Economic Theory **14**, 113–130 (1999)

Economic Theory

A formal model of theory choice in science*

William A. Brock and Steven N. Durlauf
Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706 USA (e-mail: Sdurlauf@ssc.wisc.edu)

Received: April 8, 1997; revised version: April 30, 1998

Summary. Since the work of Thomas Kuhn, the role of social factors in the scientific enterprise has been a major concern in the philosophy and history of science. In particular, conformity effects among scientists have been used to question whether science naturally progresses over time. Using neoclassical economic reasoning, this paper develops a formal model of scientific theory choice which incorporates social factors. Our results demonstrate that the influence of social factors on scientific progress is more complex than previously thought. The patterns of theory choice predicted by the model seem consistent with historical episodes of theory change.

Keywords and Phrases: Interactive discrimination, Evolution of beliefs, Scientific progress.

JEL Classification Numbers: A12, D7.

1 Introduction

> "(S)tudents of the development of science, whether sociologists or philosophers, have alternatively been preoccupied with explaining consensus in science or with highlighting disagreement and divergence. Those contrasting approaches would be harmless if all they represented were differences of emphasis or interest...What creates the tension is that neither approach has shown itself to have the explanatory resources for dealing with both...whatever success can be claimed by each of these models in explaining its own preferred

* The Vilas Trust, Romnes Trust and National Science Foundation have generously provided financial support. This paper was written while Durlauf was visiting the Santa Fe Institute, whose support and hospitality he gratefully acknowledges. We thank Kim Sau Chung and participants at the Notre Dame Conference on the New Economics of Science for comments on an earlier draft.

Correspondence to: S. N. Durlauf

> problem is largely negated by its inability to grapple with the core problem of its rivals.''
>
> Larry Laudan[1]

Since the classic work of Feyerabend (1975) and Kuhn (1970), analyses of scientific progress have been required to consider the socioeconomic environment in which scientific analysis is conducted. Whereas the predominant themes in the philosophy of science for most of the twentieth century have emphasized the development of normative criteria for the scientific process, and generally presumed from the positive perspective that scientists each possessed an identical desire to find ''truth'' as measured by these common criteria, recent trends both in the philosophy and history of science (Bloor, 1976; Latour, 1976) have been concerned with the social context in which research is conducted. This concern represents a challenge to standard accounts of science as a progressive enterprise, in which changes in accepted theories construct a trajectory towards greater verisimilitude (Popper, 1972, 1976; Newton-Smith, 1975), problem-solving capability (Laudan, 1977, 1984), or other criterion or set of criteria for theory evaluation. The basic source of this concern is the belief that those motivations of scientists which are explicitly social, be they a desire for status, success, or conformity, can lead the evolution of science away from whatever criteria constitute the appropriate goals of science.

As our quotation from Laudan indicates, much of the debate concerning positive models of scientific change can be dichotomized between those who primarily focus on conditions which support consensus about goals and methods as the norm and those who primarily study conditions under which disagreement can occur. These differences tend to mirror the distinction between a vision of science as an activity conducted by disinterested truth-seekers and one conducted by fallible and self-interested participants. A difficulty with these alternative assumptions is that they restrict the explanatory scope of their associated theories.

In this paper, we propose a model of theory choice which explicitly incorporates private as well as social influences on individual decision-making. We do this by employing a model based upon Brock and Durlauf (1995) which characterizes the behavior of binary choices in an interdependent population.[2] While the model is of course very stylized, it does provide a formalization of the sort of social environment in which theory choice occurs. In terms of antecedents, our work follows closely in the spirit of the pathbreaking study by Kitcher (1993) which formalizes science as a dynamic process in which individual scientists interact as purposeful actors.

---

[1] Laudan (1984, p. 2).

[2] See Blume (1993), Brock (1993), and Durlauf (1993) for related analyses of interaction structures and Oomes (1997) for the use of the basic framework we describe to study related issues.

At the outset, we wish to identify a number of assumptions which we employ although they are extremely controversial from the perspective of the philosophy of science. First, we assume that there is a unique scalar metric by which one theory can be judged as scientifically superior to another. Not only is there controversy between different criteria for scientific theory assessment, it seems clear that the perceived success of different theories depends upon distinct criteria (Putnam, 1994). However, since our concern is with the interaction of "social" versus "scientific" criteria for theory acceptance, the problems with this assumption do not seem to be germane.

Second, we focus on a single type of social interaction – conformity effects, by which we mean the tendency of individual scientists to place greater weight on theories which others accept than otherwise. This assumption does not by itself imply the presence of nonscientific factors, since it is certainly possible that such weighting reflects incorporation of the scientific assessments of others. However, for our purposes, we interpret conformity exclusively as a nonscientific influence. Of greater concern is the possibility that this type of nonscientific influence loses some of the richness of the literature. In particular, we omit issues of financial support and ability to publish. Both of these factors introduce the issue of how a theory is able to generate evidential support, either directly through additional research or indirectly through altering the information set employed in research by others. While our intuition is that adding such factors would not qualitatively affect our results, this has not been formally shown and would certainly represent a useful extension of our model.

