
Science Studies and the Theory of Games

Jesús P. Zamora Bonilla

Universidad Nacional de Educación a Distancia and Fundación Urrutia Elejalde (Spain)

Being scientific research a process of social interaction, this process can be studied from a game-theoretic perspective. Some conceptual and formal instruments that can help to understand scientific research as a game are introduced, and it is argued that game theoretic epistemology provides a middle ground for 'rationalist' and 'constructivist' theories of scientific knowledge. In the first part ('The game theoretic logic of scientific discovery'), a description of the essential elements of game of science is made, using an inferentialist conception of rationality. In the second part ('Sociology of science and its rational reconstructions'), some ideas for the reconstruction of case studies are introduced, and applied to one example: Latour's analysis of Joliot's attempt to build an atomic bomb. Lastly, in the third part ('Fact making games'), a formal analysis of the constitution of scientific consensus is offered.

Research for this paper has benefited from Spanish Government's research projects PB98-0495-C08-01 ('The culture of techno-science'), BFF2002-03353 ('Cognitive roots in the assessment of new information technologies'), and HUM2005-01686/FISO ('The emergence of technoscientific norms'), as well as from the Spanish Foundation for Science and Technology (FECYT) financial support. Parts of it were presented and discussed in a seminar on 'The Constitution of Science' at Alpbach European Forum, and in seminars at several universities: Padova, Erasmus (Rotterdam), UNAM (Mexico), UNED (Madrid), Coruña (Ferrol), UAB (Barcelona), and UAM (Madrid), as well as at the 12th Congress of the Division of Logic and Philosophy of Science (Oviedo). Thanks are given to the organisers and participants in those events, and specially to Max Albert, Paco Álvarez, Salvador Barberà, Robert Brandom, Fernando Broncano, Javier Echeverría, Anna Estany, José Luis Ferreira, Juan Carlos García-Bermejo, Donald Gillies, Wenceslao González, Ian Jarvie, Uskali Mäki, León Olivé, Ana Rosa Pérez-Ransanz, Carlos Solís, Mauricio Suárez, David Teira, Juan Urrutia and José Luis Zofío, and particularly to Mauro Dorato, Esther-Mirjam Sent, and an anonymous referee, who made extensive criticisms on earlier versions. Comments to the author's e-mail address (jpbz@fsof.uned.es) are welcome.

Perspectives on Science 2006, vol. 14, no. 4
©2007 by The Massachusetts Institute of Technology

1. The Game Theoretic Logic of Scientific Discovery

In this paper it is argued that some philosophically relevant aspects of science can be illuminated by considering scientific research as a type of game played by rational individuals. The main advantages that may derive from this approach are the following: first, two conceptions of science apparently in deep conflict (rationalist epistemology and post-modern constructivism) can be convincingly presented as *complementary* accounts of a single but complex phenomenon; second, an epistemology more suitable to be connected with questions about science policy can be developed; and third, some powerful conceptual tools can be developed to attempt to a more rigorous analysis of case studies and of epistemological problems. Though some important efforts have been already made in the construction of a 'game theoretic epistemology' (GTE; see esp. Kitcher 1993, ch. 8), neither philosophers nor sociologists of science have paid too much attention to them, perhaps because of the apparent difficulty in mastering game theory's formalisms. Hence, an essential goal of this paper will be to 'publicise' the basic concepts of GTE, both by showing their relevance to a philosophical understanding of the social production of scientific knowledge, and by analysing a famous case study with the help of game theory.

1.1. Scientists as players.

According to rational choice theory, when people have to make a significant choice, they reason about the advantages and disadvantages of every option they can conceive, in order to select the one whose expected net benefit is the biggest. Nevertheless, as it is clear from everyday experience, making choices is a painful activity, and much of our social life is organised in such a way that only a limited number of alternatives are to be taken into account each time, and only a limited number of criteria need to be used in choosing among them; customs, rules and habits help us in making that decisions become easier to make, often by moving us to a salient option almost without any conscious thinking. In spite of this, the need to make hard choices is unavoidable in many cases, and rational choice theory assumes that people's decisions will 'ultimately' be consistent with the most exhaustive fulfilment of their goals which is possible given their knowledge of their circumstances, whatever these goals may be. Matters become still more complicated when costs and benefits do not only depend on our own choice, but also on decisions taken by other individuals. The main complication arises because other people's capacity of making choices adds a lot more of uncertainty, not only because human behaviour may be more difficult to predict than other types of facts, but because there is a problem of circularity: to know what choice is the best one for me, I need to know what the other agents will do, but they will

need to know my own decision before deciding what choices are the best for them!

Economists call 'games' these situations of interdependence.¹ This word can be misleading, because many of those situations are far from being amusing, and not every game demands such a type of strategic reasoning by part of the players (e.g., solitaire, target shooting). However, 'game theory' has become the sanctioned name for the investigation about what rational people is expected to do when what each individual attains depends on what somebody else does. Game theory's most important idea is that a *necessary* (but not sufficient) condition for a social situation to be *stable* is that it corresponds to a *Nash equilibrium*. In a Nash equilibrium *nobody has an incentive to change her behaviour from what she is doing* (i.e., nobody has a better decision to make), *given the decisions made by the rest*. Game theory's basic assumption is that, if a situation were *not* a Nash equilibrium, then somebody would soon discover that she might improve her situation by changing her behaviour, and then the situation would change.² This solves the circularity problem because every chain of reasoning from the choice of one individual to the best option for another always points to the same result. As it is well known, the bad news are that in many cases more than one set of decisions can be a Nash equilibrium. Game theorists have made a big effort to find out some stronger conditions to determine what situation will be attained in such cases (e.g., Selten's 'subgame perfection', or Mayr's 'evolutionarily stable strategy' concepts), but here I will keep my discussion at a very elementary level, leaving for further work questions related to stronger equilibrium concepts. Some interesting philosophical consequences follow just from the assumption that the regular interactions of scientists must be Nash equilibria, as well as from the possible existence of more than one equilibrium.

Classical theories of scientific method, from Bacon and Descartes to Carnap and Popper, shared what we could call the 'Robinson Crusoe approach to scientific method', for they assumed that the rules of research have to be explicated in such a way that, *in principle*, one single individual could follow them, arriving necessarily to the same results a group of col-

1. For details, readers are referred to any of the many available introductions to game theory. Two books which offer clear discussions of the main philosophical and methodological topics related to game theory are Binmore (1992) and Hargreaves Heap and Varoufakis (1995); the last one can be particularly useful for those lacking a firm mathematical basis.

2. Or, as David Kreps puts it: *if* the players of a game have some *clear* way of playing it, then it *has* to be a Nash equilibrium (Kreps 1990, pp. 30–1). This means that in game theory we must acknowledge the possibility that players may find *no* way of playing an equilibrium, for example, if the context is too complicated. But, as I told at the beginning, the role of social norms or customs can be understood as a way to make the situation simple enough as to warrant that an equilibrium is devised and reached. See Bicchiere (2006).

laborating scientists would reach (although the latter would make it much more quickly). This is true even in the case of Popper's falsificationism, for, even if competition between researchers may have a beneficial epistemic role by incentivating the performance of severe tests of the rivals' hypotheses, this is something that a 'honest' scientist should do by herself with her own theories. Instead, sociologically oriented authors have insisted in the essential role that cooperation and competition, trust as well as strategic thinking, play in the everyday work of scientists, and particularly in the determination of *what research episodes deserve to be taken as 'discoveries'*. Just to quote some prominent examples, Kuhn showed that a researcher's decisions are affected by how she thinks her ideas will be received by the members of her community: if colleagues are very confident in their 'paradigm', heterodox hypotheses will not be taken into account, whereas if that confidence has been weakened by the proliferation of 'anomalies', new ways of looking at things can be welcome; Bloor and Shapin convincingly presented many scientific disputes as being part of larger political quarrels, so that theories are accepted or rejected according to what social interests they may serve to promote; Collins established that a scientific 'discovery', understood as a social process, is coextensive with the decisions about the existence of the discovered phenomenon; Latour and Callon explicitly described scientific argumentation as a matter of 'alliance making', and the process of looking for allies as one constituted by concessions and threats; Knorr-Centina and Galison explained consensus formation about experimental results as a continuous process of negotiation, and so forth. Displaying an impressive amount of detailed empirical evidence, gathered through historical or field case studies, some of these authors have concluded (or have not avoided speaking in a language that makes this conclusion inescapable) that what is taken as scientific knowledge is 'just' a social construction,³ i.e., something resulting mainly from the 'negotiation' between conflicting interests, with little or no constraining role left to nature itself.

Nevertheless, in spite of the frequent reference to 'strategic' or 'entrepreneurial' behaviour by part of scientists, sociological explanations of science have not included until now an explicit game-theoretic analysis of those 'negotiations' and 'conflicts'. Actually, those who have made a more systematic effort in understanding scientific research as an interaction between rational agents have been philosophers with a strong anti-constructivist orientation, as David Hull, Alvin Goldman and Philip

3. Among the cited authors, Galison would be the main exception. Many of the other authors feel uncomfortable with an accusation of 'relativism', but it is very difficult to interpret many of their statements in a different way.

Kitcher.⁴ Their main contribution has been to show that a fierce competition between recognition-seeking researchers does not entail *per se* that the 'knowledge' produced by them is of little epistemic value; rather on the contrary, under some procedures for scientific merit attribution, competing scientists can be remarkably efficient in the production of high-quality knowledge, and, as I have shown elsewhere (Zamora Bonilla 2002), recognition-seekers may even have a preference for playing under epistemically stringent rules, in order to make of science an *interesting* game to play. As I will try to show here, this type of analysis is not a mere *alternative* to sociological or historical case studies; it is, rather, an *unavoidable* step when you begin to think of scientists as intelligent agents who see themselves as engaged in a web of social connections.

