ISSN 0048-3931

PHILOSOPHY OF THE SOCIAL SCIENCES

PHILOSOPHIE DES SCIENCES SOCIALES

Published by / Publié par
Wilfrid Laurier University Press

PHILOSOPHY OF THE SOCIAL SCIENCES is an international quarterly which publishes articles, discussions, symposia, literature surveys, translations, and review symposia of interest both to philosophers concerned with the social sciences and to social scientists concerned with the philosophical foundations of their subject. PHILOSOPHY OF THE SOCIAL SCIENCES intends its focus of interest to be those issues normally considered central to the philosophy of the social sciences, that is to say, issues in general methodology (understood as metascience rather than research techniques), and the application of philosophy to the social sciences—but not excluding papers with substantial empirical content that constitute contributions to the field. No school, party, or style of philosophy of the social sciences is favoured. Debate between schools is encouraged.

PHILOSOPHIE DES SCIENCES SOCIALES est une revue internationale qui paraît quatre fois l'an et publie des articles, des discussions, des actes de colloques, des analyses de la bibliographie d'un sujet, des traductions, des recueils de comptes rendus qui présentent de l'intérêt aussi bien pour les philosophes des sciences sociales que pour les chercheurs en ces domaines qui se préoccupent des fondements de leurs disciplines. La revue entend se consacrer aux problèmes fondamentaux de la philosophie des sciences sociales, à savoir à des problèmes de méthodologie générale (au niveau de la métascience plutôt qu'à celui des techniques de recherche) et à l'application de la philosophie aux sciences sociales, sans exclure pour autant des textes comportant un contenu empirique substantiel s'ils constituent des apports à la philosophie des sciences sociales. La revue est non-partisane et cherche à stimuler la discussion philosophique.

(Shorter version of Ch. 3)

Phil. Soc. Sci. **18** (1988) 477-95

Methodological Rules as Conventions*†

CRISTINA BICCHIERI, *Philosophy Department, University of Notre Dame, Notre Dame, Ind. 46556*

1. INTRODUCTION

One of the chief merits of that broad ensemble of views called 'relativism' has been that of making us aware of the fact that science is a corporate, collective activity; an activity, furthermore, whose rules and outcomes vary across time and across cultures. This view is generally associated with the claim that our epistemic beliefs, relevant evidence, and acceptability criteria are culturally bound. In particular, there exist different and incompatible methodologies: any rational standard can carry weight only within its original cognitive-cultural context, and no standard can transcend its particular milieu.

Though relativism correctly points out the culture-dependence of our systems of beliefs and criteria for assessing them, its limitation is that it can only point to their multiplicity and change, without being able to offer an adequate rationale for their emergence. Relativists, in other words, are more preoccupied with disagreement and change than with convergence and consensus formation. The famous Kuhnian metaphors of 'conversion' and 'gestalt-switch' barely conceal the fact that we have no account of how scientists converge to a new paradigm or, for that matter, to a set of methodological rules.

In this paper, I shall restrict my attention to methodological rules. If one views these rules as informal criteria that guide scientists' practices, such as for example the criteria which define acceptable ways of conducting experiments, testing, model building and using mathematical techniques, then scientific methods come to be identified with these practices and procedures. In this case, little distinction is made between the procedures which govern normal scientific activities and the criteria according to which we justify the final products of such activities. There is, of course, some virtue in this blurring of the boundaries. One virtue lies in the rejection of the traditional separation between discovery and

* *Received* **12.11.87**

† Many of the ideas presented here were developed while I was a fellow at the Center for Philosophy of Science at the University of Pittsburgh. David Gauthier, Ernan McMullin, Vaughn McKim, Philip Quinn and the participants in the sociology of science seminar at the University of Notre Dame offered many helpful comments on earlier drafts, as did the anonymous referees of this journal.

justification, in that the same constraints function both in validation and in heuristics. To satisfy heuristic constraints in the discovery process simultaneously leads to legitimizing the result, in the sense of guaranteeing its provisional acceptance as an hypothesis worth considering. Another virtue is that scientific methods come to be considered in the context of communities of interacting individuals, as embedded in their activities, and thus subject to change.

Our problem can thus be restated in the following terms: if there exist many scientific communities and many different methodologies, how does it happen that each community shows a remarkable (if local) homogeneity with regard to method? How did convergence take place? What accounts for consensus formation, given that different groups agree on different things, but all agree on something?