Our formal model possesses two especially interesting properties. First, our model shows how scientific consensus can rapidly emerge from a period of profound disagreement. In particular, we show how social interactions in the scientific community are an essential component of this process. Second, we demonstrate that social interactions do not necessarily represent, as is often assumed in the philosophy and (especially) sociology of science literatures, an impediment to the adoption of new and better theories over their entrenched predecessors. In fact, these social influences may actually accelerate the rate at which superior theories achieve a consensus.

Both of these features stem from the way in which scientific and nonscientific influences interact. These interactions produce nonlinear effects of the sort which frequently arise in the study of complex systems. In this regard, we believe our analysis highlights the way in which formalization of arguments in the philosophy literature can lead to useful insights.

2 Framework

We consider the problem of theory choice between two theories, denoted as T_{-1} and T_1. Binary decisions of this type have been studied extensively in the economics literature. In particular, we consider a community of I scientists. Individual scientists are indexed by i; scientists i's theory choice at time t is

wi, t with associated support $\{-1, 1\}$. The collection of theory choices by all scientists in the community is $\underset{\sim}{\omega}_t$. Finally, the vector of all decisions other than that of agent i is $\underset{\sim}{\omega}_{-i,t}$.

Each scientist is assumed to possess a way of assigning numerical valuations to the adoption of a particular theory, which we will refer to as "utility."[3] The utility a scientist receives from adoption of a particular theory is assumed to be measured by a function $V_{i,t}(\omega_{i,t})$.[4] Therefore, an individual scientist's theory choice is the solution to the maximization problem

$$\max_{\omega_{i,t} \in \{-1,1\}} V_{i,t}(\omega_{i,t}) \quad . \tag{1}$$

In order to permit the analysis of the effects of private and social influences on theory choice, we place some restrictions on the utility function $V_{i,t}$.

In particular, we assume that the individual utility can be decomposed into three components

$$V_{i,t}(\omega_{i,t}) = u_{i,t}(\omega_{i,t}) - E_{i,t} \sum_{j \neq i} J_{i,j,t}(\omega_{i,t} - \omega_{j,t})^2 + \epsilon(\omega_{i,t}) \tag{2}$$

In this specification the three additive components refer to different types of utility. Specifically, $u_{i,t}(\omega_{i,t})$ represents deterministic private utility, $-E_{i,t} \sum_{j \neq i} J_{i,j,t}(\omega_{i,t} - \omega_{j,t})^2$ represents a general conformity effect, which we call deterministic social utility, and $\epsilon(\omega_{i,t})$ represents random private utility. Notice that the first and third components are idiosyncratic in the sense that they depend only on the individual's characteristics, whereas the second component is determined by the individual's characteristics as well as the (expected) choices of the rest of the population. We therefore interpret the first and third components as embodying the scientific judgments of each scientist and the second component as embodying the influence of social factors, namely conformity effects, on choice. While the individual-specific components to theory assessment could plausibly be argued to contain judgment factors which are nonscientific in nature, none of our results are fundamentally changed by this interpretation.

In formulating the individual scientist's decision problem this way, our analysis in some respect sidesteps Kuhn's (1970) argument that different scientific paradigms may be incommensurable due to different ontologies or

[3] Our use of the word "utility" to characterize the evaluation function for individuals carries no implications concerning the evaluative criteria of scientists. Also, notice that our use of a utility maximizing framework has no implications for whether theory choice is active or passive, in the sense that while choice of goods in a grocery store is active, the determination of whether one is liberal or conservative might well be passive. In the passive case, our utility function determines which attribute is possessed by a scientist. Both active and passive elements to choice are presumably present in actual theory choice.

[4] See Diamond (1988) for the use of a utility maximizing framework to understand how scientists allocate time across theories, with resultant implications for the rationality of the scientific enterprise as a whole.

epistemologies or Quine's (1951) related argument that theories are underdetermined by data. We do this as we require of scientific theories not that one can be interpreted in the ontology of another or that one theory explain phenomena that another cannot, but rather that relative to whatever goals of science a community embraces, relative theory evaluation can be made. Such an assumption is commonplace in economic models of consumer choice in which individuals have preference orderings over different bundles of commodities, such as guns and butter, in which there is no intrinsic comparability between the individual commodities.[5]

Finally, the random utility components are assumed to be extreme-value distributed and independent across individuals, so that

$$\text{Prob}(\epsilon(-1) - \epsilon(1) \leq z) = \frac{1}{1 + \exp(-\beta z)}; \quad \beta > 0 \ . \tag{3}$$

For our purposes, this assumption is used for analytical convenience as it allows us to explicitly characterize the probability measure of $\underset{\sim}{\omega}_t$. Anderson, dePalma, and Thisse (1992) provide some interpretations of and justifications for this functional form in the context of the theory of consumer choice. Observe that β indexes the degree of diversity of individual-specific theory evaluations in the community. Small values of β imply that there is wide diversity, whereas large values of β imply there is little diversity.

