained by producers.

Additional comparisons between the two sets of runs are shown in Table 2; these provide some insight into the mechanisms producing the welfare effect. The large swing in producers' surplus shown in Table 1 indicates that, in a straightforward sense, patents are effective in protecting gains from innovation. Table 2 provides some weak indication that firms respond to this 'incentive' by increasing their R&D intensity. In the evolutionary model, of course, such a behavioral change is a reflection of selection and search effects operating on behavioral alternatives actually tried, rather than of *ex ante* calculations. That is, the change is partly a matter of superior growth by firms that choose high R&D intensity as they enter, and partly a matter of imitative, adaptive policy change by firms whose profitability performance is below average. Although the five run average changes in the right direction, the extent of run-to-run variation within experimental conditions indicates that neither adaptive learning nor

TABLE 1. Surplus With and Without Patents[a]

	Total	Consumers'	Producers'
No Patents	88.0	93.2	−5.2
	(7.2)[b]	(6.3)	(1.7)
3-Year	78.1	68.4	9.7
patents	(6.4)	(6.9)	(3.0)

[a] Five run averages, expressed as per cent of reference surplus.
[b] Figures in parentheses are standard deviations.

TABLE 2. R&D and Industry Structure[a]

	R&D Policy[b]	N_H[c]	R&D expenditure[d]
No patents	3.68	16.87	369.7
3-Year patents	4.48	9.21	302.9

[a] Five run averages.
[b] Capital-weighted average of firm policies (designed as expenditure/capital ratios), end of run
[c] Herfindahl numbers equivalent, end of run.
[d] Total (undiscounted) R&D expenditure over run.

selection is a particularly powerful mechanism shaping R&D policy—the feedback is very noisy, and the cost levels are in any case small in relation to sales.

More importantly, however, there is actually over 20% more total R&D performed, on average, without patents than there is *with* patents. Relatedly, best practice productivity levels tend to be higher without patents.

The greater volume of total R&D performed in the no-patents condition is a minor, and comparatively inconsequential, reflection of the major mechanism by which patents affect welfare in these simulations. The mechanism is to restrict imitative entry and lead to a generally less competitive industry—an industry that may be somewhat more R&D intensive per unit of capital employed, but involves fewer firms, employs less capital, produces less output, and earns substantial excess returns. Also, the restriction of imitation among existing firms leads to a substantially lower ratio of average to best practice productivity than is true in the no-patents condition; the welfare impact of this effect drawfs that of the different R&D intensity.

The data on average draws in Table 3 extend this story. Three-year patents, given the context of relatively rapidly advancing latent productivity, are sufficient to almost eliminate imitation as a method of technology transfer among firms. In the absence of patents, imitation accounted for 74% of the technique adoptions by existing firms, and 87% of the techniques adopted by entrants. In the presence of 3-year patents, these numbers change to 9% and 24%. The effect mostly takes the form of reducing successful imitation draws to very low levels; hardly anything on which the patent has expired is worth imitating. It is clear that much longer patent lives would make very little difference; even moderately longer patent lives would eliminate imitation entirely. The 3-year life was, in fact, chosen to preserve some small role for imitation.

Notice that there are substantially more innovations adopted under the patents condition—68% more in fact. However, this greater 'innovativeness'

TABLE 3. Innovation and Imitation Draws[a]

	Innovation		Imitation	
	Total	Accepted[b]	Total	Accepted[b]
Existing firms				
No patents	192	27	184	77
3-year patents	142	39	156	4
Potential entrants[c]				
No patents	52	4	105	26
3-year patents	55	13	100	4

[a] Five run averages.

[b] Implies change of technique if by an existing firm, entry if by a potential entrant.

[c] Excluding founder.

basically represents the necessary recourse to independent invention as a means of acquiring technology, given that patents block imitation. It is, in fact, a reflection of the inferior industry performance the patents cause: best practice is worse, not better; average practice is much worse still; more innovations are brought into use because more of them surpass these lower standards.

6. *Concluding Comments*

The simulation results illustrate vividly the possibility that the patent system can be counterproductive from a social welfare point of view. It is easy to imagine, nevertheless, that the system might have its defenders in the simulated world. That world presents strong incentives for self-interested advocacy of the patent system, and at least some of the arguments such advocates might advance—e.g. that patents lead to more innovations—are correct. From the social welfare point of view, however, their case as a whole is incorrect. First, it overlooks the possibility [originally emphasized by Barzel (1968)], that patents can cause inefficiencies by providing private incentives to seize exogenously improving opportunities at dates that are too early from a social point of view. A related inefficiency illustrated in the present model is that the process of diffusion of the knowledge represented by the exogenous opportunities may involve an increased proportion of expensive innovative R&D relative to cheap imitative R&D. Finally, and most importantly, the assessment of patents in the context of industry evolution (and using discounted surplus as the welfare criterion) underscores their possible adverse effects on industry structure and output.

