THE ELEMENTARY ECONOMICS OF SCIENTIFIC CONSENSUS†

Jesús P. ZAMORA BONILLA*

Manuscript received: 1999.3.27.
Final version: 1999.8.3.

* Departamento de Economía, Universidad Carlos III, 28903 Getafe, Madrid. E-mail: jpzb@eco.uc3m.es

BIBLID [0495-4548 (1999) 14: 36; p. 461-488]

ABSTRACT: The scientist's decision of accepting a given proposition is assumed to be dependent on two factors: the scientist's 'private' information about the value of that statement and the proportion of colleagues who also accept it. This interdependence is modelled in an economic fashion, and it is shown that it may lead to multiple equilibria. The main conclusions are that the evolution of scientific knowledge can be path-dependent, that scientific revolutions can be due to very small changes in the empirical evidence, and that not all possible equilibria are necessarily efficient, neither in the economic nor in the epistemic sense. These inefficiencies, however, can be eliminated if scientists can form coalitions.

Keywords: economics-of-science; rationality; theory-choice; sociology-of-science; game-theory; coallitions; path-dependence.

CONTENTS

1. Introduction
2. Reasons to accept a scientific statement
3. Some objections considered
4. The social equilibrium
5. Analytical properties of the equilibria
6. Underdetermination and path-dependence
7. Scientific revolutions
8. Efficiency properties
9. Negotiation and coalitions
10. Conclusions and prospects
Appendix A. Equilibrium of rival statements
Appendix B. Consensus formation when each scientist has a different weight in the scientific community
Bibliography

1. Introduction

In recent years, the interest in the economic study of the process of scientific research has grown considerably. The crossing of some lines of research in economics, sociology and philosophy is currently creating a completely new and extremely promising approach to the study of science, which will constitute very possibly one of the main schools in this field during the next decades. The main idea behind this 'new economics of science' is that scientific researchers are *rational* people in continuous interaction and interdependence, looking *individually* after the maximisation of some 'personal utility function', but also looking *institutionally* after the efficient production of 'knowledge'. Philosophers and methodologists may find in this approach an explication of the social character of science more 'rational' than that usually found in the works of radical sociologists, and also clearer (analytically speaking) than other current naturalist approaches in philosophy; economists may have the opportunity of employing their conceptual tools to study an intriguing human activity, one which is at least as rational as the classical 'economic' aspects of life (business, consumption, saving, investment, and so on); and sociologists of science may discover an interesting way of developping in a formal fashion some of their insights about the social processes which underlay the creation of scientific knowledge.

The economic study of scientific research has, in particular, a strong motivation in view of the radical conclusions of the so called 'sociology of scientific knowledge'. According to this approach, science can not be seen as an efficient mechanism of knowledge generation ('efficient' in the sense that the knowledge so generated is 'objective' epistemically speaking), because the production of scientific knowledge is always essentially permeated by social, non-scientific interests. On the other hand, one of the fundamental ideas of economic theory is that the interdependence of rival interests *can* lead, under some especific circumstances, to an efficient social arrangement. Hence, it is not *a priori* clear that the essential presence of rival scientific and non-scientific interests in the process of scientific research precludes the attainment of an epistemically acceptable corpus of knowledge. Whether this is so or not, it is something that can only be discovered through the study of alternative economic models of scientific research.

In any case, my aim in this paper is not to offer a survey of the most interesting work in the economics of science;[1] I just want to present a very simple model (but I think an illuminating one) of the process through which individual researchers interact in order to reach a certain degree of consensus about a given proposition. I will restrict the scope of my paper to the scientists' decision of accepting or not accepting such a statement -already presented by some of them- on the basis of the experiences, previous knowledge (*i. e.*, previously accepted propositions) and the methodological and social preferences each scientist has. That is, I will ignore those decisions referring to how new statements are looked for and proposed, how new experiments, observations or calculations are chosen, designed and performed, etc. These are extremely interesting problems to analyse from the economic point of view, but the simple economics of scientific consensus (or dissensus) is, I think, an unvoidable preliminary work before trying to answer these other questions. On the other hand, my aim is simply to make an economic *interpretation* of some common ideas in the recent philosophy and sociology of science; I would be content if the conceptual schemes offered here were helpful to illuminate a little some long disputations about the rationality of science.

I will assume that scientists take their decisions with the only aim of maximising their individual 'utility', though I will make also the assumption that researchers have two different elements in their 'utility function', an *epistemic* as well as a *social* one (I will examine these elements in section 2, and consider some plausible objections in section 3). The 'social value' of a scientific statement will be seen simply as its rate of acceptance within the relevant scientific community,[2] and it will be considered as the result of an *equilibrium* in the decision patterns of each scientist, using some elementary game-theoretical tools (section 4). Section 5 will examine the three essential properties of these equilibrium solutions, namely, their *existence* or *non-existence*, their *uniqueness* or *multiplicity*, and their *stability* or *unstability*. The next three sections are devoted to derive some philosophical implications of the model: the question about the *underdetermination* of theories and the possible *path-dependence* of scientific knowledge (section 6), the possibility of *scientific revolutions*, understood here simply as sudden, big changes in the rate of acceptance of a set of interrelated statements (section 7), and the *efficiency properties* of the equilibria referred to above (section 8). In section 9, the model will be enlarged to include the possibility of forming *coalitions* among researchers, and I will show that some negative conclusions of the preceding sections cease to apply in this case. Section 10 offers some conclusions and prospects. The paper is completed with two appendixes in which I study two more complicated situations: the choice between rival statements and the possibility that each member of the scientific community has a different 'weight' in the process of consensus formation.

2. Reasons to accept a scientific statement

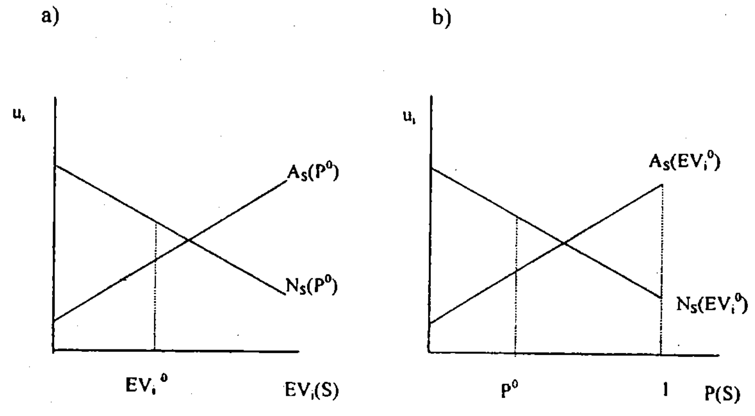
Science is a very competitive arena, in which researchers are continuously trying to make things before and better than their rivals, and trying to discover mistakes in their colleagues' work. But to put an excessive attention on this hard competition can hide the astonishing fact that science is the most successful mechanism of consensus generation ever designed by human beings.[3] For each proposition a researcher proposes by herself or puts under critical scrutiny, she needs to *accept* literally hundreds of other statements, and it is in part this acceptance what makes of her a 'good' scientist before her community, since this is usually the best public proof that she 'masters' her discipline. We can divide into two groups the reasons an individual scientist may have to accept or not to accept a given proposition or set of propositions (no assumption is made about the relative importance of each group of reasons):

a) in the first place, she may have a series of epistemic arguments (results of experiments or observations, logical connections with other accepted statements or with background assumptions, methodological preferences, and so on) which will offer a certain 'acceptability' or 'epistemic value' to the proposition for her; we can call this the scientist's *private information* about the statement;

b) on the other hand, she will also need to take into account what her colleagues manifest about that statement (especially because she recognises that her own private information is usually very limitted, but also because to conform to the decisions of her community may provide some benefits); we can call this the *public information* about the statement.