These questions require a few qualifications. Many philosophers use the term 'methodological rules' to refer to very general epistemic values such as, for example, predictive accuracy, coherence, simplicity, fertility, unifying power and explanatory power. Consensus over such values is not local, and can hardly be disputed. In the long run, these values have shown a remarkable stability, in that at least some of them have been deemed to be characteristic of good scientific theories by a wide variety of scientific communities.

However, it is also true that those epistemic values are so general as to be open to many interpretations, and different scientists may rank them differently, so that consensus might still be difficult to achieve. A well known example is the disagreement between Bohr and Einstein as to the acceptability of the quantum theory of matter, which was at least partly grounded on methodological considerations about the relative importance of different epistemic criteria. Einstein thought the new theory was not consistent with classical physics; while Bohr did not discount this defect, he counted highly the theory's predictive success. Such disagreements are fairly common in every scientific field, since a common understanding of what predictability or coherence means in a given context does not imply unanimity in ranking them in order of importance.

What I call methodological rules are rather *instantiations* of the above epistemic goals. Precisely because those goals are open to many interpretations, there might be many sets of rules satisfying them. For example, some scientists would consider an hypothesis tested only if new predictions can be drawn from it, and these predictions are confirmed. Others would be satisfied with the hypothesis' ability to explain *post hoc* what is already known. All parties, however, would subscribe to the importance of a testability criterion. When I refer to methodological consensus, then, it must be intended as the consensus of a community, or part of it, about a set of practices which embody and exemplify certain methodological rules.

Different methodological approaches do often reflect differences regarding 'fundamental' theories. But they need not do so. It may be the case that disagreement over the sources of guidance takes place within a community sharing a common set of theoretical commitments. What, then, accounts for the difference of opinion? If epistemic (internal) reasons alone cannot account for the local consensus reached (and maintained) by each group, can we find other plausible reasons? A related question I shall address is whether the reasons which may account for such local convergence can ground the rationality claims of methodologies, or if instead these are two separate issues.

A strategy often adopted to circumvent these problems consists in emphasizing the social setting of science. Scientific knowledge is seen as a social construct that reflects the structural characteristics of the society in which it is produced; hence the current standards of acceptability are made to depend on the social context within which a scientific community operates (Barnes 1977, Bloor 1982). The independent variables thus become the goals and interests of the larger community to which scientific groups belong as sub-communities.

Yet the most this approach can do is to account for the relevance of a method with respect to a local culture's values. Moreover, if it is true that a multiplicity of methodological rules may fulfill equally well the same values, choice among them remains underdetermined. If we face what might be termed an *egalitarianism of rules*, the rationality criterion proposed by relativism (i.e., that a method is rational if it fulfills some local values or interests) is necessary but not sufficient to guarantee convergence to a method.

If rational preference among rival methodologies cannot be grounded on epistemic or sociological reasons alone, this does not mean that the final choice of one of them has to be irrational. Indeed, I shall maintain that methodological choice has a social dimension, that scientific rules and practices are *also* (but not only) a social product, and that this view is not incompatible with their being rationally chosen and adhered to.

To this effect, I shall present a methodological individualist account of convergence to a method, where convergence is modelled as the successful solution of a coordination problem among scientists. The existence of a multiplicity of possible methods, none of which is clearly 'better' than the others, renders convergence somewhat arbitrary. In this sense, the methodological agreement of a scientific community is conventional, and its outcome a convention. Whether a conventional, arbitrary choice can also be rational depends on the theory of conventions and conventional behaviour one adopts.

In what follows, I shall adopt David Lewis' account of conventions, with some crucial modifications. First, I distinguish between the arbitrariness feature and the regularity feature of conventions. The emergence of a convention and its stability are two different phenomena. Second, I distinguish between a descriptive account and a rational reconstruction of conventions, and provide conditions under which a rational reconstruction is acceptable.

Finally, I show how an account of methodological rules in terms of conventions helps to explain 'local' homogeneity, or convergence, as well as to distinguish in an important way between the reason why a particular method is first adopted by a group and why it becomes stable. This view has the further advantage of providing a more satisfactory account of what it means for a scientific methodology to be 'locally' rational.

2. METHODOLOGICAL RULES AS CONSTRAINTS

A belief shared by all relativists is that rational acceptability standards are conventional.[1] Conventional does not mean totally arbitrary, though. It is argued that methodological differences ultimately depend upon differences of local culture, which constrain and structure our scientific interests and commitments. If different initial goals and interests lead to different end-points, the problem of explaining local convergence to a method is compounded by the need for a description of the nature of these different initial conditions, and how they impinge on scientific methodologies.