1.2. Scorekeeping in science.

What are, then, the basic elements in the game of science? My proposal is to describe scientific research as a game of *persuasion*. That language is extremely important to science can hardly be denied. Authors as different as Carnap and Latour would agree at least on this point, though for completely different reasons. My perspective is closer to Latour's in the sense that I will assume that interaction between researchers mostly takes place through a constant examination and evaluation of what each other *says* or *writes*, although I guess that scientists may agree to evaluate their colleagues' 'inscriptions' (to use Latour's word) by means of some criteria a Carnap would not dislike too much. This does not mean, however, that other things besides language are unimportant. Scientists also perform non-verbal actions: they earn and spend money, organise meetings, perform experiments, and so on, though it is true as well that a big part of these things is made *by* speaking or writing, that many of them are made in *expectation* of what one or the others will say, and that people's *assertions* are usually more public (and easier for others to scrutinise) than their *actions*.

That communication is central to the strategies of scientists is not only consistent with a big part of the work in the sociology of science of the last three decades, but is also close in spirit to some recent proposals in the philosophy of language. I am referring particularly to Robert Brandom's *inferentialism* (Brandom 1994). According to this theory, what makes a series of noises or marks to count as an *assertion* is the chain of inferences the

4. See Hull (1988), Goldman and Shaked (1991), and Kitcher (1993). This approach is related to some other projects in the 'new economics of science'. Some useful surveys are Stephan (1993), Sent (1999), Hands (2001, ch. 8), and Mäki (2004), as well as the papers collected in Mirowsky and Sent (2002).

speech community takes as *appropriate* to make regarding that assertion, inferences which essentially relate to the *normative* status that each participant in a conversation attributes to the others. For example, my saying 'there is a cat on my roof' can be *taken* as an assertion by my hearers if and only if we share a set of normative inferential practices which allow them to attribute to me, under specified circumstances, the 'obligation' of presenting some relevant evidence from which that sentence can be derived, as well as that of accepting the linguistic or practical consequences which, together with other commitments I have made, follow from it. Using a metaphor suggested by Wilfried Sellars, understanding an expression would amount to mastering its role 'in the game of giving and asking for reasons'. It is important to mention that Brandom's concept of 'inference' does not only cover moves from sentences to sentences, but also from 'inputs' of the language game (e.g. observations) to sentences, as well as moves from sentences to 'outputs' of the game (e.g., actions).

The aspect of Brandom's theory I want to emphasise here is that linguistic practice takes place through each speaker 'keeping score' of the commitments made by the others and of the actions allowed or commanded by those commitments, according to the *inferential rules* defining the language games which are possible within their speech community. These inferential rules need not be explicit for the players: the only relevant assumption is that they are able to recognise when a move in the language game has been made 'properly' or 'improperly', and to respond appropriately to the commitments they have made. In the case of science, I suggest to consider the 'inscriptions' produced by a researcher as her set or 'book' of commitments (her 'book', for short). There is no need that every such commitment amounts to the bare acceptance of a certain proposition (say, *A*)⁵, for it is possible to make a variety of *qualified* (or 'modalised') commitments, as 'it seems likely that *A*', 'there is some evidence that *A*', '*A* deserves some attention', and so on. *The game theoretic nature of scientific research arises because each scientist's payoff directly depends on what is 'written' on the books of the other members of her community.* This payoff is generated along three interconnected channels, which I will call 'internal score', 'external score', and 'resource allocation mechanism', all of which are determined by several types of norms. These norms are *social conventions*, in the sense that their existence amounts to their (explicit or tacit) acceptance by the members of the relevant scientific community, who may also have the capacity of contesting the norms and of making different interpretations of them.

The basic structure of the game is the following: in the first place, any

5. By the way, those inferential norms have to include some criteria to specify when several 'inscriptions' amount to 'the same' assertion.

scientific community will have adopted (tacitly or explicitly, unanimously or not)⁶ a set of *methodological norms* with which to assess the scientific value of any set of commitments, and particularly, the appropriateness of each inferential step from a given set of commitments to a new assertion or action. The coherence of a researcher's book with these norms (or, more precisely, the coherence *her colleagues say* it has) will determine the *internal score* associated to that book. In the second place, and contrarily to the case of everyday language games, in science many propositions are attached to the name of a particular scientist, usually the first one who proposed it; one of the main rewards a scientist receives is associated to the fortune that the theses (laws, models, experimental results, and the like) proposed by her have in the books of her colleagues; this 'fame' is her *external score*. Put in different words: a scientist's internal score expresses basically her *professional competence*, whereas the external score expresses the *authoritativeness* of the results advanced by her. I will call 'global score' the combination of both elements. In the third place, the community will work with a set of *norms for resource allocation* which will determine how much money, what facilities, what work conditions, what assistants, and so on, each scientist will be allotted, depending on her global score. Of course, behaviour which is not appropriate according to the allocation rules will also make downgrade a researcher's internal score.

So viewed, the game of scientific research proceeds as follows. The norms of her discipline tell each researcher what things can she do (or must she do) in order to write a book with a high *internal* score; if she succeeds in this, she will count as a more or less 'competent' researcher. These norms are about how to perform and report experiments, what formal methods to employ and how, what types of inductive or deductive inferences are appropriate, what styles of writing are acceptable, and so on. By following these norms, she will end having committed herself to the acceptance of some propositions advanced by other colleagues, hence contributing to *their* having a high *external* score. She will also have to pronounce about the coherence of her colleagues' commitments with the methodological norms of the discipline, contributing to rising or lowering *their internal* score. On the other hand, in order to get herself a high *external* score, she has to take advantage of her colleagues' being trying to attain a high *internal* score: she has to be able of devising experiments, hypotheses, or models that, *given the commitments her colleagues have already made, and given the accepted methodological norms*, these researchers would probably suffer a severe reduction of their internal score if they refused to

6. Some minimum of unanimity is needed in order to talk of *one* community, but deep disagreements may exist about some rules and about what their right interpretation is.

accept the former scientists' proposal. Nevertheless, since one researcher's external score is dependent on the decisions of her colleagues, the effect of one of her decisions on her external score will always be much more uncertain than its expected effect on her internal score. Lastly, there may be cases where the same strategy leading to a high internal score tends also to improve the external score (as in Kuhn's 'normal science'), whereas in other cases one will have to 'break' some rules (and hence probably getting a lower internal score) in order to reach a result that may warrant a high external score (as in Feyerabend's 'contrainduction'). All this simply means that methodological norms are there to be used strategically.

1.3. Playing the game.

An open question in my description of the game is where do the scientific norms of a research community come from. One possible answer is consensualist: the members of a scientific community may collectively decide under what rules to play the game, and this collective decision can obviously be analysed from a game theoretic perspective. But other approaches are also possible. For example, according to an evolutionary point of view, many 'rules' are simply the result of (nearly) unanimous commitments made according to previous norms. Note, however, that the evolutionary and the contractarian approaches are not incompatible: many norms can have *both* types of justification. Furthermore, seeing norms as a type of commitment (i.e., as an entry in a scientist's 'book') does not suppress the fundamental difference between '*making* an assertion' and '*assessing the appropriateness* of that assertion' (which is what scientific rules are for); it only means that 'asserting that one accepts a given norm' is something whose appropriateness can itself be judged by using *other* norms. In the rest of the paper, however, I will not discuss again about the origin of rules; instead, I shall put a different, and not less important question: to what extent is it rational for individual scientists to obey the rules governing their research processes?

This question has received a different answer from rationalist philosophers and functionalist sociologists, on the one hand, and from relativist philosophers and social constructivists, on the other hand. The game theoretic perspective allows to see why both answers are problematic. In the first place, the traditional 'rationalist' point of view may have been that the *cognitive* virtues of a given scientific rule are a sufficient reason for individual scientists to conform to the use of that rule. The problem is, of course, that disobedience may sometimes provide some obvious advantages, particularly if the chances of not being discovered are high. I can manipulate experimental results, or fail to put enough effort into my work, or fail to disclose some information the norms command to publish,

and so on. The sociological literature is full with case studies showing how scientists 'misbehave', at least according to the rules *they* (scientists) preach, not to say regarding the rules preached by the philosophers.⁷ The persistence of an institution like science, where most things depend on the *trust* people put on other people's assertions, demands, however, that this misconduct is severely limited, and, surprisingly, science seems to attain this goal rather well even in the absence of something like a 'police' or a 'judicial system'.

In the second place, a typical 'constructivist' attitude towards scientific norms has been to take them either as mere rhetorical devices, or as mechanisms for benefiting some privileged group. In this case the problem is that, although this approach can explain why some people may have an interest in *proposing* some norms, it does not explain why the *other* people, knowing that the norms are just rhetorical strategies for defending the interests of others, actually *obey* these norms. Stated in game-theoretic terms: the situation in which some group 'imposes' a norm to a bigger collective clearly fails to satisfy the condition for being a Nash equilibrium, for the rational thing to do for those people for which the norm is harmful would be to disobey it, particularly if they grossly outnumber the other group. Actually, a system of norms will be *stable* only if it constitutes a Nash equilibrium, i. e., only if, under the assumption that the others are obeying the norms, anyone's best option is also to obey. For example, given that most people speak a certain language in a country, it will be in my, and *every other's* interest to do the same; given that judges and policemen do efficiently their work according to the prevailing civil and criminal laws, it will be in my and *their* interest to obey these. As it is clearly shown in this example, when 'obeying certain norms' includes 'punishing those who do not obey', general compliance with the rules can be expected (cf. Axelrod (1984), Elster (1989)). In the case of science, this reflects in the fact that a researcher's book is permanently being evaluated by other colleagues in their respective books, which are evaluated by other scientists, and so on; for example, I will be punished if my model violates the law of energy conservation, but also if I *fail to criticise* a colleague whose model (which I may be using or discussing) makes this mistake. So, the fact that a norm is followed by most of my colleagues makes disobedience very costly for me.