I will assume, *ceteris paribus*, that a normal researcher will be more willing to accept a proposition S the higher is for her its epistemic value and the higher is the proportion of members of the relevant community who already accept it. Let $EV_i(S)$ and $P(S)$ represent these two concepts. The former can be different for different researchers, that is, not every scientist must give the same epistemic value to each proposition, since each one will have her own methodological criteria, her own private experiences, and will not necessarily accept the same statements besides *S*. On the other hand, the idea that *P(S)* influences the decision of the individual scientist can be seen, from the economic-theoretical point of view, as a rather *ad hoc* assumption, lacking sufficient 'micro-foundations'. To understand its consistence with the rationality principle, take into account that each researcher only has a limited private information about each statement (and, in many cases, it has *no* private information at all); so, the number of scientific propositions she could assert 'by herself' would be simply ridiculous: if the direct or indirect aim of scientists is to maximise their knowledge of the world (or their contribution to the stock of knowledge possessed by society, or the rewards based on this contribution), they need trust in the word of their colleagues, making of science a *collective* enterprise, where each one is willing to accept the propositions presented by others only if she has some reason to believe that her colleagues are presenting those propositions because they have some private information supporting them. The more colleagues accept a given statement, so, the less probable it is that all the private information they have is misleading. So, *P(S)* can be seen as a *signal* (though perhaps an imperfect one) of the non-observable private information about *S* held by the other researchers. This is simply an example of the necessity of mutual confidence when a social institution presupposing some division of labour is created.[4] Something very similar takes place in the market: you can specialise in the production of a single good only because you *believe* that others will produce the other goods you desire and will accept your money, and you will not have any other signal supporting this belief besides the others' observed behaviour. Moreover, 'accepting' one statement is not only an *epistemic* decision: it has not to be identified necessarily with 'believing' in its truth (or, at least, in its 'approximate truth', or in the truth of its empirical consequences). As it was indicated by Kuhn (1962), a scientist can 'accept' a proposition simply in the sense that she makes use of it as a common tool to solve some kind of scientific problems. We must take into account that the highest number of statements that a researcher uses while writing a paper, for example, corresponds to those statements propositions that she *takes for granted*, not to those she is arguing for or against; the former are statements which she founds in textbooks and other papers, and that she accepts mainly because many other colleagues have already accepted them (supposedly on sound epistemic reasons). *In this sense,* statements or families of statements are to be seen more like 'tools' or 'technologies' than like '(possible) beliefs'. A scientific statement, hence, can show something similar to what economists call 'increasing returns to adoption': the more people uses the same technology, the lower the costs and the higher the benefits of deciding to use it.[5]

The researcher who is confronted to the decision about to accept or not to accept a proposition *S* will then compare the utility levels associated to both options, and these will depend on the actual values of $EV_i(S)$ and $P(S)$. This situation is graphically represented in fig. 1.a-b, where $EV_i(S)$ and $P(S)$ are idealised as continuous variables, and where the lines 'A_S' and 'N_S' represent respectively the utility levels associated to accepting and to not accepting *S* for the *i*-th researcher (for concreteness, I will assume that, if both utility levels coincide, then *S* will be accepted). These lines will shift upwards or downwards when a change occurs in the abscise levels of the *other* diagram; for example, if $P(S)$ increases, passing from P^0 to a higher value P^1, then the utility of accepting *S* will be higher for *each* value of $EV_i(S)$, and the utility of not accepting *S* will be lower (see fig. 1.c-d); this sometimes will have the consequence that, though *S* was not accepted by *i* when only a fraction P^0 of her community accepted the statement, it becomes accepted by her if a higher proportion P^1 of colleagues accept it. In the same way, an increment of the epistemic value that *S* has for *i* will displace analogously the curves A_S and N_S in fig. 1.b. Take into account that both fig. 1.a and fig. 1.b represent exactly the *same* decision, and so, given a pair of values of $P(S)$ and $EV_i(S)$, the height of the curve A_S must be the same in both diagrams for this combination of values (and correspondingly the height of N_S). Note also that 'not accepting *S* has not to be confused with 'accepting not-*S* (the choice between two contradictory statements is analysed in appendix A).



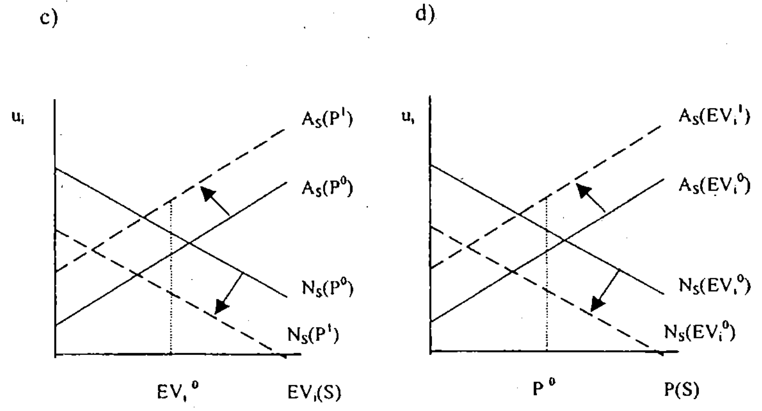


Figure 1

I do not make any assumption concerning whether the researchers' utility functions are identical or not. We can suppose that formal education tends to make them more or less similar within a single discipline, or at least, within each 'school'. But what is really important is that, even having the same utility function, individual researchers will tend to take different decisions because of having different private information about the epistemic value of each statement. The aim of my model is, precisely, to analyse how this different, and perhaps conflicting, pieces of private information interact in order to form a consistent social state.[6]

3. Some objections considered

A possible objection to my model at this stage could be that nothing forces the scientist to take a decision between 'accepting' and 'not accepting' *S*: she could simply express something like '*S* has for me an epistemic value of EV_i'[7]; or there could be more options than simply to accept or not to accept each proposition (she could affirm '*S*', or '*S* is highly probable', or 'there seems to be some evidence in favour of *S*...). But I take it as an empirical datum that scientists *explicitly accept* many statements during their work and in their communications, perhaps because their institutional role is to offer these statements to the rest of society as pieces of uncontroversial knowledge, or simply because human minds can not work properly with mere probabilities, and need to take at least some statements as true in spite of the risk of making mistakes. Furthermore, if we consider the possible 'modalisations'[8] with which a statement can be presented (for example, 'certain facts seem to count in favour of *S*'), we could also take *this* longer proposition as one which the researcher has to decide whether to accept or not to accept.

Another reasonable objection is that what a researcher will actually take into account will not usually be the *proportion* of colleagues who are accepting a proposition, but, rather, *who* is exactly doing what; some researchers' decisions will be surely more influential than others'. In fact, I have chosen to use function '*P*' simply because of its simplicity, but it can be easily adapted to the case where each researcher has a different 'weight' within the scientific community, if this weight is recognised for all, because, in this case, the expression '*P(S)*' can refer to the *sum* of the weights of those researchers who accept *S*. The model becomes more complicated instead if each researcher did not give the same weight than her colleagues to each member of the community. These possibilities are studied in appendix B.

A third objection asserts that, for some researchers at least, the slopes of A_S and N_S could be inverted in fig. 1.b.[9] These researchers (which we can call 'the sceptics') would naturally tend to suspect that 'something goes wrong' with a proposition if everybody accept it. I assume that this can be the case, but also that sceptics will usually not be many (what entails that their existence will not disturb my model's conclusions too severely), and, what is more important, that their scepticism will be directed only towards a few, though significant propositions. Moreover, scepticism will be more usual when somebody intends to *propose* a new statement, a kind of action which I have left aside of my model.