Both the quantitative and qualitative patterns in the simulation outcomes reflect, of course, the assumptions of the model and the particular parameter values chosen. In considering what significance to attach to these numerical examples, it is important to distinguish two quite different questions: (i) are these results likely to be representative or typical of the actual welfare effects of the patent system as a whole? (ii) are these results illustrative of real mechanisms that may imply significant negative consequences for the patent system in some industrial contexts? My own responses are a tentative *no* for question (i) and a more vigorous *yes* for (ii). Both responses are consistent with the general view that there are large and important differences among industries with respect to the characteristics of productive knowledge, the sources of technological opportunity and the mechanisms of appropriability.[11] There is abundant reason to think that there are real situations that have little in

[11] For illustrative evidence in support of this point, see Levin *et al.*, 1987. The Yale survey partially reported in that report contains a good deal of additional evidence along the same line.

common with the simulated environment, little reason to think that the simulation outcomes correspond to the central tendency of real outcomes (whatever that might mean), but nevertheless good reason to think that the key features of at least some real environments may correspond sufficiently closely to the simulated world to make the simulation results relevant.

Four features of the simulated world principally account for the poor showing of the patent system. The first, of course, is the exogeneity of technological opportunity. Returns to innovative R&D are diminishing at any point of time as exogenously given opportunities are exhausted; this obviously limits the benefit from the patent system's stimulus to innovative R&D. Secondly, the assumed existence of background R&D means that some innovative R&D takes place, producing results sufficiently practical to motivate the founding of firms, regardless of whether the private returns are sufficient to cover the R&D costs. This reduces the marginal significance of the additional R&D that is funded in anticipation of net private gains. Relatedly, since it is background R&D that gives rise to the birth of the industry, the surplus transfers brought about by the patent system do not create effective incentives to found the industry earlier—this might conceivably be a significant source of benefit from the patent system in reality.

The two features just noted might be subsumed under the heading 'major influences from the broader social context of technological change'. The industry analyzed is not an island complete unto itself from a technological point of view. Perhaps some industries can reasonably be thought of as such islands, but certainly the theoretical literature has drastically (albeit implicitly) overemphasized the importance of these cases relative to industries whose technological development is significantly intertwined with that of the wider society.

A third feature underlying the results is the fact that the patents characterized in the model serve only to bar imitation, and do not serve as a basis for licensing agreements. (If they did, presumably such agreements would eliminate the inefficient substitution of innovative for imitative R&D.) This feature of the model may be rationalized by reference to a point made previously: the patents in the model do not fully disclose the technology; a prospective licensee needs the undisclosed technology as well as the licensed technology, and the prospective licensor is reluctant to strike such a deal because secrecy is considered more reliable than the legal protections of the patent system. Whatever the specific rationale, there are good theoretical and empirical grounds for thinking that in *some* sectors of the economy, patents function about as they do in the model.[12]

[12] Based on survey of high-level R&D executives, Levin *et al*. (1987) emphasize the diversity across industries in the working of the patent system and other mechanisms for protecting the gains from innovation. Overall, the executives scored patents as more effective in 'preventing duplication' than in 'securing royalty

Finally, the negative impact of patents on discounted total surplus is partly attributable to the slowing of the industry's early growth caused by the suppression of imitative entry. Much of this effect would not occur if the innovator-founder were assumed to enter at profit-maximizing scale, 'saturating' the market at P_0. This would require that the innovator-founder's decision rules reflect knowledge of the market demand at P_0; under the assumptions of the model this is not the case. Given this restriction, a tentative approach to initial entry scale seems plausible. With uncertainty about the demand at P_0, an orthodox model deriving entry scale from optimization calculations would probably yield a similar result: the more contenders for the market, the larger initial industry capital would tend to be.[13]

Changing these various assumptions would change and perhaps overturn entirely the simulation results. As noted above, while different assumptions might be more realistic for some industrial contexts, they would be less realistic for others.

Those who call for stronger protection of intellectual property—a common theme particularly among US policymakers in recent years—often seem to regard the desirability of such change as virtually axiomatic, perhaps on a par with the desirability of less crime. As economists are well aware, even the desirability of less crime is not axiomatic if the only available means to the end is the devotion of additional resources to law enforcement. The present paper emphasizes that the desirability of stronger intellectual property protection is far from axiomatic, even abstracting from the significant issue of enforcement. It illustrates the fact that specific features of the knowledge environment of an industry may be critical in determining how that particular industry is impacted by a policy change. A more flexible and discriminating approach to the modeling of technological opportunity is called for if economists are usefully to come to grips with these important issues.