The question is, hence, whether the mechanism of mutual check described in section 1.2 is strong enough for deterring researchers from systematically disobeying the prevailing rules. If the answer were 'no', so

7. See, for example, Altman and Hernon (1997). See also Wible (1997) for a good rational choice approach to the study of scientific fraud.

that researchers faced permanently a ‘prisoner dilemma’ when deciding whether to obey the norms,⁸ then either the public trust in scientific results would be much more fragile than what is usually presumed by the scientific rhetoric, or the apparent stability of so many portions of scientific knowledge would just be based on scientists’ exceptional honesty. I hope, however, that the toy model presented in this section opens a possibility for avoiding those conclusions. The model must not be understood as a description of empirical cases, but just as a kind of ‘how possibly’ argument, intended to show that *in a community playing a game similar to the one described in the past section, it is reasonable to expect that the norms will not be disobeyed ‘too often’.*

Let f be the frequency with which a researcher disobeys the norms, and suppose, for simplicity, that all infringements are equally important (if this is not the case, then f can be alternatively interpreted as a normalised average of an individual’s infringements). Let $u_i(f)$ (> 0) be the utility received by scientist i if she disobeys the norms with frequency f and is *not* discovered, and let $-v_i(f)$ (< 0) the disutility she gets if discovered and hence punished. The probability of being discovered ($p_i(f)$) is an increasing function of f . An individual’s expected utility from disobeying the norms with frequency f will hence be $EU_i(f) = (1 - p_i(f))u_i(f) - p_i(f)v_i(f)$, and the optimum infringement frequency for her will correspond to that value of f which maximises $EU_i(f)$. On the other hand, it is reasonable to assume that an individual’s utility also depends on the frequency with which the norms are disobeyed by *other* researchers: the more frequently norms are infringed by your colleagues, the less utility will you get from the same action, so that, in principle, a situation where f is low for all is better for everyone than a situation where f is high for all. The question is whether, in the equilibrium, the f ’s will be ‘high’ or ‘low’. In order to answer this question, I will add some simplifications. First, suppose that $p_i(f)$ is just equal to f (i.e., the probability of being discovered is the same as your frequency of infringement). Second, assume that $u_i(f)$ and $v_i(f)$ are linear functions of f ; in particular, $u_i(f) = a_i + b_i f$, and $v_i(f) = c_i f$ (with $a_i, b_i, c_i > 0$; these coefficients may be different for each researcher, for these may have different opportunities or abilities for engaging in successful infringement of norms; for simplicity, I will omit the subindex i when no confusion can arise). Lastly, i ’s utility will also depend on the *average* frequency of infringement within the rest of her community, \mathbf{f} , so that $u(f, \mathbf{f}) = (1 - \mathbf{f})(a + b\mathbf{f})$ (i.e., even if i ’s infringements are *not* discovered, she gets

8. In the ‘prisoner dilemma’ game, each player finds always more profitable to defeat the other players, no matter what these may do. For an interesting application of this game to scientific examples, see Luetge (2004).

a null utility if norms are always disobeyed by her colleagues), whereas $v(f)$ does not depend on f (i.e., she will be equally punished by her infringements, independently of how frequently her colleagues disobey the norms; since ‘punishment’ basically consists in depriving you from some resources, your colleagues will always be interested in ‘punishing’ you as soon you give them a chance).

The ‘solution’ of this game is expressed in the following theorem (the proof is given in the appendix; $f_i^*(\mathbf{x})$ is the optimal frequency of infringement for scientist i when the average infringement frequency is \mathbf{x}):

- (i) There is only one Nash equilibrium. It occurs when the average frequency \mathbf{e} of disobedience is such that the average of all the $f_i^*(\mathbf{e})$ equals \mathbf{e} . (1)
- (ii) $\mathbf{e} \leq (b - a)/(2b + c) < 1/2$.
- (iii) This equilibrium is stable.

The good news are that a stable equilibrium exists (the game must be played in a coherent way, and only in that way), and that mispractices are ‘limited’: the more valuable are the advantages you have from obeying the norms (a), and the more severe is the punishment from disobeying them (c), the less average infringement there will be in the equilibrium. The bad news are, of course, that some positive degree of infringement will always exist (save if $a \leq b$ for all researchers), but it wouldn’t have been realistic to expect that real scientific communities are ‘perfect’.

2. Sociology of Science and Its Rational Reconstructions

2.1. Looking for Nash equilibria in case studies.

In this second part of the paper I will defend a particular, and not too costly way, in which the game-theoretic approach to science can be put to use. I hope it may give rise to a number of works allowing us to better understand examples of scientific practice, and to make a normative assessment of case studies. The idea is very simple. First, we can take any paper or book which describes a research episode, and try to extract from it the *goals or preferences* of each relevant actor (scientists and non-scientists), the options we can reasonably think they believed to have, and finally, the results they could expect for each possible combination of decisions. This amounts simply to formally reconstruct the episode under the form of a strategic game. Second, we can try to see, for each actor, whether her *actual* decisions were the best ones *from the point of view of her own goals*, given the actions of the other agents. This amounts to checking whether the actual history of the episode was a Nash equilibrium. If the answer to the second

question were 'no' for at least one of the actors, then we would have to make a reasoned choice among the following four options:

- a) *our game-theoretic reconstruction* of the sociological or historical description of the episode is flawed; we will surely have to pay more attention to some details;
- b) our reconstruction is all right, but the *initial sociological or historical description* of the episode was not correct: either the goals, the options, or the choices of some researchers were not exactly as presented in that case study;
- c) both our reconstruction and the original description of the episode are all right, but *at least one participant was behaving irrationally* (i.e., she was not wise enough to determine which was the best option for her, and to act accordingly); and
- d) the original description, our reconstruction of it, and the behaviour of the participants, are all of them all right, but some convincing reasons can be given for why an outcome which is not an equilibrium can be 'reasonable' in this case.

On the other hand, if the answer to our question is 'yes' for every actor, the new questions will be *whether other different equilibria were possible*, and (if this was the case) *why the actual equilibrium was obtained instead of others*. This question is particularly relevant if we are able of showing that one of the other equilibria would have been more beneficial for every, or for most of the people engaged, or from any other normative point of view. In particular, we may add to the goals of those agents the values and preferences of any other people whose point of view we *freely decide* to take as relevant although they were not actual players of the game (for example, some 'oppressed group', or 'the average citizen', when we can make a definite sense of that expression, or 'the epistemologist we are', if what we want is to make a cognitive evaluation of the research process), and see whether the outcome of the game has been 'beneficial' for those other agents or not. I think this would really be the most important advantage of these reconstructions, for this strategy would simply force 'science studies' to assert *explicitly* in what a sense the actual course of a research episode has been 'good' or 'bad' for exactly what people, and, more importantly, *how would it have been better*, what decisions the agents engaged in the episode *should have taken* in order to produce the most beneficial outcome from the type of values we have decided to put forward. This would lead us to suggest mechanisms that might have changed the structure of the game in such a way that our 'privileged' values would have been benefited. Of course, this would not be an 'objective' evaluation, for an evaluation is always made according to *somebody's* values, but the game-theoretic strategy does not

force us to be 'objective' in a transcendental sense, only to be explicit about the point of view from which we are making our assessments. In the next subsection I will exemplify this game theoretic approach by offering a simplified reconstruction of a famous case study analysed by Bruno Latour. It must be understood that mine is not directly a reconstruction of the historical events Latour refers to (if this expression means something at all from Latour's perspective), but just of his own actor-network-theoretic reconstruction of it.

2.2. The *École des Mines* game.

The 'actor-network theory' developed by Michel Callon and Bruno Latour, with its insistence on the essential role of 'negotiations' between different actors, is perhaps the sociological approach that can more easily be translated into the conceptual framework of game theory. Furthermore, this translation will show that many of those authors' most radical assertions can be understood, as they have many times declared, in a way that makes them very close to common sense. The case I have chosen is Latour's study of Frédéric Joliot's attempts to build an atomic reactor by the end of the 30's (Latour 1999, ch. 3), in which two basic ideas of 'actor-network' theory are clearly exemplified: the phenomenon of the mutual 'translation' of the interests of different agents, and the thesis that humans and non-humans play not too different a role in the construction of scientific knowledge (so that both can be placed under the single category of 'actants'). In both cases I will try to show that these ideas admit a natural game theoretic explanation.

Latour defines the concept of *translation* as referring "to all the displacements through other actors whose mediation is indispensable for any action to occur (. . .); chains of translation refer to the work through which actors modify, displace, and translate their various and contradictory interests" (1999, 311). As I understand it, this notion reflects the fact that, contrarily to some simplistic descriptions of scientific activity, the connection between the goals of a researcher and the means she can use is never (or not usually) a direct one, but needs to be constructed through a continuous negotiation with other people who are simultaneously pursuing their own goals. In its turn, the notion of an *actant*, as applied to non-humans, refers to the fact that these are not something whose behaviour can be taken for granted before the research begins, but which 'emerge' in this process through the series of trials to which they are subjected. Regarding our case study, two significant goals of the (human) actors considered by Latour are, roughly put, Joliot's desire of being the first in attaining a controlled chain-reaction (and, amongst other things, a Nobel prize thanks to that), and the then French Minister of Armaments, Raoul Dautry, who at-

tempted to build a powerful weapon for promoting France military independence. Both actors were placed in a position where an arrangement between them was possible, but, as in most bargainings, both have to give up something, redefining and changing some of their goals during the course of the negotiation. As Latour puts it (p. 88),

when Joliot met Dautry he did not particularly try to change Dautry's goal, but to position his own project in such a way that Dautry would see the nuclear chain reaction as the *fastest* and most certain way of achieving national independence . . . This transaction is not of a commercial nature. For Joliot is not a question of selling nuclear fission, since it doesn't even exist yet . . . Both men believe that, since it is impossible for either to achieve his goal directly, political and scientific purity are in vain, and that it will thus be best to negotiate an arrangement that modifies the relation between their two original goals.