4. *The social equilibrium*

In order to obtain a 'social equilibrium', we can aggregate the individual decisions of fig. 1.b in a fashion analogous to the 'bandwagon model' of demand and to other 'herding' models.[10] Imagine we ask each researcher whether she would be willing to accept *S* if $P(S) = 0$; perhaps nobody would accept *S* under these circumstances, but any other answer is also possible, depending on the private information each researcher has about *S*. If we repeat the question for each possible value of *P(S)*, we will obtain as a result a non decreasing line represented in fig. 2 as the 'reaction curve of *S*', *RC(S)*.[11] The reason why this curve is non decreasing is because of the slopes of A_S (increasing) and N_S (decreasing) in fig 1.b, for these slopes entail that passing from a lower value of *P(S)* to a higher one can only increase the number of researchers who accept the statement. This would not be necessarily true if there were relatively *many* 'sceptics' in the scientific community (*i. e.*, more people who pass from accepting *S* to not accepting it, when *P(S)* increases, than people who do the contrary), for in this case some reaction curves could have some decreasing portions; I shall nevertheless ignore this possibility in the rest of the paper.

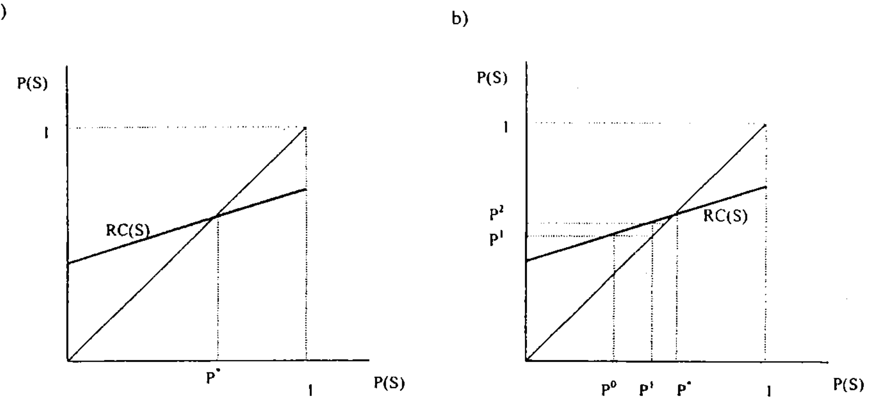


Figure 2

The only possible social equilibria are those points where *RC* crosses the identity line of 45 degrees, *i. e.*, the 'fixed points'. These points are Nash equilibria, since they correspond to situations where everybody is taking her best option given the decisions taken by the others. To understand why it can be expected that the social situation is one of equilibrium, think that, in principle, it is possible that any other value of *P(S)* obtains; let, for example, be P_0 the original situation; according to *RC(S)*, if P_0 researchers accepted *S*, then the proportion of individuals who would prefer to accept it would be $P_1 > P_0$, and so, a proportion $P_1 - P_0$ would change their original decision of rejecting *S*, once they discover their 'mistake'. If P_1 is still not an equilibrium, the process will continue until P^* is reached. A parallel argument shows how the equilibrium is reached if we start from some point to the right of P^*. Of course, what guarantees that the equilibrium is reached is a fluent communication among individual researchers; institutions facilitating this fluency are, so, essential to the proper working of science. Complete consensus about a statement *S* will be obtained, obviously, when its reaction curve reaches an equilibrium at which *P(S)* equals 1 or, if we are not very strict, if it is 'close' to 1.

5. Analytical properties of the equilibria

The main analytical properties that are looked for in economic equilibria are their *existence*, *uniqueness* and *stability*. With respect to the first of these properties, it is easy to prove that some equilibrium must exist, recalling that there are only a finite number of researchers (say, N) in the relevant scientific community, and hence, that the domain and the values of RC can only be rationals of the form n/N, where n is a natural number between 0 and N. Since $RC(0) \geq 0$, then 0 is an equilibrium if $RC(0) = 0$; if it is a possitive number, then there are two possibilities: either $RC(n/N)) > n/N$ for every n such that $0 < n < N$, or $RC(n/N) \leq n/N$ for some n; in the first case, $RC((N-1)/N)$ will necessarily be 1, and hence 1 (*i. e.*, N/N) will be an equilibrium, since RC is non-decreasing. In the second case, if there are some n between 0 and N for which $n/N >$ RC(n/N), let m be the first such number; hence we have that $RC(m/N) \geq RC((m-1)/N) \geq (m-1)/N$, and then, $m/N >$ RC$((m-1)/N) \geq (m-1)/N$; so, RC$((m-1)/N) = (m-1)/N$, and hence $(m-1)/N$ is an equilibrium.

Things are more problematic regarding the uniqueness or multiplicity of equilibria. Obviously, if $RC(S)$ can be idealised a straight line, there will be only one equilibrium, like if fig. 2, but nothing guarantees that this will be the shape of the reaction curve. A more plausible case is that, among the population of researchers, the points of indifference in fig. 1.b (*i. e.*, the points where A_S and N_S cross) are more or less 'normally' distributed around some single peak; this will produce an S-shaped reaction curve, as that represented in fig. 3. In fact, I will assume that this is the habitual form of these curves.[12] If $RC(S)$ is S-shaped, it can have a maximum of three equilibria (note that point B in fig. 3, contrary to points A and C, does not need to be an equilibrium: suppose, for example, that $RC(n/N) = n\text{-}1/N$ and $RC(n+1/N) = n+2/N$, with $B = (n+(2/3))/N$).

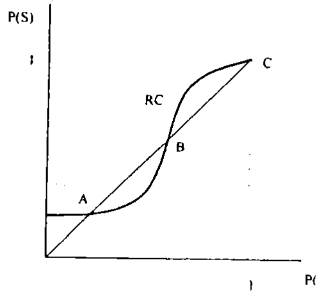


Figure 3

Lastly, with respect to the stability problem, it can be shown that an equilibrium is stable if and only if the slope of the reaction curve at that point is less than 1 (in absolute value). In fig. 2.b, it can be seen that, starting from any point in the vicinity of P^*, the underlying dynamics will force the social situation to return to the equilibrium. On the other hand, if the slope of the reaction curve is higher than 1, as in fig. 4 (in which only a part of *RC(S)* has been drawn), starting from any situation close to P^* (like P^0) we will move away from the equilibrium (to P^1, P^2, and so on). So, of the three equilibria in fig. 3, *A* and *C* will be stable, and *B* will be unstable provided it is actually an equilibrium.

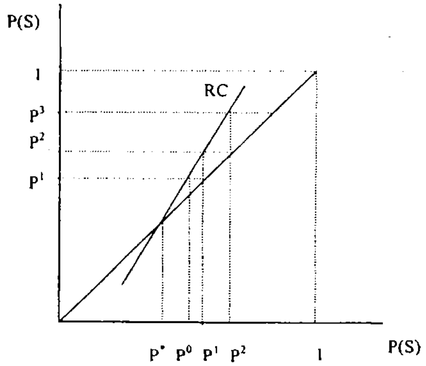


Figure 4

6. *Underdetermination and path-dependence*

The possible existence of many equilibria has some interesting consequences, which I will present in this and in the next two sections. In the first place, which equilibrium is actually reached at a certain moment will depend basically on which equilibrium was obtained at the previous moment. Imagine a situation like that depicted in fig. 5, with an S-shaped reaction curve which has moved upwards, for example because of the accumulation of empirical data in favour of *S*. What will be the *actual* equilibrium in the new reaction curve *RC'*, *B* or *D*? The answer will depend on which equilibrium, *A* or *C*, had been actually achieved in *RC* (technically, given *RC'*, *B* is a local attractor for *A* and *D* is a local attractor for *C*). Of course, the problem is the same with the old reaction curve: its actual equilibrium will have depended on the equilibria existing in the past. It must be also taken into account that changes in *RC* will usually depend on the present proportions of researchers accepting each scientific statement, since these proportions will have a strong influence on the decisions of researchers concerning what experiments, observations or calculations to do, what research lines to follow, and so on. Note also that changes in each researcher's private information do not necessarily transport *RC* monotonically upwards or downwards; that is, the reaction curves of one statement at different moments of time may cross; this will be the normal situation when, for example, some scientists have discovered evidence in favour of *S* while others have discovered evidence contrary to that statement.