These assumptions are sufficient to imply (following formal arguments developed in Brock and Durlauf, 1995) that at each t, scientist i's theory choice possesses the probability structure

$$\text{Prob}(\omega_{i,t} | E_{i,t}(\underset{\sim}{\omega}_{i,t}))$$

$$\sim \exp\left(\beta h_{i,t}\omega_{i,t} + \sum_{j \neq i} \beta J_{i,j,t}\omega_{i,t}E_{i,t}(\omega_{j,t})\right) \ .^{6} \tag{4}$$

In this formulation $h_{i,t} = \frac{1}{2}(u_{i,t}(1) - u_{i,t}(-1))$ and so provides a sufficient statistic for the private deterministic component of the comparative evaluation of the two theories. Note that incommensurability of scientific evaluative criteria can explain differences across individuals in $h_{i,t}$, in the sense that different individuals may assign different scientific weights to theories as a result of differences of beliefs concerning factors such as which phenomena

[5] Laudan (1996), chapters 1–3, develops the related argument that while theories may be deductively underdetermined, theories may still be "ampliatively" determined, by which he means that nondeductive factors such as simplicity, explanatory scope, or relationship with other theories, may nevertheless imply that a single theory dominates another. In the context of our utility specification, these ampliative factors are subsumed in the V function. See Hands (1994) for a related discussion of the relationship between Laudan's approach to modelling scientists' decisionmaking and economic formulations of utility maximization.

[6] The term "$\sim$" means "is proportional to" and is employed to avoid the need to write cumbersome normalizations.

are most important for a theory to explain or the mechanisms by which theories are evaluated.

Since the random components of the individual utility functions are independent, the collection of theory choices is characterized by the joint probability structure

$$\text{Prob}\left(\underset{\sim}{\omega}_t | E_{1,t}\left(\underset{\sim}{\omega}_{-1,t}\right), \ldots, E_{I,t}\left(\underset{\sim}{\omega}_{-I,t}\right)\right) = \Pi_i \text{ Prob}\left(\omega_{i,t} | E_{i,t}\left(\underset{\sim}{\omega}_{-i,t}\right)\right) \sim$$
$$\Pi_i \exp\left(\beta h_{i,t}\omega_{i,t} + \sum_{j \neq i} \beta J_{i,j,t}\omega_{i,t}E_{i,t}(\omega_{j,t})\right) \tag{5}$$

Our model of the evolution of theory choice will be complete once we specify the properties of $\beta, h_{i,t}, J_{i,j,t}$, and $E_{i,t}(\cdot)$.

3 Leading case analysis

i. Model specification

We initially consider the case in which all dynamics are determined for fixed evaluation weights. Formally, this means that there exists a constant J such that

$$J_{i,j,t} = \frac{J}{I-1} \tag{6}$$[7]

and a constant h such that

$$h_{i,t} = h \tag{7}$$

This means that any differences in the way in which members of the community make scientific evaluations of the competing theories are embedded in the $\epsilon_{i,t}(\underset{\sim}{\omega}_{i,t})$'s. Following Brock and Durlauf (1995), it is straightforward to show that

$$\text{Prob}\left(\underset{\sim}{\omega}_t | E_{1,t}\left(\underset{\sim}{\omega}_{-1,t}\right), \ldots, E_{I,t}\left(\underset{\sim}{\omega}_{-I,t}\right)\right)$$
$$\sim \Pi_i \exp\left(\beta h \underset{\sim}{\omega}_{i,t} + \beta J \omega_{i,t} m^e_{i,t}\right) \tag{8}$$

where

$$m^e_{i,t} = \frac{1}{I-1}\sum_{j \neq i} E_{i,t}(\omega_{j,t}) \tag{9}$$

The properties of this expression can be fully understood in two steps. First, assume that all scientists possess common expectations of the mean choice of others.

$$m^e_{i,t} = m^e_t \quad \forall\, i \tag{10}$$

[7] The $I-1$ which appears in the denominator of this expression acts as a normalization factor.

Under this assumption, one can verify that the sample average choice level, $\bar{m}_{I,t}$, will, as the number of scientists becomes arbitrarily large, obey

$$\lim_{I \Rightarrow \infty} \bar{m}_{I,t} = \tanh(\beta h + \beta J m_t^e) \tag{11}$$

Second, assume that the common expectation of the average choice is self-consistent in the sense that the common expected average choice level is one that is actually realized,

$$m_t^e = \lim_{I \Rightarrow \infty} \bar{m}_{I,t} \ . \tag{12}$$

Self-consistency, combined with equation (11), implies that the steady state behavior of the average choice level is any value m^* such that

$$m^* = \tanh(\beta h + \beta J m^*) \tag{13}$$

The solution m^* to (13) depends only on β, h and J.

This simple equation provides some insights into the interaction of private and social influences on theory choice. In particular, following Brock and Durlauf (1995), the following will hold:

Proposition 1. Existence of multiple versus unique steady states

Under the assumptions of the "leading case"

i. If $\beta J < 1$, there exists a unique solution to (13).

ii. If $\beta J > 1$, there exists a threshold H, (which depends on βJ) such that

a. for $|\beta h| < H$, there exist three roots to eq. (13), one of which has the same sign as h, and the others possessing opposite sign.

b. for $|\beta h| > H$, there exists a unique root to eq. (13) with the same sign as h.

In the multiple steady state case, we will denote the three equilibria as m_-^*, m_m^* and m_+^* in order to distinguish the extremal equilibria by sign.