The operation of translation consists of combining two hitherto different interests . . . to form a single composite goal . . . Neither of the parties . . . will be able to arrive at *exactly* his original goal. There is a drift, a slippage, a displacement, which, depending on the case, may be tiny or infinitely large.

Besides the naiveté of assuming that, for an arrangement to be a 'commercial' one, the good which is sold must already exist, the 'economic' nature of the negotiation between Joliot and Dautry is transparent, since they are simply exploiting 'the gains from trade': each one can do something that can benefit the other, and both are better off *after* the exchange than before—or so they expect by the time the arrangement is being made—, in spite of having had to renounce to something (Dautry gave up lots of valuable resources, and risked that Joliot did not put as much effort in building the bomb as the former wanted; Joliot had to reorder his own research priorities, and had to tolerate other intrusions by the military). The arrangement between both actors could be reconstructed in a very simplified way as in the game depicted in sequential form in fig. 1. In the first place, Dautry has two options: to fund Joliot's project (F), or not (N). If he funds it, then Joliot has also two options: to change his priorities and investigate first about the chances of building the bomb (B), or to develop just his first plan about the controlled chain-reaction (C). Assuming that Dautry will have the chance of monitoring to some degree what Joliot is doing, the Minister has again a couple of options later on: to keep on with the funding (K), or to stop it (S). The numbers associated to the outcomes simply reflect the order of preferences of each actor (first, Dautry; second, Joliot), assigning arbitrarily the level 0 for the case where

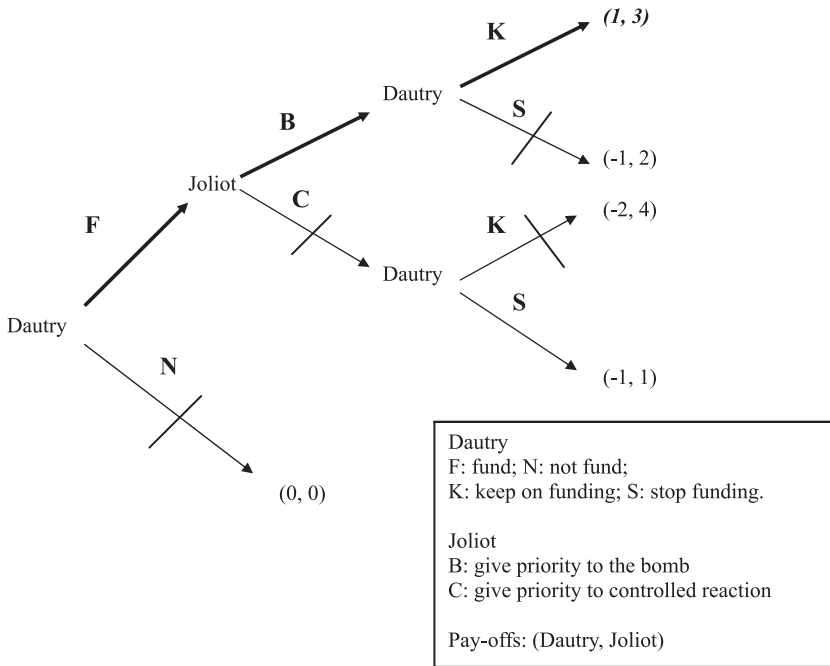


Figure 1.

no funding is given. The game is solved by ‘backwards induction’: thinking, at each final node, what would be preferred by the agent who is choosing at that point, eliminating the options not preferred (these options are crossed in Figure 1.a), and putting the same question at the previous nodes. The solution, and hence the equilibrium of the game (thick arrows), is when Dautry funds Joliot, Joliot makes the research preferred by Dautry, and Dautry keeps on funding Joliot.

In Figure 2 the game is represented in strategic form. The strategies of Joliot are B and C, and those of Dautry have the form ‘fund, keep on funding if Joliot chooses B, and stop if he chooses C’ (FKS), ‘fund, stop if he chooses B and keep on if he chooses C’ (FSK), ‘fund, and keep on no matter what he chooses’ (FKK), ‘fund, and stop no matter what he chooses’ (FSS), and finally, not fund (N).⁹ The solution appears more clearly here as

9. More formally, we would have to include strategies of the form ‘not fund, and keep-on funding if Joliot chooses B’, and so on, for a strategy, in orthodox game theoretical terms, has to include a prevision for all points of the game in sequential form, even if what is cho-

Dautry

		FKK	FKS	FSK	FSS	N
Joliot	B	1	I	-1	-1	0
	3	3	2	2	0	
C	-2	-1	-2	-1	0	
	4	1	4	1	0	

Figure 2.

the only Nash equilibrium of the game, for B (adjusting to Dautry's plans) is the best strategy for Joliot if Dautry chooses FKS (funding Joliot, and keeping on funding if and only if Joliot has chosen B), and this is the best strategy for Dautry if Joliot chooses B.

Of course, this game is extremely simplified, and is not faithful to the large amount of detailed decisions which the actors involved had to make, nor to the presence of other relevant actors; but my aim here is just to show how a plausible reconstruction could look like, and how it can illuminate some important aspects of actor-network theory. For example, the fact that, in a 'translation', no one of the actors can usually reach his most preferred option: for Joliot this is shown by the fact that the best outcome would be where Dautry funds him and he could carry out his own original plans (i.e., when Dautry chooses F and K, and Joliot chooses C), whereas for Dautry, the option he reaches in the equilibrium of the game is the best among those available in the game, but he would have preferred to attain his goals while keeping the resources given to Joliot (reaching a utility level of 2, say). Another important aspect of the translation concept is, however, not as clearly represented in my example: it is the *communication process* which allows each actor to know most of the options available to the others. This, nevertheless, could in part be analysed as a *previous* game, in which each agent has to decide what information to *reveal* regarding the

sen in one part of a strategy precludes reaching some other points. In our case, however, all those strategies give the same pay-off for both players (0,0), and the solution of the game does not change if we reduce all of them to the strategy N of the fifth column in fig. 1.b.

possibilities and opportunities he knows or he has conceived, the problems associated to them, and so on.

Other assertions of actor-network theorists are shown to have a shakier basis. For example, the game-theoretic reconstruction makes us see that, when we think in terms of *preferences* instead of *goals*, the aims of the agents do not really ‘change’ through the negotiation with others (save, perhaps, when radically new opportunities are conceived or communicated). What really happens is that, being the payoffs obtained by each player dependent on the decisions made by the others, one can not simply pick up ‘directly’ her favourite option, but has to enter a process of strategic thinking and negotiation. The result of this is not ‘a new goal’ formed by the combination of the *goals* of the agents, but the outcome of a combination of *choices*, an outcome which will be placed, if the agents are rational, in a not too low rank in their orders of preferences. Regarding Latour’s insistence in that the result of these translations is a mix of ‘scientific’ and ‘non-scientific’ goals, which can not be isolated or purified, the game theoretic reconstruction divides this ‘mixing’ into two different aspects: first, the outcome of a game is determined by the preferences of *all* the players, be they scientists or not, although these preferences are not ‘mixed’; and second, talking about preferences instead of about goals makes it clearer that all agents, scientists included, will assess each possible outcome according to a combination of criteria, some of them ‘epistemic’, some of them ‘economic’, some of them ‘politic’, and so on. I shall return to this question below.

With respect to the notion of ‘actants’, a game theoretic translation also shows transparently what is sensible (even common-sensible), and what is just a *fasson de parler* in the thesis that scientists have to ‘negotiate’ with ‘non-humans’ *in the same sense* in which they have to negotiate with other people. In Joliot’s example, he attempted, amongst a myriad of other things, to measure the number of neutrons which were produced in each nuclear disintegration, and to develop a physical mechanism which can serve to control the chain-reaction. According to Latour,

Joliot’s labours could not of course be confined to ministerial offices. Having gained his laboratory, he now had to go and negotiate *with the neutrons themselves*. Was it one thing to persuade a minister to provide a stock of graphite, and quite another to persuade a neutron to slow down enough to hit a uranium atom so as to provide three more neutrons? Yes and no. For Joliot it wasn’t very different . . . Containing the minister and the neutrons in the same project, keeping them acting and keeping them under discipline, were not really distinct tasks. *He needed them both* (pp. 89–90).

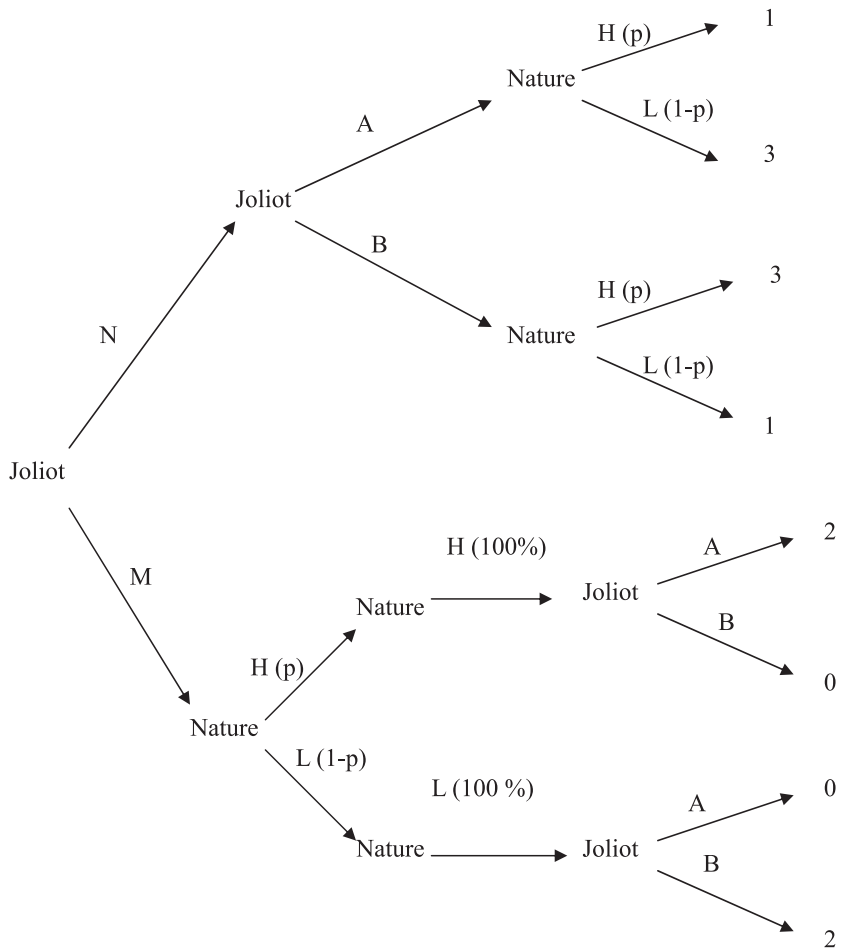


Figure 3.