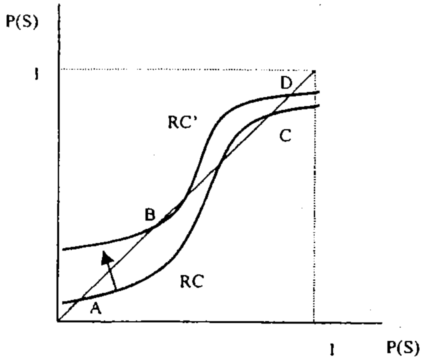


Figure 5

All this means that *the evolution of scientific knowledge will be 'path-dependent'*.[13] The particular shape and possition of *RC* at a particular moment will basically depend on the *set* of private experiences possessed by each researcher at that moment, but the equilibrium point actually reached in that *RC* will depend on the *order* in which those experiences have accumulated. The possible existence of many equilibria can also be related to the thesis of the underdetermination of theories by data. According to this thesis, for any corpus of 'empirical evidence' there are always several theories compatible with it,[14] and hence, selecting one of them can not be a question of logic alone, but will be based on other kinds of values (epistemic, social, or whatever). According to our model, *even* when these values are taken into account (through functions EV_i and P), there is room for several configurations of statements accepted by the scientific community, and the selection of one particular configuration will lastly depend on some 'historical accidents'. Empirical evidence, epistemic values and social interests may be not sufficient conditions, even when they are taken together, to explain why a certain degree of consensus about a theory has come to exist.

7. Scientific revolutions

As I have indicated in passing in the last section, the accumulation of arguments in favour of a given statement tends to move upwards its reaction curve (the reason will be explained in more detail in the next section). It can be reasonably assumed that, previously to the presentation of serious arguments supporting a theory *S*, its reaction curve has an actual equilibrium at 0, even if it had not a constant null value along its domain (see, for example, RC^0 in fig. 6.a). When new evidence in favour of the theory appears, *RC* will be progressively displaced upwards, passing first to an equilibrium like P^1 and later to P^2 (though RC^2 has another possible stable equilibrium, it is not possible to reach it starting from P^1). The three reaction curves depicted fig. 6.a are not the only periods in the first part of a theory's history, but they are merely three selected moments of a more or less continuous accumulation of evidence supporting that theory, an accumulation which displaces upwards its reaction curve also in a more or less continuous fashion. The story follows in fig. 6.b; here, the evidence favouring the theory has taken the equilibrium point to P^3. But this is a tangency equilibrium (it is only stable to the left, not to the right); any minimal addition to the evidence in favour of the theory will make it now that the new reaction curve has only one equilibrium, corresponding to a much higher value of *P(S)*. So, although the passing from 0 to P^3 as been continuous and smooth, the passing from P^3 to P^4 (or, more strictly, to some point to the right of $P^{3\prime}$ but infinitesimally close to it) is a sudden one.

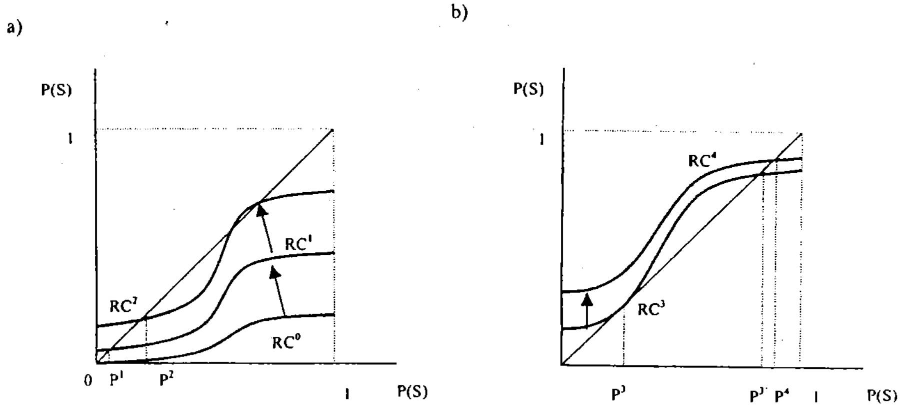


Figure 6

This phenomenon is obviously analogue to a 'scientific revolution', especially if P^3 is very low and $P^{3'}$ is very high. Take also into account that the acceptance of a statement by an individual researcher has an influence on the epistemic value that other propositions have for her. So, a sudden change in the rate of acceptance of a statement can produce an appreciable change in the rates of acceptance of others, particularly when the first one is a very basic theoretical principle. In order to employ a terminology more consistent with the usual one, we can talk about a scientific revolution not in every case when a sudden and big change occurs in the degree of consensus about one single statement, but particularly when such a change in a few basic propositions induces a similar change in a wide range of other statements. This revolution is nothing similar to a 'religious conversion', nor depends on irrational psychological factors; it is simply the result of the rational and conscious decisions of individual scientists.[15]

8. *Efficiency properties*

The next point of my exposition will be the examination of the efficiency properties of those social situations allowed by our model as stable equilibria. In the context of scientific research, we can make a distinction between two senses of the term 'efficiency'. On the one hand, it is possible to talk about 'efficiency' in the sense in which the term is commonly used in economic theory, especially in welfare economics (*i. e.*, the so called 'Paretian efficiency'). In this sense, a social equilibrium will be efficient if and only if there is no other situation where somebody gets a higher utility and nobody suffers a utility loss with respect to the first situation. On the other hand, since we are analysing in particular the social production of *knowledge*, we can talk about the '*epistemic efficiency*' of the equilibria. In this sense, efficiency will mean that, if S has for everybody at least the same epistemic value than T, then the former should not be less accepted than the latter (we can call this second concept *'Paretian epistemic efficiency'*).[16]

With respect to 'welfare efficiency', the main results are the following:

(1) *If there is only one equilibrium, or if there is no equilibrium in which* P(S) *equals 0 or 1, then all the equilibria will be efficient.*

Proof: Imagine an equilibrium like A in fig. 7.b. If we move to the left, to another point like C, there will be some researchers who still do not accept S. Since the line N_S of fig. 7.a is assumed to be decreasing, the utility of those researchers will be lower in *C* than the utility they had in *A*. The same is valid for every point to the right of *A* and different of 1 (be it an equilibrium or not).

If, on the other hand, 1 is not an equilibrium, then, if *P(S)* = 1 were the actual situation, there would be at least some researcher accepting *S* in that point, but who would prefer not to accept it given the choices of the rest; that is, her utility accepting *S* at 1 is lower than not accepting it, which, on the other hand, is lower than the utility she would have at point *A* also not accepting *S*.

The argument is analogous for all the points to the left of *A*, or if *A* is placed at 0 or 1. In conclusion, under the conditions stated in the theorem, any move of the social situation from *A* makes some researcher worse than she is at *A*, and so, *A* is Paretian-efficient. *Q.E.D.*

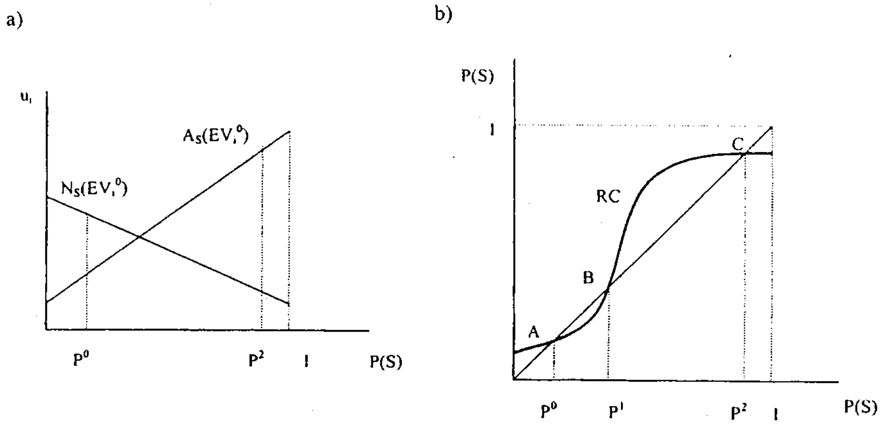


Figure 7

(2) *If there are more than one equilibrium, and one of them corresponds to* P(S) = *0 or* P(S) = *1, then the other equilibria may be not efficient.*

Proof: If *RC(1)* = 1 (*i. e.*, if 1 is an equilibrium, like in fig. 3) and there is at least another equilibrium *A*, the utility of those researchers who reject proposition *S* in *A* can be higher when they accept *S* at point 1, because A_S in fig. 7.a is increasing. The proof is analogous if *RC(0)* = 0 and there are other equilibria. *Q.E.D.*

With respect to epistemic efficiency, the main results are the following:

(3) *If both the reaction curve of* S *and that of* T *have many equilibria, it is possible that* $EV_i(S) < EV_i(T)$ *for every* i *and* $EV_i(S) < EV_i(T)$ *for some* i, but $P(T) < P(S)$.