Proposition 1 is interesting, as it provides a precise relationship between the strength of individual and social factors in determining whether the average theory choice of a scientific community is or is not a unique function of the set of scientific evaluations of the individual scientists. In particular, it illustrates that for any strength of conformity effects (as measured by J), there is a level of evidential support (as manifested in h) such that two communities of scientists will come to similar average conclusions on the relative merits of two theories. At the same time, for any level of evidentiary support h, there is some conformity effect level J such that social interactions can lead a community consensus away from that theory which by scientific criteria is superior.

This feature has four implications for debates on the progressiveness of science. First, it suggests that the consequences of the introduction of social factors in the choice of scientific theories has no necessary implication for claims concerning the progressiveness of the evolution of theories.

Second, so long as there exist sufficiently decisive evidentiary differences between two theories (which in our model is equivalent to a sufficiently large

value of $|h|$), strong social factors do not act to the detriment of progress in scientific models. Notice that this conclusion is not dependent upon any assumption about whether evidentiary support is theory neutral or not nor does it depend upon commensurability between theories (in terms of common or translatable definitions and objects of explanation) per se. The proposition only depends on the capacity of the scientific community to engage in relative theory evaluation.

Third, it suggests that the analysis of theory progression needs to distinguish between "local" and "global" progressiveness. What we mean is the following. The relationship between h and βJ means that so long as the relative scientific merits of two theories are large enough, social factors will never impinge on emergence of the consensus around the superior theory. However, for any theory and $\beta J > 1$, there will exist a local alternative theory (defined as one in which is h "small" but negative) which represents a dominant steady state choice for the community despite the presence of the superior alternative.

Fourth, the model suggests that the diversity of evaluative criteria and/or evaluative evidence within the population has important implications for the progressiveness of science. In particular, observe that a small enough β will, for any strength of social interactions, eliminate the multiplicity of steady states. Intuitively, a small β means that the dispersion of private theory evaluations is high, in the sense that a substantial fraction of individual theory choices is driven exclusively by private considerations, i.e. the relative scientific differences between the theories is so large that any conformity effect is overwhelmed. The presence of such "extreme" private beliefs in turn means that for the remaining members of the community, there will be insufficient capacity for consensus to allow for multiple steady states.

The qualitative results of the baseline model are robust to the assumption that all scientists agree on the relative scientific merits of the two theories. To see how the basic model can be generalized to account for fixed individual heterogeneity (as opposed to the heterogeneity associated with $\epsilon(\omega_{i,t})$), we simply reintroduce distinct values of $h_{i,t}$ for each scientist. In this case, for any common set of expectations, average theory choice will obey

$$\lim_{I \Rightarrow \infty} \bar{m}_{I,t} = \int \tanh(\beta h_{i,t} + \beta J m_t^e)\, dF_{h_{i,t}} \quad . \tag{14}$$

where $dF_{h_{i,t}}$ denotes the probability measure characterizing $h_{i,t}$. Imposing self consistency implies that the average choice level at t is any value m_t^* such that

$$m_t^* = \int \tanh(\beta h_{i,t} + \beta J m_t^*) dF_{h_{i,t}} \tag{15}$$

It is easy to see that when the $h_{i,t}$ values all possess the same sign, there will be a unique self consistent solution when the magnitude of the $|h_{i,t}|$ are sufficiently large.[8]

[8] This follows immediately from the continuity and monotonicity of the $\tanh(\cdot)$ function.

4 Dynamics

In order to analyze the dynamics of the baseline model, we make the following two assumptions

$$m_t^e = m_{t-1} \tag{16}$$

Imposition of eq. (16) may be interpreted either as meaning that expectations of community beliefs are adaptive, so the expectation of the mean today is whatever transpired the previous period, or that theory choice at t is influenced by a desire to conform to average beliefs at $t-1$. Note that since we assume that the idiosyncratic $\epsilon(\cdot)$ terms are invariant (except for the choice of $\omega_{i,t}$), individual scientists do not flip opinions randomly once a steady state average has been achieved.

This assumption on expectations means that the dynamics of the model are described by the sequence of m_t values consistent with

$$m_t = \tanh(\beta h_t + \beta J m_{t-1}) \ . \tag{17}$$

i. Stability

Brock and Durlauf (1995) verify that under these dynamics, the extremal equilibria are stable in the multiple equilibrium case, whereas the middle equilibrium is not.

Proposition 2. Stability of steady state mean choice levels

Under the assumptions of our "leading case",

i. If eq. (13) possesses a unique root, that root must be locally stable.

ii. If eq. (13) possesses three roots, then the steady state mean choice levels m_-^* and m_+^* are locally stable whereas the steady state mean choice level m_m^* is locally unstable.

With reference to theory choice, Proposition 2 means that social factors can lead a community to make collective choices which on average differ from the choices which would be made due to purely "scientific" factors. Such choices will depend on the initial distribution of theory choices in the community. This establishes a type of path dependence in the evolution of theory choices.

ii. Nonlinearity

This analysis of stability is incomplete in the sense that it treats the relative scientific content of the theories (h) as fixed and looks at the dynamics of the mean conditional on this. Of course, scientific theories evolve in the presence of changes in evidence and associated theories. Hence, we consider how the steady state average theory choice evolves in response to changes in h.