Besides the truism that both factors were necessary for the project, and even besides the similarities between ‘negotiating with the minister’ and ‘negotiating with the neutrons’, it is also true that significant differences exist between both things, differences which can deeply affect our epistemological consequences. Figure 3 shows a simplified game Joliot might be playing ‘against’ the neutrons (which, in a more realistic reconstruction, will be a *part* of a bigger game where other human and non-human players also intervene). Joliot would have in the first place two op-

tions, which are to construct the chain reactor without attempting to measure first the number of neutrons produced per atomic disintegration (N), or to make this measure before constructing the reactor (M). To simplify, imagine that, before making the measure, he considered only two possibilities regarding the behaviour of neutrons: 'high number' (H) with probability p and 'low number' (L) with probability $1 - p$; each one of these possibilities would demand two different types of mechanisms for the reactor. If Joliot chooses the option N, then he can either plan to build a reactor adapted for the possibility H (let call this option A), or one adapted for L (B). If he chooses the option M (i.e., making the measure first), then 'nature' will 'select' H or L, and *then* Joliot will choose between A and B, now knowing for certain whether the reactor will 'work' (obviously, this is a simplification, for new sources of uncertainty will arise in real, more complicated cases). The final pay-offs that can be expected before taking any decisions are the following (again, numbers only indicate the order of preference): 3, if a 'good' reactor is built without having the cost of measuring the number of neutrons first; 2, if a 'good' reactor is built, but incurring the costs of measuring; 1, if a 'bad' reactor is build, but having saved the costs of measuring neutrons; and 0, if the reactor is 'bad', and the measure has also been made. What the most desirable option is at the first node will depend on Joliot's estimation of the probability p , and of the precise levels of preference attached to each possible result.¹⁰

The fundamental difference between the games in figures 1 and 3 is the fact that, whereas in fig. 1 the *preferences* of Dautry were essential to determine the equilibrium of the game, now *the player called 'nature' has no preferences at all*, nor any capacity to *decide* what to do at the nodes corresponding to it. This entails that Joliot can not expect, e.g., that the 'choices' of nature in the lower part of fig. 3 are *influenced* by nature's expectations about Joliot's own choices afterwards, as Dautry expected that Joliot's choices in the upper part of fig. 1 were influenced by Joliot's beliefs about what decisions could make Dautry in the last nodes. In other words, *non-humans can not behave strategically, nor can humans play strategically 'against' them* (except, perhaps, in the case of some living beings; but, even in this case, the equilibrium of the game, when several of them are possible, can-

10. Let $a = u(3)$ (i.e., the utility of reaching a result ordered with the number 3), $b = u(2)$, $c = u(1)$ and $d = u(0)$, such that $a > b > c > d$. In this case, at the second node of the upper line, A will be preferable to B if and only if $p > \frac{1}{2}$ (if $p = \frac{1}{2}$, then A and B will give the same expected utility); at the first node, it can be proved that, if $p > \frac{1}{2}$, then N will be better than M if and only if $(b - c)/(a - b) < p/(1 - p)$, and if $p > \frac{1}{2}$, then N will be better than M if and only if $(b - c)/(a - b) < (1 - p)/p$. If $p = \frac{1}{2}$, then N is better than M if and only if $a - b > b - c$.

not be selected by a previous process of *communication* with them, which is essentially what a negotiation amounts to). The hard job for Joliot regarding his ‘negotiation’ with neutrons is getting an *objective* estimation of the *probabilities* of each possible outcome of nature’s ‘choices’, whereas in the negotiation with the minister the main problem for Joliot was rather that of seeing the situation *from the point of view* of Dautry, as well as that of what information to provide to him during the communication process. Of course, an essential part of Joliot’s job, as well as of any other scientist, is to *make* things behave in a desired way, but, contrarily to what we can make with other people, this can not be done by ‘persuading’ those things, but only by *manipulating* them, finding out how they *regularly* react under several circumstances (by the way, conceiving what of these circumstances—or ‘trials’—can be most useful for the humans’ interests is one of the most difficult tasks for a scientist). In discovering the best possible way of playing a game like that of fig. 3, Joliot will have to strive for getting *true* knowledge about the behaviour of things, i.e., for having a representation as *accurate* as possible of the probabilities with which some empirically testable events may happen under certain circumstances, since this will be the only way in which the interests of Joliot *in that game* may be satisfied, whatever these interests are. On the contrary, in a game like that of fig. 1 (or, by the way, in a more realistic game, where both humans and non-humans take part), Joliot will not only have to figure out what decisions can Dautry or the other human players make, but will also have to persuade them of making certain choices on their turn, instead of other choices they might make.

Nothing in the stories told by Latour about how these ‘trials’ are made to succeed entails that the process departs an inch from the ideas of, say, Hempel or Popper about the empirical testing of hypotheses, save, perhaps, by Latour’s insistence in the role of experiments as *tools of persuasion*. So, even if scientists can not behave strategically with respect to non-humans, they *can* take into account, when playing ‘against nature’, strategic reasons which have to do with the possible decisions of other *humans* who are playing in the same game; for example, Joliot may have preferred to carry out a particular experiment (which is a game against nature) instead of others, because he expected that its results would more easily ‘force’ their colleagues to accept certain facts. Nevertheless, the proponents of actor-network theory have not provided a convincing explanation of *why are the colleagues persuaded by certain experiments more likely than by others* (instead of, for example, systematically rejecting to be persuaded by any argument). This is just the problem we met at the end of section 1.2, and one for which the only solution I think can work is—as I argued there—to assume that the members of a scientific community have *agreed* (tacitly at

least) to play the persuasion game according to a more or less definite set of methodological rules, an agreement which has to have the property of being self-enforceable (or self-enforceable 'enough', as in section 1.3). If this view is right, one of the main questions for future research in the sociology of scientific knowledge would be the following: *how exactly do the ('social' or 'epistemic') interests and preferences of individual scientists lead these to accept, not just the truth of certain hypotheses or theories, but, more importantly, the validity of certain methods?* Surely, both empirical and game-theoretical research will be essential for finding a convincing answer to this question.

3. Fact Making Games.

In this final section I will present a game theoretic model of the way a research community reaches a certain degree of consensus about what propositions must be accepted in its area. Stated in other way: I will try to show how the 'construction of a scientific fact' arises as the outcome of a game theoretic equilibrium. 'Consensus' is not employed here to refer to the unanimous acceptance or rejection of a thesis, but as a matter of degree, i.e., just as an indication of which members of the community accept the thesis and which ones reject it.

Since any proposition will have been proposed by some researcher, a pertinent first question is *why has she publicly proposed this proposition instead of other* (from the set of statements she is able to conceive). According to the approach developed in this paper, an obvious answer is that she will propose a certain statement expecting to get a high external score thanks to it, i.e., according to the expectation she has that her colleagues will accept it, given how strongly they are committed to the prevailing methodological norms, and *given the facts they have already accepted* (for those norms may *command* to accept a certain proposition if it is shown to be in certain relations with those facts). Hence, the wider is the consensus about the validity of a certain fact F , the stronger is the incentive of any scientist to invent a hypothesis H for which F can be taken as a reason (i.e., such that the 'inference' from F —and other accepted facts—to H is appropriate according to the community's methodological norms). This will make the scientist to accept F as well, since proposing H without accepting F will tend to reduce her internal score.

On the other hand, if she happens to invent a theory that the methodological norms command to *reject* in the face of some commonly accepted facts, that theory will not be accepted unless she manages to persuade her colleagues that these facts are not acceptable, but this will only be possible if she shows that *other* commonly accepted propositions command to reject those ones which are problematic for her theory. This strategy will have a cost for our scientist, since she will have to devote time and effort to

finding out the needed arguments and facts. In a nutshell, *the more accepted is a fact within your scientific community, the stronger your incentive will be for accepting it, other things being equal.*¹¹ After all, when scientists compete for the resolution of a problem, they must basically agree on what the problem is, and this is just impossible without agreeing at least about some facts.

3.1. Individual decisions.

Let N be a group of researchers competing for the discovery of some aspects of the world. We will take any subset A of N as the set of scientists who are accepting an empirical fact E which is relevant for their research. For the reasons just offered, the bigger is the set of scientists which accept E , the more interesting it will be for any member of N to accept it. We can formally represent this assumption as follows (' $u_i(X)$ ' represents researcher i 's expected utility if X is the set of researchers accepting E):

Suppose $A \subseteq B$. Then, (2)

- (a) if $i \in A$, then $u_i(A) \leq u_i(B)$
- (b) if $i \notin B$, then $u_i(B) \leq u_i(A)$.

The utility level $u_i(A)$ will also depend on the *epistemic value* that E has for i , that is, on the relation E has with other propositions i has accepted, as well as on other properties E has for i from the cognitive point of view. This factor will be considered again in the next section. By the moment, I will assume that the epistemic value of E is fixed (though not necessarily identical) for every researcher.