Proof: If the epistemic value of *T* is higher than that of *S* for somebody and is not lower for anybody, then the reaction curve of *T* can not be under the reaction curve of *S* at any point, for, given any value of *P(S) = P(T)*, there will be at least as more people willing to accept *T* than willing to accept *S*. But, if both curves have many equilibria, then it is possible that the actual equilibrium of *T* (say, *A*) is below that of *S* (say, *B*), as depicted in fig. 8. *Q.E.D.*

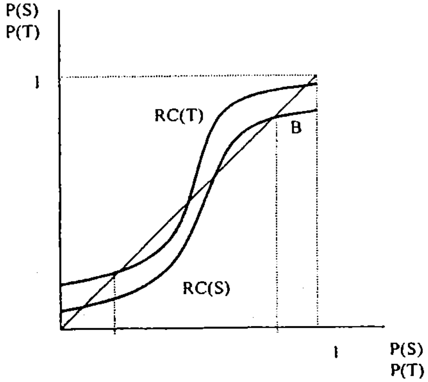


Figure 8

This theorem is one of the most pessimistic conclusions of our model. It means that there is no guarantee, in principle, that the scientific community will chose a 'good' consensus from the epistemic point of view (one which reasonably corresponds to the epistemic value that the statement in question has for each researcher), even if its members' decisions are individually rational. For example, everybody can be internally convinced that a commonly accepted proposition has a very low epistemic value, while they follow accepting it merely because the rest do the same. This is an example of what Banerjee (1992) calls a 'herd externality'. The following theorem is, however, more optimistic.

(4) *It is possible that* $EV_i(S)$ *increases for somebody (because of the result of new experiments, for example) and does not decrease for anybody, but* P(S) *decreases.*

Proof: Given an increment in the epistemic value of *S* for at least some researchers, and no decrement for the rest, the reaction curve of *S* will move upwards, as in the passing from RC_1 to RC_3 in fig. 9. If this move were made without any intermediate steps, the equilibrium would pass from point *A* to *C* or from *B* to *E*. But, if the passing from RC_1 to RC_3 has been made through RC_2 (for example, because, before discovering the new evidence highly favouring *S*, some more problematic findings had been made), in this case, the equilibrium will move from *A* or *B* to *D* to *E*. So, if the original equilibrium were *A*, it could be possible that the new equilibrium were lesser, in spite of the fact that *S* is now at least as good as before for everybody. Q. E. D.

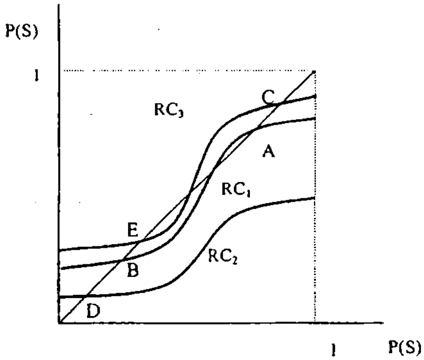


Figure 9

This theorem indicates that changes in the rate of acceptance of a proposition do not necessarily respond in a 'sound' way to changes in the epistemic valuations that researchers do of that proposition. The *order* in which the arguments in favour or against a statement have been presented is relevant for determining its actual degree of consensus, and this can make it that, what in the aggregate is an epistemic improvement of a theory produces a loss of 'social confidence' in that theory.

9. Negotiation and coalitions

One of the most shocking ideas introduced by constructivist sociologists of science two decades ago was the concept of 'negotiation'.[17] I confess that, as a philosopher of science, and before beginning to study the economic aspects of research, it was difficult for me to understand in a precise way what those authors meant with that term, and I suspect that this was also the situation of many other philosophers. The economics of scientific consensus offers an opportunity to illuminate that concept with a more precise meaning, although I think that it is not the meaning which is applicable to the main examples provided by those sociologists (since these examples refer more to the *production* of new scientific statements in the laboratory, an aspect of research which I have left aside of my model). I propose to interpret the term 'negotiation' as referring to *collective decisions*, that is, situations in which the individual researcher does not take her decision isolatedly (assuming the decisions of others as given), but discussing with others what decision to take together and simmultaneously.

If we allow for the possibility of collective decisions, that is, of decisions taken by consensus among a group (not necessarily identical to the whole scientific community under study), then there is a possibility of 'jumping' from one equilibrium to another. Let us suppose, in fig. 7.b, that the scientific community is originally at equilibrium *A*, and that there is a proportion of researchers higher than $P_1 - P_0$ such that they are rejecting *S* at *A*, but who would prefer to be at equilibrium *C* accepting that statement (*i. e.*, for each member *i* of that group, the height of the curve A_S in diagram 7.a is bigger at P_2 than the height of N_S at P_0). In this case, if this group forms a coalition and takes collectively the decision of accepting *S* (a decision which would be irrational for each member individually, without the other members' compromise of changing their decisions in the same way), then the unstable equilibrium *B* will be surpassed, and the community will move towards *C*. It is easy to see that, the closer are P_1 and P_0, the more probable is the existence of a coaliton interested in 'jumping at once' to *C*, since the necessary size of the group will be smaller. Note also that those researchers most interested in 'jumping' will be those who are, in *A*, almost indifferent between accepting and not accepting *S*, since for them the utility associated to the acceptance of *S* in *C* will be surely higher than the utility of rejecting it in *A* (which, by hypothesis, is close to the utility of accepting *S* in *A*). Furthermore, it can be reasonable assumed that 'negotiation' is not free of costs, and that these costs will increase with the size of the coalition necessary to 'jump'. So, these 'jumps' will more probably be observed when the distance from P_0 to P_1 is relatively small. On the other hand, the model of collective decisions becomes more complicated if we assume that scientists do not know exactly the shape of *RC*. In this case they will take their decisions on the ground of their *expectations*, and the path to an equilibrium will be more aleatory. I leave for another occasion the study of this possibility.

It is interesting to note that, if the jump from *A* to *C* is possible, the inverse jump will not. The reason is the following: we can divide the community into four groups, which are *a)* those researchers who would accept *S* both at *A* than at *C*, *b)* those who would not accept it neither at *A* nor at *C*, (these two groups do not have any influence in the jump's success), *c)* those who would accept *S* at *C* and would not do it at *A*, but prefer *A* to *C*, and *d)* those who would accept *S* at *C* and would not do it at *A*, but prefer *C* to *A*. Of course, if the group *c* is bigger than $P_2 - P_1$, then the jump from *A* to *C* will be possible and the other jump not, and if the group *d* is bigger than $P_1 - P_0$, then the converse will be true.[18]

The possibility of collective decisions allows to avoid the three worst conclusions of the preceding sections. In the first place, underdetermination disappears, since only one equilibrium is now possible in each reaction curve (provided they had two stable equilibria). In the second place, welfare inefficiency is also eliminated, since, if a community was at point *A* in fig. 3, for example, but everybody preferred to be at 1, then those not accepting *S* at *A* could take the collective decision of jumping from the inefficient equilibrium to the efficient one. And finally, the epistemic inefficiencies also disappears, though now the proof is not as evident as before. Recall that these inefficiencies lied in the fact that, in a case like that of fig. 8, the actual equilibrium of *RC(S)* could be higher than the one of *RC(T)*, while everybody might agree that *T* is better than *S*; if a point like *B* has been chosen collectively in the case of *S*, it is because there are more researchers who preferred that point to the lower equilibrium of *S* than people whose preferences were the other way; but, since *T* has at least the same epistemic value for everybody than *S*, all those individuals who prefer the higher equilibrium of *S* to the lower will also prefer the higher equilibrium of *T* to the lower (for the difference in utility between accepting *S* at the higher equilibrium and not accepting it at the lower –recall fig. 1.d- is lesser than the difference corresponding to *T*). The argument is parallel for the case described in fig. 9.