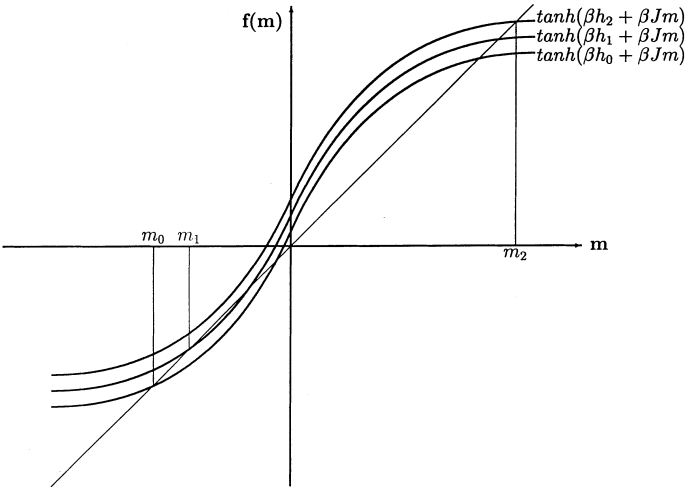

Figure 1. Evolution of average theory choice

Given the multiple steady states which exist in the model, it is essential to distinguish between marginal and nonmarginal changes in h. A marginal change in h will alter the steady state mean according to

$$\frac{dm^*}{dh} = \frac{\beta(1 - \tanh^2(\beta h + \beta J m^*))}{1 - \beta J(1 - \tanh^2(\beta h + \beta J m^*))} \tag{18}$$

The sign of this function is of course positive. This function is highly nonlinear and means that the magnitude of the impact of changes in the evidentiary support of one theory versus another will depend on the current mean as well as the various behavior parameters of the model.

A second and more interesting type of nonlinear behavior can occur in the model when one considers a sequence of nonmarginal changes in h. Nonmarginal changes lead to the possibility that in addition to movement of a particular steady state, there may be a change in the existence of the steady state itself. This feature is illustrated in Figure 1, in which there is a set of equal size movements in h_t.[9] In this figure, it is assumed that the community of scientists is characterized by mean m_0, which corresponds to

[9] We assume that the evolution of h_t is monotonic in favor of one theory. While monotonicity is not important for the subsequent analysis, the assumption that more and more evidence accumulates in favor of one theory (which by the evaluative criteria of community members is superior to the other) is essential. Kitcher (1993) provides a model of the allocation of scientific resources which can naturally generate this assumption as an equilibrium in scientific effort. To be clear, there is no necessary reason to suppose that evidence in favor of a superior theory accumulates in this fashion, and certainly pathological episodes such as Lysenkoism indicate that the opposite could even happen. However, we regard this assumption as consistent with those scientific practices with which we are familiar and would certainly agree with Kitcher and others that the reward structure for scientific activity essentially achieves this over time.

the case where social factors have caused community consensus to converge on the theory which the scientific valuation would have (on average) rejected. An increase in evidentiary support for theory 1, measured as a movement from h_0 to h_1, induces a shift in the average toward theory 1, although most scientists still choose theory -1. However, when the level of evidentiary support shifts the relative valuation from h_1 to h_2, the multiplicity of steady state outcomes disappears. The unique steady state of the model, m_2, possesses a sign consistent with that of h.[10]

The dynamics described in the Figure match the empirical challenge described by Laudan at the beginning of this paper. Suppose that an accepted theory T_{-1} is challenged by an alternative theory T_1 which from the perspective of scientific criteria is superior, so that h_0 is positive. Social factors can lead the continuing acceptance of T_{-1}. As evidence accumulates in favor of T_1, i.e. h_t increases, there will exist a point where a discontinuous change in the community occurs, corresponding to the elimination of the self consistent steady state in support of the old theory. There is no conflict between an extended period of disagreement and a rapid emergence of theoretical consensus. Hence the "tension" between disagreement and consensus which Laudan describes is naturally addressed within our framework.[11]

An additional implication of this analysis is that decisive evidence in a scientific controversy is contextual, i.e. the impact of an increment on the evolution of scientific consensus is determined by the distribution of relative weights at the time the new evidence is introduced in combination with the strength of the social interaction effect βJ. The most extreme version of this holds when an incremental piece of evidence eliminates two of the possible self-consistent equilibria, as illustrated in Figure 1. Hence, the importance of Eddington's eclipse observations for the general acceptance of relativity were a function not just of the findings per se, but also of the totality of observational evidence and the ampliative virtues of the theory relative to Newton's. For example, Eddington's observations were interpreted by the scientific community against a background in which the aberrations in Mercury's orbit and the Michelson-Morley experiments were already known.[12]

---

[10] This type of discontinuous behavior is common in models with positive interaction effects: see Brock and Durlauf (1995) for discussion and references.

[11] Laudan (1996), chapter 13, argues that the emergence of consensus arises due to the evolution of new pieces of evidence which are able to satisfy scientists with different evaluative criteria from those who initially accept a theory. In our framework, this can be done by allowing the $h_{i,t}$'s to differentially shift across time. Notice that Laudan's account does not include any conformity effects and so does not endogenously produce the threshold consensus formation found in our model.