We can now define the *reaction set* of A as the set of researchers who would prefer to accept E , rather than not accepting it, if it were accepted just by the members of A . Formally:

$$r(A) = \{i \mid u_i(A - i) \leq u_i(A \cup i)\} \quad (3)$$

Note that, if $i \in A$, she will belong to $r(A)$ if and only if $u_i(A - i) \leq u_i(A)$, whereas, if $i \notin A$, she will belong to $r(A)$ if and only if $u_i(A) \leq u_i(A \cup i)$.¹²

Since the expected utility of every researcher does not only depend on her own decision, but also on the decisions taken by the rest, we can ex-

11. Brock and Durlauf (1999) and Zamora Bonilla (1999) present two models in which scientists only take into account the *number* of colleagues accepting each proposition, though not *who* these colleagues are.

12. The proofs become more complicated if it is possible that $u_i(A) = u_i(B)$ for some i , A and B . In (3) it is tacitly assumed that, if $u_i(A - i) = u_i(A \cup i)$, i will prefer to accept the fact rather than not, but other mechanisms can be imagined.

Table 1.

X	\emptyset^*	a	b	c	ab*	ac	bc	abc
Player								
a	4	1	2	2	3	3	1	5
b	4	2	1	2	3	1	3	5
c	7	6	6	1	5	2	3	4
r(X)	\emptyset	b	a	ab	ab	ab	ab	ab

pect that the actually observed situations will be those which are Nash equilibria, i. e., those where every scientist is making her best possible choice given the choices made by her colleagues. These equilibria can be described by the following simple condition:

$$\text{A Nash equilibrium is a subset } A \subseteq N, \text{ such that } A = r(A). \quad (4)$$

This simply means that A is an equilibrium if and only if, given that the members of A accept E , and the other scientists reject it, neither those accepting the fact E would prefer to reject it, nor those rejecting it would prefer to accept it. In order to examine the possible existence and the properties of these equilibria, the following two results are useful (the proofs are in the appendix):

$$\text{If } A \subseteq B, \text{ then } r(A) \subseteq r(B). \quad (5)$$

$$\text{If there is a set } A \text{ such that } A \subseteq r(A), \text{ or such that } r(A) \subseteq A, \quad (6)$$

then there is at least one Nash equilibrium.

The following theorems express the existence and fundamental properties of the equilibria:

$$\text{There is at least one Nash equilibrium.} \quad (7)$$

Proof: It derives directly from (6), since $\emptyset \subseteq r(\emptyset)$ and $r(N) \subseteq N$. \square

$$\text{There can be more than one equilibrium.} \quad (8)$$

Proof: Table 1 exemplifies this possibility, for in that case the sets marked with an asterisk (\emptyset and $\{a, b\}$) are equilibria. \square ¹³

13. The numbers in the cells are the utility levels associated by each researcher—the third first rows—to each possible consensus—columns—. Readers can check that these utility assignments satisfy (2). The last row gives the reaction set for each possible consen-

There can be Pareto inefficient equilibria. (9)

Proof: In table 1, the equilibrium $\{a, b\}$ is Pareto inefficient, for every researcher prefers \emptyset to $\{a, b\}$. \square

- (a) There can be a set A such that, for every n , $r^n(A)$ is not an equilibrium.
- (b) If this is the case, then for some n, m , $r^n(A) = r^{n+m}(A) = r^{n+2m}(A) = \dots$

Proof: For (a), in table 1 we have that $r(\{a\}) = \{b\}$, whereas $r(\{b\}) = \{a\}$. This entails that, if n is an odd number, $r^n(\{a\}) = \{b\}$, and if n is an even number, $r^n(\{a\}) = \{a\}$, and the opposite for $\{b\}$. So, starting either from $\{a\}$ or from $\{b\}$, no equilibrium is attained through the dynamics induced by function r . For (b), the proof is direct taking into account that N is finite, so, the series $r^1(A), r^2(A), \dots$, must form a cycle. \square

3.2. Collective decisions.

Our three last results have negative implications for the traditional ideas about the objectivity of science, and they can be seen as a game-theoretic translation of several constructivist thesis which are popular within science studies. Theorem (8) can be seen as a radicalisation of the underdetermination thesis, for it tells that the actual consensus about a scientific fact could have been different, not only under the same amount of empirical information possessed by every researcher, but also under the same attributions of authority among the members of a scientific discipline. Basically, the actual equilibrium consensus on E will depend on which *was* the equilibrium before the new empirical information was added (as we will see in section 3.3); so, the possibility shown in this theorem can be expressed by the lemma ‘*history matters*’. Theorem (9) indicates that a community may reach a particular consensus even if all of its members would have preferred another one. For example, in table 1, the three researchers are not very persuaded about the presumed fact, and particularly c is very unpersuaded (she would not accept E in any case), but both a and b can be accepting it just because the other has accepted it. I propose to call this possibility ‘*the naked emperor effect*’. Lastly, (10) asserts that it is even possible that no equilibrium consensus is attained, but the research community

sus, in the following way: for subset X (column) and scientist i (row), we check whether i gets a higher utility from X than from $X-i$, if $i \in X$, or whether she gets a higher utility from $X \cup i$ than from X , if $i \notin X$; if the answer is ‘yes’, then i belongs to $r(X)$; the argument is then repeated for each scientist.

is instead permanently oscillating between several non-equilibrium situations; this can be called ‘*the sidewalk crossing effect*’.

The good news for friends of rationality is that these negative conclusions can be counteracted by the addition of new empirical information, as well as by the possibility of *collective choices*, i. e., decisions made through the agreement of a set of researchers, or a ‘coalition’, not necessarily equal to the full community. I shall examine in the first place the effect of coalitions.

- (a) If there is a cycle $A_1, A_2 (= r(A_1)), \dots, A_n (= r(A_{n-1})), A_1 (= r(A_n))$, then there are sets $B \{ \subseteq \bigcap_{(1 \leq j \leq n)} A_j \}$ and $C \{ \supseteq \bigcup_{(1 \leq j \leq n)} A_j \}$, such that B and C are equilibria and can be reached by the collective choice of the members of $(\bigcup_{(1 \leq j \leq n)} A_j) - (\bigcap_{(1 \leq j \leq n)} A_j)$.
- (b) In this case, for every $i \in (\bigcup_{(1 \leq j \leq n)} A_j) - (\bigcap_{(1 \leq j \leq n)} A_j)$, it happens that $u_i(B) \leq u_i(A_j)$ (for all A_j such that $i \notin A_j$, and for all or A_j such that $i \notin r(A_j)$), and $u_i(C) \leq u_i(A_j)$ (for all A_j such that $i \in A_j$, and for all A_j such that $i \in r(A_j)$).

This theorem shows that, in principle, cycles can be avoided by means of the making of (not necessarily unanimous) collective decisions: the ‘doubting’ researchers, i.e., those who accept the fact at some point in the cycle but not at all points, may form a coalition and decide simultaneously either to accept or to reject it. The first decision will lead the community to a new equilibrium (C), at which all the scientists who accepted the fact at some point of the cycle will be better than in those cases where they were accepting it or were willing to accept it. The second decision will lead to an equilibrium (B) which is included into the set of scientists who accepted the fact at all the points of the cycle, and which all researchers taking part in the cycle prefer it to the cases when they did not accept E , or were not willing to accept it. Unfortunately, from (11) it does not automatically follow that scientists will *actually* be interested in forming such a coalition, for some of them may prefer equilibrium B , and others equilibrium C , and perhaps no agreement is possible between both groups. Nevertheless, the possibility that coordination serves to avoid ‘sidewalk crossing’ suggests that this phenomenon will not be very common in science. The next theorem refers to the way that coalition formation can avoid the presence of the ‘naked emperor effect’.

If there is an inefficient equilibrium A , then there will also be an efficient equilibrium B which can be reached from A by means of a collective choice. (12)

This result can serve as a counterargument to the idea that the ‘negotiated’ character of scientific facts entails that they are not ‘objective’, at

least in the sense of their acceptance not being consistent with sound methodological criteria. Rather on the contrary, *it is precisely the ability to negotiate a collective decision what allows the scientific community to avoid inefficient equilibria*. We must take into account that your getting a low level of utility may be due to a small chance of getting your theories accepted (a low expected value for your external score), and also to a bigger difficulty in fulfilling the chosen methodological criteria while defending a proposition (a low expected value for your internal score). Efforts for attaining a high external score constitute a very competitive ('zero-sum') game, since in the race for discovery 'the winner takes it all'. Hence it is very unlikely that, if *all or almost all* scientists' utilities may increase by passing from an equilibrium to another, this could be explained by an increment in *every* scientist's chance of getting a higher *external* score. It is much more likely that it is due to a general improvement in the possibility of successfully fulfilling the methodological norms of the community, and hence, to an increased chance of getting a higher *internal* score. For example, it can be due to the rejection of some false, but previously accepted results or principles, in such a way that the inconsistencies they led to are also removed.

3.3. Responding to changes in epistemic valuations.

As I mentioned above, utility levels regarding the acceptance of a statement *E* also depend on the *epistemic value* the proposition has for each individual researcher, which can obviously be different from the value it has for the others. I suggest to decompose this epistemic value into three different factors:

a) the strength with which the prevailing methodological norms command to accept *E*, given what other propositions a scientist has accepted (hence, not accepting *E* will make decrease your internal score); this is the *epistemic support* of *E*;

b) the possibility of using *E* to justify the statements or models proposed by you (hence, accepting *E* may help to increase your external score); this is the value of *E* as an *epistemic tool*; and lastly

c) the psychological conviction you have about the validity of *E*, given the reasons you happen to have (theoretical arguments, experimental evidence, and so on); this is the *cognitive value* of *E*.

By calling 'epistemic value' to the *private* information each scientist has about every proposition, I am far from denying that knowledge assessment is a collective process. Rather on the contrary: what I am trying to show is *how* this collective assessment is made. The essential point of a game theoretic explanation of such a decision is that every agent has to take *simultaneously* into account any 'private' information she may have, as well as the claims her colleagues are publicly making.