We can resume these conclusions in the following proposition:

(5) *If the formation of coalitions is allowed and has no costs, and if the reaction curves are convex, concave or S-shaped, then there will be only one feasible equilibrium for each reaction curve, and it will be efficient in the welfarist and in the epistemic sense.*[19]

Of course, the 'negotiations' considered in this section are only a formal instrument of analysis, and it can be doubted whether they have some correspondence to the reality of scientific communities. In the first place, my analysis is grounded on the rather strong assumption that all the possible equilibria are known by everydoby; given the other assumptions of my model about the information that researchers possess and are able to communicate, it is difficult to see how this knowledge can be actually achieved; at most, they might form an idea of the shape of the reaction curve, through the experience of the adjusting process which led to the last equilibrium (see fig.2.b and 4). In any case, the moral of my analysis is that all institutional arrangements which serve to generate a public knowledge of the full reaction curve will increase the efficiency of the process of consensus formation.

In the second place, real mechanisms of collective decision may be based on a more complex institutional structure than that assumed in my model; for example, not everybody may have the same freedom to 'negotiate', 'transaction costs' have to be taken into account, and other criteria besides individual utilities may have evolved to settle the 'negotiations' (for example, concrete epistemic or social values may be favoured by the scientific institutions). But, considering that the idealised kind of collective decisions analysed above (so to say, 'individualist' and with 'freedom of entry') tends to produce efficient equilibria, it will be interesting to study if the empirically given mechanisms of social choice existing within the scientific communities are also, and to what extent, efficiency promoting.

10. Conclusions and prospects

The simple model offered in this paper has tried to illuminate the process through which scientific propositions are accepted or rejected by the members of a scientific community. The model is based on the assumption that researchers act rationally when they decide to accept or to reject a given proposition, though, as they are not omniscient nor asocial subjects, they will usually make their choices taking into consideration the choices made by their colleagues; stated differently, they will employ not only their 'private' information on the epistemic value of a statement and their own methodological preferences, but will also take the actions of others as a 'signal' of that epistemic value. The model has shown that, as many sociologists of science defend, if 'scientific knowledge' is what a scientific community accepts as such, then the epistemic criteria of individual researchers are not enough to explain what has to be taken as 'scientific knowledge': it is also necessary to take into account the mutual interdependence between different individuals. But it has also shown that, even when this interdependence is considered, there can exist several 'social configurations of knowledge', and which is the actual one will normally depend upon more or less accidental events in each research process. The model has also shown that 'little' changes in the private evidence each researcher has can produce 'revolutionary' changes in the rates of acceptance of scientific statements. The possible existence of many equilibria entails also that the actual equilibrium needs not be 'efficient', neither in the welfarist nor in the epistemic sense of the word, though the possibility of forming coalitions can provide an escape from inefficient equilibria. Epistemic inefficiency is especially problematic in the case of a social institution such as science; it entails that, even when everybody recognises 'privately' a theory as better than another, the scientific community can 'prefer' the theory which is worse for everybody. This is due to the fact that the individual researcher does not know the 'private' preferences of her colleagues, only their 'public' decisions. The study of what institutional mechanisms (such as the coalitions of the preceding section) could promote efficiency in the epistemic sense is undoubtedly an interesting task for the 'new economics of science', and one in which the economic approach has obvious advantages over the radical sociologic approaches who firstly insisted on these inefficiency effects.

The model has completely ignored the reasons why new statements are devised and new theoretical or empirical arguments defending or attacking old and new propositions are proposed by scientists. But it is reasonable to assume that what matters in these cases for the individual researcher is especially *the social level of acceptance of the statements proposed by her.* So, our model can also serve as a preliminary background to analyse the behaviour of scientists as *producers* of knowledge, and not only as *evaluators* of statements.

Appendix A. Equilibrium of rival statements

Though I have simplified the decision researchers have to take about one statement as 'to accept it or not to accept it', things are, of course, not so simple in reality. What scientists have normally to do is to select one among a set of rival statements; even when no more than one proposition S has been set forth to interpret a set of phenomena, they have to choose between accepting *S* and accepting *not-S*. Similarly, what moves each researcher to choose her optimal decision about *S* is not only her own private information on this statement and what her colleagues manifest about it, but also what these other researchers manifest about the alternative statements which have been proposed. In the most simple case, when only two propositions are being discussed (say, *S* and *T*), this means that we should to add to fig. 1.a-b a third graphic whose abscise is *P(T)*, with the proviso that *P(S)* + *P(T)* should be at most equal to 1. The aggregation of all individual decisions is shown in fig. 10, where *RC(S,0)* is the reaction curve of *S* when *P(T)* = 0, and *RC(S,1-x,x)* is the curve formed by the final points of each reaction curve *RC(S,x)*; so, for each value of *P(T)*, the curve *R(S)* shows what value of *P(S)* would correspond to an equilibrium of *RC(S)*. For every point (*x*, *y*) to the left of *R(S)* (being *x* the value of *P(S)* and *y* the value of *P(T)* corresponding to that point), we obtain that *RC(S,x,y)* > *x*, while for points point (*x*, *y*) to the right of *R(S)*, *RC(S,x,y)* < *x*. The description is analogous for *T*. A global equilibrium occurs when *R(S)* crosses *R(T)*, as in point *A* of fig. 10. The situation is more complicated, but not essentially different, when the reaction curves of *S* or *T* have more than one possible equilibrium.

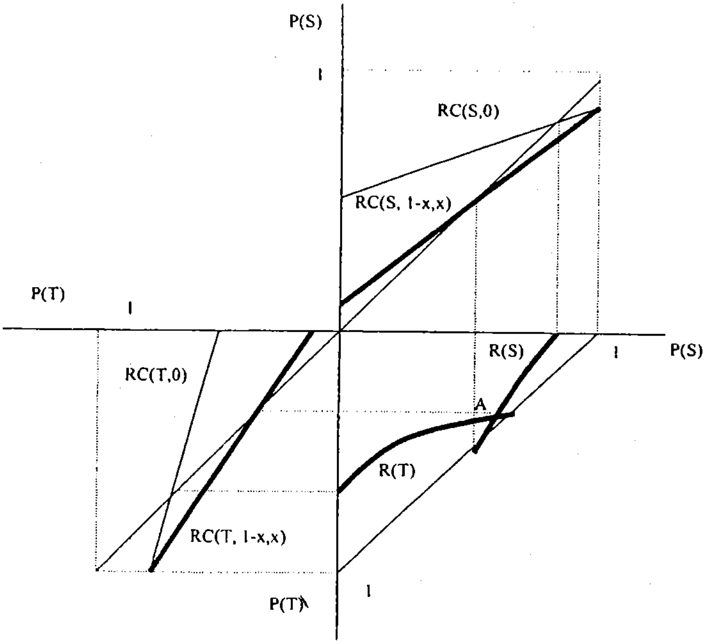


Figure 10

That some equilibrium must exist is shown by *reductio ad absurdum* in fig. 11, where *R(S)* and *R(T)* are assumed not to cross. If this were the case, we would have, for a point like *B*, of the form $(x_0, 1\text{-}x_0)$, and lying to left of *R(S)* and above *R(T)*, that $RC(S,x_0,1\text{-}x_0) > x$ and $RC(T,1\text{-}x_0,x_0) > 1\text{-}x$, and so, $RC(S,x_0,1\text{-}x_0) + RC(T,1\text{-}x_0,x_0) > 1$, which is impossible.