[12] See Durlauf (1997) for a related discussion of how the interpretation of how the presence of particular alternative theories defines the context in which evidence evaluation can occur.

iii. How does diversity of opinion affect scientific consensus?

The introduction of heterogeneity in $h_{i,t}$ allows a straightforward analysis of the effects of an increase in the diversity of scientific evaluations of the relative merits of the two theories. We do this by introducing a mean preserving spread in the $h_{i,t}$'s. Suppose that

$$h_{i,t} = h_t + u_{i,t}$$

where $u_{i,t}$ is *iid* across scientists. When the variance of this shock, σ_u^2, is zero, the steady states of the model will correspond to those of eq. (13). When the variance is increased from zero, the effect on the average theory choice will depend on the slope of the $\tanh(\beta h_t + \beta J m_t^*)$ where recall that m_t^* is the initial self consistent average. Assuming that we are at a locally stable steady state and given that these states, by Proposition 2, are those where the $\tanh(\beta h_t + \beta J m_t^*)$ function is locally concave, this implies that the population average will move towards the origin. Hence a mean-preserving spread in the relative theory evaluations will lead to a reduction of the average choice level in the direction of zero (i.e. towards half the community choosing each of the two theories). It is even possible for an increase in heterogeneity to cause the population to flip to a new steady state, (and hence overshoot the origin) for reasons already described in the discussion of the influence of β in the previous section. Taken together, the implications for the distribution of beliefs, as measured by the variance of $h_{i,t}$ and the level of β, indicate that increased heterogeneity of beliefs will act to mitigate the potential for social factors to cause a community to form a consensus around an inferior theory.[13]

iv. Do social factors impede the acceptance of more successful theories?

Our analysis of the progressiveness of science suggests that the role of social factors is more complex than is often recognized. From the analysis of stability, it is clear that social factors can impede the acceptance of a new, scientifically superior theory. This impedance is manifested by the existence of a negative mean steady state.

However, once the mean choice level possesses the same mean as h, the presence of social factors will act to *increase* the degree of consensus around the scientifically preferable theory.[14] This makes intuitive sense, since once the consensus of the community is centered on the superior theory, this consensus will influence its rapid acceptance. Notice that this holds even if there are multiple steady states in theory choice. In this case, social factors increase the absolute value of the extremal steady states, creating the pos-

[13] See Kitcher (1993) for a complementary analysis in which heterogeneity diversifies the research output of a community, which in our notation, will lead the distribution of $h_{i,t}$'s to evolve.

[14] This is an immediate consequence of the monotonicity of the tanh(.) function and the assumption that $\beta J > 0$.

sibility that, depending on which steady state is realized, social factors can lead to widespread consensus around either of the two theories.

These features are summarized in the following proposition, which assumes that h is positive.

Proposition 3. Interaction of social factors and average theory choice

i. Conditional on the other parameters in the model, the greater the value of J, the greater the average theory choice in the model, when the average choice level and h possess the same sign.
ii. Conditional on the other parameters in the model, the greater the value of J, the lower the average theory choice in the model, when the average choice level and h possess the opposite sign.

5 Extensions

i. Leaders and followers

Our baseline model assumes that each scientist possesses an identical weighting formula when considering theory choices of other members of his community. A natural modification, discussed and formalized in a very different context by Kitcher (1993) is to decompose the community of scientists into two types, leaders and followers, which are distinguished by subscripts l and f respectively. The average theory choice for the scientific community will then be a weighted average of the two groups, Letting $m_{l,t}$ and $m_{f,t}$ denote the means of the leader and follower groups, the overall mean m_t will be the population weighted average of the group specific means

$$m_t = \lambda m_{l,t} + (1 - \lambda) m_{f,t} \tag{19}$$

Leaders and followers may be distinguished with respect to the parameters h, β and J. When each scientist possesses an h and β which depends on his status as a leader and follower as well as a separate social interaction weight which depends on the leader/follower combination involved in any pairwise interaction, the mean of the leaders and followers will be described by any joint solution to the pair of equations

$$m_{l,t} = \tanh(\beta_l h_{l,t} + \lambda \beta_l J_{l,l} m_{l.t} + (1 - \lambda)\beta_l J_{l,f} m_{f,t}) \tag{20}$$

$$m_{f,t} = \tanh(\beta_f h_{f,t} + \lambda \beta_f J_{f,l} m_{l,t} + (1 - \lambda)\beta_f J_{f,f} m_{f,t}) \tag{21}$$

One special case of the leader follower model may be obtained from the further assumptions

$$J_{l,l} = J_{l.f} = 0 \tag{22}$$

$$J_{f,l} > J_{f,f} > 0 \tag{23}$$

Together, these assumptions mean that leaders make decisions independent of a desire to conform whereas followers experience conformity effects which

are especially strong relative to the average behavior of leaders. In terms of qualitative differences with the baseline model, there are several features worth noting. First, in this formulation, since the leaders of a discipline are assumed to make theory choice exclusively according to scientific merit whereas followers are influenced by social as well as scientific factors, the model can exhibit a complementary explanation to the disagreement/rapid consensus phenomena discussed in the previous section. In the leader/follower case, the development of scientific evidence supportive of theory 1 can, once it induces a sufficient consensus among leaders, generate a mean shift of followers towards the new theory. Of course, such dynamics only make sense if the leaders in the community do not have a vested interest in the continued acceptance of the old theory, which they have presumably helped to develop.