An intuitive assumption is that, the higher is the epistemic value of E , the lower it will be the expected utility associated to *rejecting* it (either for the increased probability of not fulfilling the methodological norms by doing so, or for the costs of looking for new arguments in favour of E , or for the psychic costs of denying your own beliefs), and the higher the expected utility associated to *accepting* it. We can formally express this as follows (u and u' are utility levels at two different moments, t and t'):

If, for researcher i , the epistemic value of E is higher at t' than at t , then: (13)

(a) If $i \in A$, $u_i(A) \leq u'_i(A)$.

(b) If $i \notin A$, $u'_i(A) \leq u_i(A)$.

This assumption allows to deduce our last theorem:

If E has a higher epistemic value at t' than at t for every member of N , and if A was an equilibrium at t , then there will be an equilibrium B at t' such that $A \subseteq B$. (14)

The philosophical relevance of this theorem is clear, for it means that, if every member of a scientific community agrees that the epistemic value of a scientific fact has grown (for example, because of the realisation of new empirical tests whose results happen to favour E), then each possible equilibrium consensus will be transformed into one which is at least as comprehensive as the former. Of course, the opposite takes place when the epistemic value of E decreases for everybody. We can assert, then, that the growth of the epistemic value of a fact (assessed individually by each member of a scientific community) can be an explanation of the general acceptance of the fact among a scientific community. Obviously, there can be other possible explanations as well, but *the burden of the proof is for those who argue that scientific consensus is normally reached by other types of reasons.*

In general, it seems also reasonable to expect that, for every scientific proposition, there will be a certain amount of empirical evidence such that the only possible equilibrium for E is unanimity or near unanimity. Table 2 illustrates this possibility. Table 2.a precedes from table 1 in the following way: an improvement of the epistemic value of E for every researcher has made utility levels increase by one unit where $i \in X$, and decrease in the same amount where $i \notin X$, coherently with (13); a similar change takes place in passing from table 2.a to 2.b. In table 2.a, the equilibrium $\{a, b\}$ has disappeared, substituted by N ; \emptyset is still an equilibrium, but it can be contested by a collective choice of a and b , who, if they decide to form a coalition to accept E , can lead the community to N , where both have a higher utility than at \emptyset (c will not benefit from this jump, for his

Table 2a.

Player \ X	∅*	a	b	c	ab	ac	bc	abc*
a	3	2	1	1	4	4	0	6
b	3	1	2	1	4	0	4	6
c	6	5	5	2	4	3	4	5
r(X)	∅	b	a	ab	abc	ab	ab	abc

Table 2b.

Player \ X	∅	a	b	c	ab	ac	bc	abc*
a	2	3	0	0	5	5	-1	7
b	2	0	3	0	5	-1	5	7
c	5	4	4	3	3	4	5	6
r(X)	ab	abc	abc	ab	abc	abc	abc	abc

utility is higher at ∅; but he can not do anything to prevent it). In table 2.b, the epistemic value of *E* has increased so much that the only possible equilibrium is *N*, which has become, besides, the situation preferred by all researchers.

References

Altman, E., and P. Hernon (eds.). 1997. *Research Misconduct: Issues, Implications, and Strategies*. London: Ablex Publishing.

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Binmore, K. 1992. *Fun and Games: A Text on Game Theory*. Lexington (Mass.): D. C. Heath.

Bloor, D. 1976. *Knowledge and Social Imaginery*. London: Routledge and Kegan Paul.

Brandom, R. B. 1994. *Making It Explicit. Reasoning, Representing, and Discursive Commitment*. Cambridge (Ma.): Harvard University Press.

Brock, W. A., and S. N. Durlauf, 1999. "A Formal Model of Theory Choice in Science". *Economic Theory* 14: 113–30.

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

- Elster, J. 1989. *The Cement of Society: A Study of Social Order*. Cambridge: Cambridge University Press.
- Hands, D. W. 2001. *Reflection without Rules: Economic Methodology and Contemporary Science Theory*. Cambridge: Cambridge University Press.
- Hargreaves Heap, Sh. P., and Y. Varoufakis. 1995. *Game Theory. A Critical Introduction*. London: Routledge.
- Goldman, A. I., and M. Shaked. 1991. "An Economic Model of Scientific Activity and Truth Acquisition". *Philosophical Studies* 63: 31–55.
- Hull, D. 1988. *Science as a Process*. Chicago: The University of Chicago Press.
- Kitcher, P. 1993. *The Advancement of Science*. Oxford: Oxford University Press.
- Kreps, D. M. 1990. *Game Theory and Economic Modelling*. Oxford: Clarendon Press.
- Latour, B. 1999. *Pandora's Hope*. Cambridge, Ma.: Harvard University Press.
- Luetge, Ch. 2004. "Economics in Philosophy of Science: A Dismal Contribution?". *Synthese* 140: 279–305.
- Mäki, U. 2004. "Economic Epistemology: Hopes and Horrors". *Episteme* 3: 211–222.
- Mirowski, P., and E.-M. Sent (eds.). 2002. *Science Bought and Sold: Essays in the Economics of Science*. Chicago: The University of Chicago Press.
- Sent, E.-M. 1999. "The Economics of Science: Survey and Suggestions". *Journal of Economic Methodology* 6: 95–124.
- 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.
- Zamora Bonilla, J. P. 1999. "The Elementary Economics of Scientific Consensus". *Theoria* 14: 461–88.
- . 2002. "Scientific Inference and the Pursuit of Fame: a Contractarian Approach". *Philosophy of Science* 69: 300–23.

Appendix

Proof of (1) (section 1.3).

Under the assumptions made in the text, expected utility is given by

$$(15) EU_i(f, \mathbf{f}) = (1 - f)(a + bf)(1 - \mathbf{f}) - f(cf) \quad (15) \\ = -f^2(b(1 - \mathbf{f}) + c) + f(b - a)(1 - \mathbf{f}) + a(1 - \mathbf{f})$$

Individual maximisation of (1) is reached when $\leq EU/\leq f = 0$, and, expressing the optimum f as a function of \mathbf{f} , this happens when:

$$f_i^*(\mathbf{f}) = \{(b - a)(1 - \mathbf{f})\}/2(b(1 - \mathbf{f}) + c) \quad (16)$$

From this it follows that:

$$(i) f_i^*(\mathbf{f}) \leq (b - a)/(2b + c) < \frac{1}{2} \text{ (if } a < b) \quad (17)$$

$$(ii) f_i^*(\mathbf{f}) = 0 \text{ (if } a \leq b)$$

$$(iii) f_i^*(1) = 0$$

$$(iv) df_i^*/d\mathbf{f} = - (b - a) c / \{(2(b(1 - \mathbf{f}) + c))^2\} (< 0, \text{ if } a < b)$$

$$(v) |df_i^*/d\mathbf{f}| < 1 \text{ (if } \mathbf{f} \leq \frac{1}{2})$$

(17.i) is immediate. (17.ii) follows from EU being decreasing within the interval $\{0, 1\}$. (17.iii) asserts that, when your colleagues always disobey the norms, your optimum strategy is to obey them; though this may sound strange, take into account that in this case $EU(f, 1)$ equals $-cf^2$, which is negative for $f > 0$, and 0 for $f = 0$; i.e., you do not get anything either from obeying or from disobeying the norms, but are still punished when you are discovered disobeying them. On the other hand, (17.iv) entails that your optimum frequency of infringement *decreases* as the average frequency rises. This result, as well as the previous one, essentially derives from the assumption that you are not *less* punished for your transgressions when your colleagues commit *more* infringements in the aggregate. Regarding other types of social norms, this need not be true; for example, when the police works *less* efficiently, it is *more* probable that you will not be punished because of your crimes (although you can be 'punished' even if you do not commit any crime), and this provides a reason to commit more crimes. However, in the case of science, researchers want essentially to have a global score *higher* than their colleagues', and this entails that they will hardly miss the opportunity of denouncing your infringements, even when they also commit many of them. So, our assumption that \mathbf{f} affects u but not v simply means that, the higher is \mathbf{f} , the less willing will your colleagues be to recognise your merits, though they will always be equally prone to punish you. [Another plausible assumption is that you are indeed *more* 'punished' as \mathbf{f} increases, *independently of how frequently you misbehave*; i.e., the more your colleagues disobey the laws, the more frequently will you get an *unjustified* penalty. This can be represented making v to depend *positively* on \mathbf{f} , for example, by making $v = cf + d\mathbf{f}$ ($d \leq 0$). It

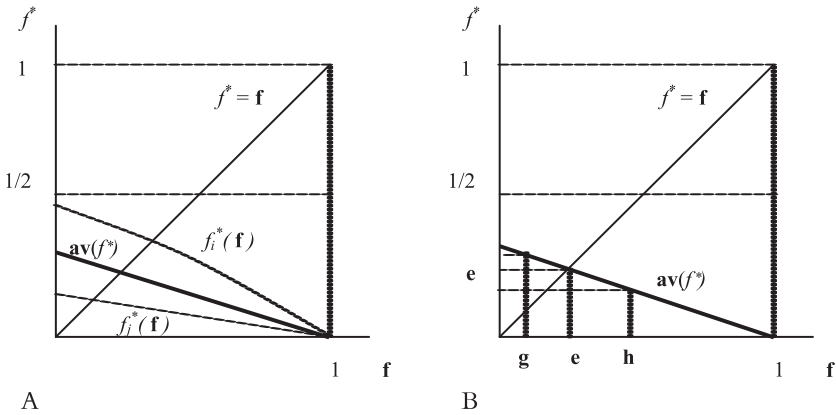


Figure 4.

can be shown that (17.i-iv) are still valid with this modification of the *EU* function, though the proof is slightly more complicated. In particular, if $d \gg 0$, the optimal f for every researcher will be still lower than when $d = 0$. Finally, (17.v) is easy to compute.