References

Arrow, K. J. (1962), 'Economic Welfare and the Allocation of Resources for Invention,' in R. R. Nelson (ed.), *The Rate and Direction of Inventive Activity*, Princeton University Press: Princeton.

Barzel, Y. (1968), 'Optimal Timing of Innovations,' *Review of Economics and Statistics*, 50, 348–355.

David, P. A. (1974), *Technical Choice, Innovation and Economic Growth*. Cambridge University Press: London.

Dosi, G. (1984), *Technical Change and Industrial Transformation*. St. Martin's Press: New York.

income'. The belief that licensing could generally circumvent the adverse effects of patents on diffusion is one that fails to come to grips with the transactional difficulties in the market for intellectual property.

[13] Alternatively, the effect would at least be diminished if the imitative entry blocked by the innovator's patent were somehow transformed into innovative entry capable of 'inventing around' the patent. This would, of course, be a departure from the model's 'background R&D' story of how innovative entry occurs.

Downs, G. W. and L. B. Mohr (1976), 'Conceptual Issues in the Study of Innovation,' *Administrative Science Quarterly*, 21, 700–712.

Hirshleifer, J. (1971), 'The Private and Social Value of Information and the Reward to Inventive Activity,' *American Economic Review*, 61, 561–574.

Horner, S. (1977), *Stochastic Models of Technology Diffusion*. Unpublished dissertation, University of Michigan.

Iwai, K. (1984a), 'Schumpeterian Dynamics: Part I, An Evolutionary Model of Innovation and Imitation,' *Journal of Economic Behavior and Organization*, 5, 159–190.

Iwai, K. (1984b) 'Schumpeterian Dynamics: Part II, Technological Progress, Firm Growth and "Economic Selection" ', *Journal of Economic Behavior and Organization*, 5, 321–351.

Jaffe, A. B. (1986). 'Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits and Market Values,' *American Economic Review*, 76, 984–1001.

Levin, R. C. and P. Reiss (1988), 'Cost-reducing and Demand-creating R&D with Spillovers,' *Rand Journal of Economics*, 19, 538–556.

Levin, R. C., A. K. Klevorick, R. R. Nelson and S. G. Winter (1987), 'Appropriating the Returns from Industrial Research and Development,' *Brookings Papers on Economic Activity: Special Issue on Microeconomics*, 783–820.

Nelson, R. R. (1980), 'Production Sets, Technological Knowledge and R&D: Fragile and Overworked Constructs for Analysis of Productivity Gowth?' *American Economic Review*, 79, 62–67.

Nelson, R. R. and S. G. Winter (1978), 'Forces Generating and Limiting Concentration under Schumpeterian Competiton,' *Bell Journal of Economics*, 9, 524–548.

Nelson, R. R. and S. G. Winter (1982), *An Evolutionary Theory of Economic Change*. Harvard University Press: Cambridge, MA.

Nordhaus, W. D. (1969)*, Invention, Growth and Welfare: A Theoretical Treatment of Technological Change*. MIT Press: Cambridge, MA.

Pavitt, K. (1986), ' "Chips" and "Trajectories": How Does the Semiconductor Influence the Sources and Directions of Technical Change?', in R.M. MacLeod (ed.), *Technology and the Human Prospect: Essays in Honour of Christopher Freeman*. Frances Pinter: London.

Reinganum, J. (1988), 'The Timing of Innovation: Research, Development, Diffusion,' in R. Schmalensee and R. Willig (eds), *Handbook of Industrial Organization*. North-Holland.

Romer, P. M. (1986), 'Increasing Returns and Long-run Growth,' *Journal of Political Economy*, 94, 1002–1037.

Romer, P. M. (1990), 'Endogenous Technological Change,' *Journal of Political Economy*, 98, S71–S102.

Rosenberg, N. (1969), 'The Direction of Technological Change: Inducement Mechanisms and Focusing Devices,' *Economic Development and Cultural change*, 18, 1–24.

Scherer, F. M. (1972). 'Nordhaus Theory of Optimal Patent Life: A Geometric Interpretation,' *American Economic Review*, 68, 422–427.

Solow, R. M. (1970), *Growth Theory: an Exposition*. Oxford University Press: New York.

Winter, S. G. (1982), 'An Essay on the Theory of Production,' in S. H. Hymans (ed.), *Economics and the World Around It*. University of Michigan Press: Ann Arbor.

Winter, S. G. (1984), 'Schumpeterian Competition in Alternative Technological Regimes,' *Journal of Economic Behavior and Organization*, 5, 287–320.