ii. Schools of thought

The leader/follower framework can be easily modified to account for schools of thought. In the simplest case, suppose that there are two schools of researchers, A and B. These schools are distinguished in that members of each school place different conformity weights on the theory choice of members of the same versus the alternative school. If each group weights internal and external conformity in the same way, the within school means $m_{A,t}$ and $m_{B,t}$ can be expressed as the solutions to:

$$m_{A,t} = \tanh(\beta h_t + \beta J_s m_{A,t} + \beta J_D m_{B,t}) \tag{24}$$

$$m_{B,t} = \tanh(\beta h_t + \beta J_D m_{A,t} + \beta J_s m_{B,t}) \tag{25}$$

where J_s is the same-school interaction weight and J_D is the different-school interaction weight.

The important difference between this model and the previous one is the possibility that different schools of thought end up in different steady states, depending on the particular weights attached to the two communities. In the extreme case, suppose that $J_D = 0$. In this case, the communities will independently replicate our original model. So long as there exist multiple steady states, then it is possible for different communities to reach different conclusions based upon the same scientific evaluations.

iii. Endogenous evolution of h_t

A third modification of the basic framework concerns the evolution of evidentiary support distinguishing the two theories. One way to do this is to assume, for reasons ranging from relative ease of funding to the possibility that those who accept a particular theory are more likely to find evidence supportive of it than those who prefer another theory, that the rate of change in evidentiary support is a function of the percentage of individuals who accept it at a point in time. Formally, this would mean that

$$h_t - h_{t-1} = \phi(m_{t-1}) \ . \tag{26}$$

We assume that $\phi(\cdot)$ is positive and monotonically increasing in m_{t-1} as we are again interested in the case where theory 1 dominates theory -1 in terms of comparative scientific evaluation, this suggests a second mechanism by which nonlinear theory dynamics can occur. When a consensus initially exists around a given theory, the introduction of a superior alternative will not necessarily produce a rapid accumulation of evidence in its favor. As evidence builds up in favor of one theory, this will feedback into the rate of change of the level of evidence, creating standard nonlinear dynamics.[15]

iv. Alternative dynamics

Finally, our model can accommodate different dynamics from those implied by the expectation formation equation (16). In particular, we consider the case where the expected value of the average community choice at t equals a simple average of past realizations,

$$m_t^e = t^{-1} \sum_{j=1}^{t} m_{j-1} \ . \tag{27}$$

In this case, the sequence of expectations possesses a recursive structure in that

$$\begin{aligned} m_{t+1}^e &= \frac{t}{t+1} m_t^e + \frac{1}{t+1} m_t \\ &= \frac{t}{t+1} m_t^e + \frac{1}{t+1} \tanh(\beta h_t + \beta J m_t^e) \ . \end{aligned} \tag{28}$$

In order to understand the limiting properties of this equation, we proceed as follows. First, straightforward manipulation of eq. (28) yields an adjustment equation for expectations of the form

$$m_{t+1}^e - m_t^e = \frac{1}{t+1} (\tanh(\beta h_t + \beta J m_t^e) - m_t^e) \tag{29}$$

Second, we decompose $\tanh(\beta h_t + \beta J m_t^e) - m_t^e$ as

$$\tanh(\beta h_t + \beta J m_t^e) - m_t^e = E_{t-1}(\tanh(\beta h_t + \beta J m_t^e) - m_t^e) + u_t \tag{30}$$

where E_{t-1} denotes the conditional expectation operator given all information available by $t-1$ and u_t is (by construction) a martingale difference sequence. Together, we can rewrite the expectations adjustment process as

[15] To give but one example, Holling and Sanderson (1996) provides an example of a mathematically similar story for ecological systems.

$$m^e_{t+1} - m^e_t = \frac{1}{t+1}(E_{t-1}(\tanh(\beta h_t + \beta J m^e_t) - m^e_t) + u_t) \tag{31}$$

This equation may be analyzed using results for a broad class of dynamic processes studied by Benaïm and Hirsch (1994). In particular, they show that for a general process of the form

$$m^e_{t+1} - m^e_t = g_{t+1}(F(m^e_t) + u_t) \tag{32}$$

where g_{t+1} is a sequence of positive numbers decreasing to zero, u_t is a martingale difference sequence, the limiting behavior of this system is characterized by the differential equation

$$\frac{dm^e}{dt} = F(m^e) \tag{33}$$

In our case, which is an example of (32) when h_t is constant, the relevant differential equation is

$$\frac{dm^e}{dt} \tanh(\beta h + \beta J m^e) - m^e \tag{34}$$

Hence, Benaïm and Hirsch (1994) implies that with probability one, under the expectations process eq. (27), the average population choice will converge to the one of the locally stable steady states described by Proposition 2.