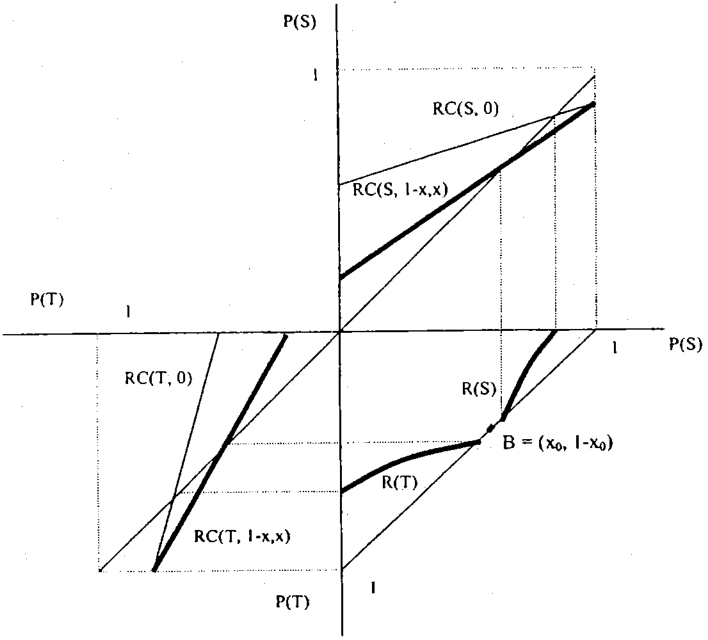


Fig. 11

Appendix B. Consensus formation when each scientist has a different weight in the scientific community

Let us examine firstly the situation arising when each scientist has a different weight w_i whithin her community, and this weight is commonly recognised by all her colleagues. In this case, we can change the variable *P(S)* in the utility function of each researcher for $\Sigma_{i\in A(S)}w_i/\Sigma_{i\in I}w_i$, that is, the proportion of the sum of the weights of those who accept *S* to the sum of all weights (this, on the other hand, may be normalised to be equal to the unity). The only possible values that this sum can take correspond to those rational numbers $a/\Sigma_{i\in I}w_i$ for which there is some subset *J* of weights such that $a = \Sigma w_{i\in J}$, and there are only a finite number of these subsets. The existence of an equilibrium is proven in this case analogously to the proof presented in section 5: if neither 0 nor 1 are equilibria, the first value of $a/\Sigma_{i\in I}w_i$ which is lesser than $RC(a/\Sigma_{i\in I}w_i)$ will be necessarily preceded by another value which is an equilibrium.

Let us suppose now that each scientists has a different weight in the utility function of the others, but that different researchers can attach a different weight to the same colleague. In this case, the utility functions of accepting and of not accepting S is better deffined on the possible *subsets* of the scientific community. The assumption that A_S is increasing with $P(S)$ and N_S decreasing is now transformed in the following: if i accepts S when it is accepted by the subset Q, then she will also accept S if it is accepted by any subset R in which Q is included. For any subset Q of the community, let $r(Q,S)$ be the subset of those who would decide to accept S if it were accepted by all the other members of Q (I will ignore henceforth the reference to S). This entails that, if $Q \subseteq R$, then $r(Q) \subseteq r(R)$. On the other hand, a social equilibrium is now a subset A of the community such that $A = \mathrm{r}(A)$.

In order to prove that some equilibrium must exist in this situation, suppose firstly that there is some subset A of the whole community N such that $r(A) \subseteq A$ (see fig. 12 for this and the following two cases). Let define $r^n(A)$ inductively in the following way: $r^1(A) = r(A)$, and $r^{n+1}(A) = r(r^n(A))$. In this case, for every n it will occur that $r^{n+1}(A) \subseteq r^n(A)$. Since the number of researchers is finite, we will have that $r^{m+1}(A) = r^m(A)$ for some m, and so, $r^m(A)$ will be an equilibrium. In the second place, suppose that there is some set B such that $B \subseteq r(B)$; in this case, we will have that $r^n(B) \subseteq r^{n+1}(B)$ for every n, and, for the same reason than before, $r^{m+1}(B) = r^m(B)$ for some m. Lastly, if C is not included in $r(C)$ nor *vice versa*, define ${}^1r(C)$ as $r(C)$ and ${}^{n+1}r(C)$ as $r(C \cup {}^nr(C))$. It can be shown that ${}^nr(C) \subseteq {}^{n+1}r(C)$ for every n, and so, that ${}^{m+1}r(C) = r(C \cup {}^mr(C)) = {}^mr(C)$ for some m. If C is included in ${}^mr(C)$, the former identity will be equivalent to $r({}^mr(C)) = {}^mr(C)$, and so ${}^mr(C)$ will be an equilibrium. If C is not included in ${}^mr(C)$, note that ${}^mr(C)$ is included in $C \cup {}^mr(C)$, and so $r(C \cup {}^mr(C))$ $[= {}^mr(C)] \subseteq C \cup {}^mr(C)$, what is an example of the first supposition (*i. e.*, the case where $r(A) \subseteq A$); so, there will be an m' such that $r^{m'+1}(C \cup {}^mr(C)) = r^{m'}(C \cup {}^mr(C))$, and this set will be an equilibrium.

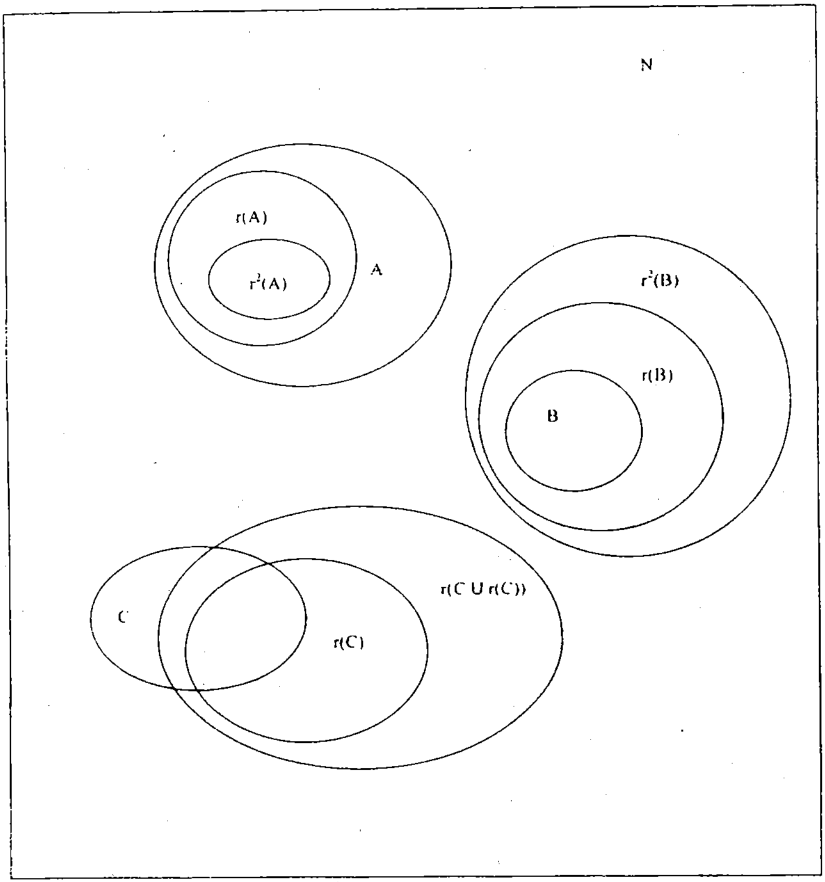


Fig. 12

Unfortunately, the existence of some equilibria does not guarantee now that one of them is necessarily reached. For example, we can have a pair of subsets *A* and *B*, such that $A \not\subset B$, $B \not\subset A$, A = *r(B)* and *B* = *r(A)*. In this case, if one of these subsets has been reached, the social situation will oscillate between both, without reaching an equilibrium (at least, while no changes in the empirical evidence take place).