Lastly, the proof of (1) proceeds as follows: In figure 4.a, $f_i^*(f)$ and $f_j^*(f)$ represent, for each possible value of f , the optimum choices of f that a couple of members of a scientific community would make. The function $av(f^*)$ gives, for each value of f , the average of the optimum choices that would be made by *all* the community members. Figure 4.b shows e as the equilibrium point, where the $av(f^*)$ function crosses the identity line. If the average choice were smaller (like in g), the optimum choices of the scientists would lead to an average choice bigger than e , and, *a fortiori*, bigger than g ; if the average choice were bigger than e (like in h), then the average of the optimum choices would be smaller than e , and hence, smaller than h . So, that the average of the optimum choices is e is a necessary condition for an equilibrium to obtain. On the other hand, if the choices of the scientists had an average equal to e , but with a distribution different from the one deriving from the functions $f_i^*(f)$, this would be inconsistent with the assumption that scientists are making their optimum choice; hence, of all the distribution of choices whose average is e , the only one which corresponds to a Nash equilibrium is the one where each scientist i chooses the frequency $f_i^*(e)$. Lastly, that $e < 1/2$ derives from (17.i), and stability is guaranteed by the fact that the slope of the f 's is less than one (and hence, displacements from the equilibrium will tend to approach the decisions to that point again).

Proof of (5) (section 2.1)

Suppose that $A \subseteq B$, and $i \in r(A)$. If $i \in A$, then by assumption $i \in B$, and $u_i(A) \leq u_i(B)$ [by 2.a]; since $i \in r(A)$, then $u_i(A - i) \leq u_i(A)$ [by 3], and since $A - i \subseteq B - i$, then $u_i(B - i) \leq u_i(A - i)$ [by 2.b]; hence, $u_i(B - i) \leq u_i(B)$, i. e., $i \in r(B)$ [by 3]. If $i \notin A$, $u_i(A) \leq u_i(A \cup i)$ [by 3, and the assumption that $i \in r(A)$]; now it is possible that i belongs to B or not; if $i \in B$, then both $u_i(A \cup i) \leq u_i(B)$ and $u_i(B - i) \leq u_i(A)$ [by 2, and the assumption that $A \subseteq B$]; so, $u_i(B - i) \leq u_i(B)$, and hence $i \in r(B)$ [by 3]; lastly, if $i \notin B$, then $u_i(B) \leq u_i(A)$ [by 2.b], but, since $i \in r(A)$, $u_i(A) \leq u_i(A \cup i)$ [by 3], and since $A \subseteq B$, $u_i(A \cup i) \leq u_i(B \cup i)$ [by 2.a]; hence $u_i(B) \leq u_i(B \cup i)$, which means that $i \in r(B)$. So, if $A \subseteq B$ and $i \in r(A)$, then $i \in r(B)$.

Proof of (6) (section 2.1).

Let $r^1(A)$ be $r(A)$, and define inductively $r^{n+1}(A)$ as $r(r^n(A))$. Assume that $A \subseteq r(A)$; then by (5), $r^n(A) \subseteq r^{n+1}(A)$ for every $n \leq 1$; but, since N is finite, there will be some m for which $r^m(A) = r^{m+1}(A)$, and hence $r^m(A)$ is an equilibrium. An analogous proof is valid if $r(A) \subseteq A$.

Proof of 11 (section 3.2.).

With respect to (a), if, while the community is oscillating within the cycle, the members of the set $(\cup_{(1 \leq j \leq n) A_j}) - (\cap_{(1 \leq j \leq n) A_j})$ collectively decide to *reject* statement E , the community will pass to $r(\cap_{(1 \leq j \leq n) A_j})$; but, since $\cap_{(1 \leq j \leq n) A_j} \subseteq A_j$ for every j , it follows that, for every j , $r(\cap_{(1 \leq j \leq n) A_j}) \subseteq r(A_j)$ ($= A_{j+1}$; recall that $r(A_n) = A_1$) [by 5], and this entails that $r(\cap_{(1 \leq j \leq n) A_j}) \subseteq \cap_{(1 \leq j \leq n) A_j}$; from this it follows that there is an equilibrium which is a subset of $\cap_{(1 \leq j \leq n) A_j}$ [by 6]. On the other hand, if the members of $(\cup_{(1 \leq j \leq n) A_j}) - (\cap_{(1 \leq j \leq n) A_j})$ collectively accept E , the community will pass to $r(\cup_{(1 \leq j \leq n) A_j})$; since $A_j \subseteq \cup_{(1 \leq j \leq n) A_j}$ for all j , we have that $r(A_j)$ ($= A_{j+1}$) $\subseteq r(\cup_{(1 \leq j \leq n) A_j})$ for all j [by 5], and then, $\cup_{(1 \leq j \leq n) A_j} \subseteq r(\cup_{(1 \leq j \leq n) A_j})$; this entails that there is an equilibrium which is a superset of $\cup_{(1 \leq j \leq n) A_j}$ [by 6].

With respect to (b), for any $i \in (\cup_{(1 \leq j \leq n) A_j}) - (\cap_{(1 \leq j \leq n) A_j})$, (2.b) and the construction of B entail that $u_i(B) \leq u_i(\cap_{(1 \leq j \leq n) A_j}) \leq u_i(A_j)$ (for those A_j containing i), whereas (3) and (2) entail that $u_i(B) \leq u_i(A_j - i) \leq u_i(A_j)$ (for those A_j such that $i \in A_j$ and $i \notin r(A_j)$) and that $u_i(B) \leq u_i(A_j) \leq u_i(A_j \cup i)$ (for those A_j such that $i \notin A_j$ and $i \notin r(A_j)$). On the other hand, (2.a) and the construction of C entail that $u_i(C) \leq u_i(\cup_{(1 \leq j \leq n) A_j}) \leq u_i(A_j)$, for those A_j not containing i , and (3) and (2) entail that $u_i(C) \leq u_i(A_j) \leq u_i(A_j - i)$ (for those A_j such that $i \in A_j$ and $i \in r(A_j)$) and that $u_i(C) \leq u_i(A_j \cup i) \leq u_i(A_j)$ (for those A_j such that $i \notin A_j$ and $i \in r(A_j)$).

Proof of 12 (section 3.2).

If A is inefficient, this means that there is a set C such that every $i \in N$ prefers C to A . In the first place, if $C = \emptyset$, then the members of A can collectively reject the statement, and the equilibrium \emptyset will automatically be reached (that \emptyset is an equilibrium is shown by the fact that, for every $i \in A$, $u_i(\emptyset) \leq u_i(A)$ [by assumption] $\leq u_i(\{i\})$ [by 2.a], and hence $i \notin r(\emptyset)$ [by 3]; for every $i \notin A$, $u_i(\emptyset) \leq u_i(A)$ [by assumption] $\leq u_i(A \cup i)$ [by 3, because by assumption A is an equilibrium] $\leq u_i(i)$ [by 2.a], and hence $i \notin r(\emptyset)$ [by 3]; so, $r(\emptyset) = \emptyset$). In this case, $B = C$.

In the second place, an analogous argument shows that, if $C = N$, N is an equilibrium and can be reached by the collective decision of accepting E made by the members of $N - A$ (and $B = N$).

In the third place, if $N \neq C \neq \emptyset$, then it can be proved that we have neither $A \subseteq C$, nor $C \subseteq A$, because in any of these cases, either the members of $N - A$, or the members of C , respectively, will be worse off at C than at A [by 2], and so C can not be Pareto superior to A . We can show then that $A \cup C \subseteq r(A \cup C)$, for, if $i \in C - A$, we have $u_i((A \cup C) - i) \leq u_i(A)$ [by 2.b] $\leq u_i(C)$ [by assumption] $\leq u_i(A \cup C)$ [by 2.a], hence $u_i((A \cup C) - i) \leq u_i(A \cup C)$, and $i \in r(A \cup C)$ [by 3]; if $i \in A$, we have $u_i((A \cup C) - i) \leq u_i(A - i)$ [by 2.b] $\leq u_i(A)$ [by 3, because A is an equilibrium, and hence $i \in r(A)$] $\leq u_i(A \cup C)$ [by 2.a]; hence $u_i((A \cup C) - i) \leq u_i(A \cup C)$, and $i \in r(A \cup C)$ [by 3]. So, if the members of $C - A$ decide collectively to accept E , the situation will pass from $A \cup C$ to $r(A \cup C)$, and an equilibrium $B (\supseteq A \cup C)$ will be reached [by 6].

Lastly, we have to show that an *efficient* equilibrium will be reached. If \emptyset is Pareto superior to A and N is not, then \emptyset is efficient, because, for any D not identical with \emptyset or with N , the members of $N - D$ will be worse off in D than in \emptyset [by 2.b]. An analogous argument shows that N is efficient if it (but not \emptyset) is superior to A . If both \emptyset and N are Pareto superior to A , either one of them is superior to the other, in which case this one can be collectively chosen, or neither of them is, in which case the two are efficient. On the other hand, if neither $A \subseteq C$, nor $C \subseteq A$, the equilibrium B which can be collectively chosen from A might be inefficient, but if this were so, a new collective decision could jump to another equilibrium superior to B ; nevertheless, since N is finite, the process can not be repeated indefinitely, and so, it must stop at an efficient equilibrium.

Proof of (14) (section 3.3).

Let A be an equilibrium at t . If $i \in A (= r(A))$, then $u_i(A - i) \leq u_i(A)$ [by 3], and, since $u_i(A) \leq u'_i(A)$ and $u'_i(A - i) \leq u_i(A - i)$ [by 10], it follows that $u'_i(A - i) \leq u'_i(A)$; so, $i \in r'(A)$ (the reaction set of A at t'). Hence, $r(A) = A \subseteq r'(A)$, and there will be an equilibrium at t' in which A is included [by 6].

Copyright of Perspectives on Science is the property of MIT Press and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.