How does the evolution of h_t influence this conclusion? The evolution of h_t introduces a way in which path dependence in theory choice is influenced by the evolution of evidence. Suppose that h_t evolves towards some fixed $\overline{h}$, achieving this limit at some finite T. The specific sample path of h_t will determine (along with the sample path of u_t) which of the steady state solutions is realized. The evidentiary noise associated with the deviations of h_t from $\overline{h}$ can thus mean that either of the deterministic steady states is reached with positive probability.

Finally, as shown in Bena and Hirsch (1994), it is possible to retain the qualitative features of this analysis if eq. (27) is replaced with an alternative averaging mechanism which assigns greater weight to more recent observations.

6 Conclusions

This paper has attempted to contribute to the analysis of the sociology of science by constructing a very stylized model of community theory choice. Despite the many simplifications, the model was shown to be able to replicate an important nonlinearity in the evolution of scientific paradigms – the movement of a community of scientists from a period of extended disagreement to a period of rapid consensus formation. This feature was shown to naturally fall out of a framework which assumes that theory choice is *comparative*, in that the community is choosing between a pair of alternative theories. In addition, the model suggests that the influence of social factors in

scientific theory choice may assist rather than hinder scientific progress, at least as measured by progressiveness in the evolution of theories. While we do not establish any presumption that social factors facilitate rather than impede scientific progress, our analysis does indicate that any such conclusions will depend upon a range of specific details of community interactions. Hence we are skeptical that there is any generic empirical regularity to be found in the role of social factors in the evolution of scientific theories; rather careful case-by-case historical studies need to be conducted.

This analysis, of course, does not speak to the question of how alternative theories emerge for consideration in the first place. In Popper's language, we have discussed the community-level logic of justification rather than the logic of discovery. Any conclusions we draw on the possible facilitating role of conformity effects on movement towards superior theories is strictly conditional on the available theory options. Given the viability of scientifically inferior local theory alternatives at any point in time, it is possible that evolution of theory choices could lead to long run path dependence in scientific knowledge.[16] Extension of our model to endogenize the evolution of theory components is an important complement to the current analysis.

References

Anderson, S., de Palma, A., Thisse, J.-F.: Discrete choice theory of product differentiation. Cambridge: MIT Press 1992

Bena, M., Hirsch, M.: Dynamics of Morse-Smale urn processes. Mimeo, Department of Mathematics, University of California at Berkeley (1994)

Bloor, D.: Knowledge and social imagery. London: Routledge and Kegan Paul 1976

Blume, L.: The statistical mechanics of strategic interaction. Games and Economic Behavior **5**, 387–424 (1993)

Brock, W.: Pathways to randomness in the economy: emergent nonlinearity and chaos in economics and finance. Estudios Economicos **8**, 3–55 (1993)

Brock, W., Durlauf, S.: Discrete choice with social interactions. Mimeo, Department of Economics, University of Wisconsin at Madison (1997)

Chalmers, A.: What is this thing called science? Indianapolis: Hackett 1994

Diamond, A.: Science as a rational enterprise. Theory and Decision **24**, 147–167 (1988)

Durlauf S.: Nonergodic economic growth. Review of Economic Studies **60**, 349–366 (1993)

Durlauf, S.: Limits to science or limits to epistemology? Complexity **2**, 31-37 (1997)

Feyerabend, P.: Against method. London: Verso 1975

Feynman, R.: The character of physical law. Cambridge: MIT Press 1965

Friedman, M.: The methodology of positive economics. In: Essays in positive economics. Chicago: University of Chicago Press 1953

Hands, D. W.: Blurred boundaries: recent changes in the relationship between economics and philosophy of science. Studies in the History and Philosophy of Science **20**, 751–772 (1994)

[16] The failure to explicitly consider whether and how small local differences between theories at a point can lead to global differences over time is, in our opinion, a serious weakness of attempts to argue that social factors have long term effects on the development of science. For example, none of Bloor's (1976) conjectured alternative mathematics notions are ever shown by him to possess any substantive implications for the growth of mathematical knowledge per se.

Holling, C. Sanderson, S.: Dynamics of (dis)harmony in ecological and social systems. In: Hanna, S. (ed.) In rights to nature. Washington, DC: Island Press 1996
Kitcher, P.: The advancement of science. Oxford: Oxford University Press 1993
Kuhn, T.: The structure of scientific revolutions, revised edn. Chicago: University of Chicago Press 1970
Latour, B.: Science in action. Cambridge: Harvard University Press 1987
Laudan, L.: Science and values. Berkeley: University of California Press 1984
Laudan, L.: Beyond positivism and relativism. Boulder: Westview Press 1996
Newton-Smith, W.: The rationality of science. London: Routledge and Kegan Paul 1981
Oomes, N.: Market failures in the economics of science. Mimeo, Department of Economics, University of Wisconsin (1997)
Popper, K.:Objective knowledge. London: Oxford University Press 1972
Popper, K.: Conjectures and refutations. London: Routledge and Kegan Paul 1976
Putnam, H.: The diversity of the sciences. In: Words and life. Cambridge: Harvard University Press 1994
Quine, W. V. O.: Two dogmas of empiricism. Philosophical Review **51**, 20–43 (1951)