Notes

† Research for this paper has been funded by the Spanish Government's DGICYT, as part of research project PB 95-0125-C06. A preliminary version was presented at the Permanent Seminar on Economic Methodology organised by Juan Carlos García-Bermejo at Universidad Autónoma de Madrid, in February 1998. I am grateful to prof. García-Bermejo for giving me the opportunity of discussing publicly my ideas. I have received also interesting and helpful comments from Francisco Alvarez, Miguel Beltrán, José Antonio López-Cerezo, Uskali Mäki, Juan Urrutia and José Luis Zofío.

[1] Some recent surveys are Dasgupta and David (1994) and Stephan (1996), as well as the monograph Wible (1998). Kitcher (1993) offers in its last chapter a detailed economic model of the scientific researchers' decisions, which is until now the most elaborate proposal from the philosophical side.

[2] Of course, two statements with the same rate of acceptance can actually have very different epistemic or social values; but these differences will not be important for the questions my model tries to answer, and so, I will ignore them here. On the other hand, by 'social value' I mean just the 'aggregate' value that a proposition has for the scientific community, as opposed to the value it has for an individual scientist; I do not refer here, of course, to the 'economic' value of a piece of knowledge for society in general.

[3] Of course, I refer only to those mechanisms not grounded on coercion. Kitcher (1993) correctly presents the study of science as the study of the evolution and working of its 'consensus mechanisms', but the formation of this consensus is an element not sufficiently present in his own economic model. In particular, he does not study the economic properties of that consensus.

[4] This insistence in the importance of the 'division of cognitive labour' is also an essential component of Kitcher's model.

[5] On the concept of 'increasing returns to adoption', see Arthur (1989). An intuitive example is the fax: think of the utility that buying a fax would have for you if you were the first person in adopting that innovation, and compare it with the utility of buying a fax if millions of other people already had one. In the same way, using the proposition S (in the context of an argument in favour or against *another* proposition) will be more useful for you if many people accept S than if it is a much more controversial statement.

[6] A different, but simmilar formalisation of this situation has been developped by Brock and Durlauf (1997). The main differences between my model and theirs are, first, that they assume a definite mathematical form of the utility functions, while my conclusions are based on more general assumptions, and are, hence, more 'robust'; second, that they design the researcher's decision as a choice between two theories, not allowing, hence, for the possibility of 'suspension of judgement', even if the epistemic values of both theories are very small; third, they base the researcher's decision on her expectations about her colleagues' decisions, instead of, as I do, on her observation of these decisions (this makes my model less sophisticated); and fourth, most of their conclusions essentially depend on the unrealistic assumption that the ratio between the epistemic values of each theory is the same for every researcher.

I thank professors Brock and Durlauf for having let me know their paper.

[7] A problem with this possibility is that, since I have assumed that EV_i is essentially 'private' and 'non-quantitative', it does not allow to make 'interpersonal comparisons of epistemic utility', so to say, and hence these private epistemic values would be hardly communicable in a public fashion. An obvious consequence of this is that EV_i can not be read simply as a Bayesian probability function (for example, two statements can have the same subjective probability and still have two different epistemic values, because one of them is simpler, or more general, and so on).

[8] For an explanation of this idea, see Woolgar (1988), ch. 5.

[9] I owe this observation to Juan Urrutia.

[10] See, for example, Banerjee (1992), where some scientific decisions (those corresponding to which lines of research to choose) are taken as typical examples of 'herding'. Banerjee's main conclusion (that herding can cause a negative externality by making the agents not to employ optimally all the information they privately possess) is indeed very similar to one of my efficiency results in section 8.

[11] Of course, *RC* is not a continuous curve, but a series of discrete points; nonetheless, if the relevant community is 'large', a line can serve as a useful graphical representation.

[12] If the peak of the distribution of indifference points is at one extreme (*i. e.*, at $P(S) = 0$ or 1), the reaction curve will be concave or convex, respectively.

[13] For the concept of path-dependence, see especially Arthur (1989) and David (1984). Though they talk mainly about technologies, most of their comments can be applied to knowledge in general, and to scientific knowledge in particular.

[14] In some stronger versions, many theories are equally well confirmed by the same empirical data.

[15] Another plausible interpretation of this phenomenon is that offered by Brock and Durlauf (1997): after a long period of discussion, consensus about a theoretical statement may emerge more or less suddenly, due to the effects that the *last* small changes in the empirical evidence have on the *total* evidence.

[16] An alternative approach could be to construct a 'social utility function' and a 'social epistemic utility function' (*e. g.*, as the sum of the utilities and epistemic values, resp., associated by each individual researcher to the acceptance of each statement), and to define the efficiency of a social situation as the maximisation of some of these functions. The conclusions from this alternative approach may be very different from those explained below, but I am very pesimistic about the possibility of defining those social utility functions in a consistent and operational way, since individual utilities and epistemic valuations are simply the numerical expression of comparative orderings, and not really additive quantities. Of course, my pessimism does not preclude that an interesting use of that approach might be made.

[17] See, for example, Latour and Woolgar (1979) and Knorr-Cetina (1981).

[18] I have ignored the possibility that some researchers are indifferent between *A* and *C*, because their number will be normally lesser than that of the groups *c* and *d*.

[19] Recall that this 'efficiency' is always understood in a Paretian way, and not by reference to the attainment of a particular epistemic goal.

BIBLIOGRAPHY

Arthur, B.W.: 1989, 'Competing Technologies, Increasing Returns and Lock-in by Historical Events', *The Economic Journal* 39, 116-31.

Banerjee, A.V.: 1992, 'A Simple Model of Herd Behavior', *The Quarterly Journal of Economics* 107, 797-817.

Brock, W.A., Durlauf, S.N.: 1997, 'A Formal Model of Theory Choice in Science', Social Systems Research Institute, Working Paper 9707, University of Wisconsin, Madison.

Dasgupta, P., David, P.A.: 1994, 'Toward a New Economics of Science', *Research Policy* 23, 487-521.

David, P.A.: 1984, 'Clio and the Economics of QWERTY', *American Economic Review* 75, 332-37.

Kitcher, P.: 1993, *The Advancement of Science*, Oxford, Oxford University Press.

Knorr-Cetina, K.D.: 1981, *The Manufacture of Knowledge*, Oxford, Oxford University Press.

Kuhn, T.S.: 1962, *The Structure of Scientific Revolutions*, Chicago, The University of Chicago Press.

Latour, B. Woolgar, S.: 1979, *Laboratory Life. The Social Constructions of Scientific Facts*, London, Sage.

Stephan, P.E.: 1996, 'The Economics of Science', *Journal of Economic Literature* 34, 1199-1235.

Wible, J.R.: 1998, *The Economics of Science. Methodology and Epistemology as if Economics Really Mattered*, London, Routledge.

Woolgar, S.: 1988, *Science: The Very Idea*, London, Ellis Horwood.

Jesús P. Zamora Bonilla is doctor in Philosophy and graduate in Economics by the Universidad Autónoma de Madrid. He is currently Associated Professor at the Department of Economics and in the Programme of Doctorate in Humanities, in the Universidad Carlos III de Madrid. His main present line of research is the economics of scientific knowledge, though he has also made research on the problem of verisimilitude, having published the book *Mentiras a medias. Unas investigaciones sobre el programa de la verosimilitud* (Ediciones U.A.M., 1996), and several articles, for example 'Verisimilitude, structuralism and scientific research' *(Erkenntnis,* 1996), 'An invitation to methodonomics' (*Poznan studies in the philosophy of science*, 1997) and 'Truthlikeness, rationality and scientific method' (*Synthese*, forthcoming).