Sociological reasons are often invoked to make sense of the diversity of initial conditions which accounts for the difference in assumptions and methodological rules.[2] What is not made explicit is whether cultural and social conditions *uniquely* lead to a method or whether the same initial conditions might be compatible with different alternative methodologies. This distinction is crucial, since it underlies two quite distinct views of conventions and two correspondingly different views of scientific rationality.

The sociological brand of relativism seems to support methodological uniqueness rather than multiplicity. Scientific methodologies are seen as uniquely determined by a given culture and social environment. A good example of this view is found in a case study of the controversy concerning the character of heredity and evolution between 'biometricians' and 'Mendelians' which occurred in England in the early twentieth century (MacKenzie and Barnes 1975). Biometricians held evolution to be a predictable continuous process, occurring by the selection of small 'continuous' differences, while Mendelians treated evolution as a discontinuous phenomenon, thus allowing for the unpredictable emergence of new individual traits or properties.

1 See, for example, D. Bloor, *Knowledge and Social Imagery*. London 1976, p. 37.

2 A good survey of current approaches in the sociology of knowledge is M. Mulkay (1979).

The two opposing theoretical and methodological views could not be imputed to differences in training and competence, nor to the parties' possessing differential evidence or not sharing the same theoretical framework. Nor could a positivistic ideal explain—say—the biometrician's use of statistical models, or the operational reinterpretation of concepts such as heredity or Galton's 'Ancestral law'. The respective epistemic goals, MacKenzie and Barnes conclude, were more a rationalization than a determinant of their choices. The two alternatives, it is suggested, are to be conceived as forms of culture, 'carried on by communities with particular goal-orientations and interests. Such goal-orientations and interests serve to constrain and structure technical commitments and procedural choices within science' (ibid., p. 201).

The interests and goals in this case have a sociopolitical nature. The biometrical view would provide scientific credibility to eugenics, a programme for gradually improving the human race by modifying the relative fertility of different social groups. In turn, eugenics is interpreted by the authors as a typical manifestation of the social interventionism characterizing the English professional middle class at that time. The Mendelian doctrine, on the contrary, is seen as functional to the preservation of conservative interests, since its stress on discontinuity and unpredictability could lend some legitimation to those groups who opposed all forms of social engineering.

Still, all that this and similar case studies succeed in establishing is that a doctrine, theory, or method can be and often is *used* to legitimize extra-scientific views, which in turn may represent wider social interests. They do not, however, succeed in showing that those interests *explain* scientific choices, nor that there are no alternatives that could equally well serve the same interests.[3]

This is a common problem with functional explanations, of which the above case study is an example. Very generally, a functional account of an institution, behaviour, or social pattern B explains it by means of its beneficial function F for group G. I reproduce here a number of conditions that occur fully or in part in functional explanations as Jon Elster listed them (Elster 1979, pp. 28-29):

(1) F is an effect of B.
(2) F is beneficial for G.
(3) F is unintended by the actors producing B.
(4) F (or the causal link between B and F) is unrecognized by the actors in G.
(5) F maintains B by a causal feedback loop passing through G.

3 That social and political interests are intended to be explanatory variables is made quite clear in the following passage: '. . . the narrowing and structuring of predictive concerns, and thus of specific judgement, arose from the operation of broadly based social factors' (ibid., p. 205).

What sociological accounts of scientific episodes consistently maintain are points (1)-(3). A scientific commitment or choice, for example, is justified by exhibiting its beneficial consequences for some social group, which may or may not be identical with the group making the choice. While the social function of a scientific view or method may be explicitly recognized by the proponents, this function is only acknowledged *ex post*; that is, it is never the scientists' primary reason for upholding that view or method.

Otherwise, it would have to be claimed that those scientists' choices were *intended* to produce this or that particular social result, and were thus intentionally biased. But this is not a claim the sociologist of science is willing to make. Finally, point (5) deserves special attention, since it is paradoxical to explain a scientific commitment only *ex post*, after its unintended consequences have proven to be socially desirable. What can be admitted, instead, is that a reason (or maybe *the* reason) why a view is maintained, or a pattern established, is that it is functional to some group's interest.

Even if the sociologist of science were to adopt (1)-(5), however, convergence to a method would remain unexplained. What could be explained, on the contrary, would be the permanence of a method, provided it could be conclusively demonstrated that (*a*) there is in fact a causal link between the method and its alleged social consequences, and (*b*) there exist no other methods that could have equally well served the same function. However, it is often hard to find significant correlations between scientific views and conditions of local culture, or the particular social interests that these views might serve. It is even harder to offer this type of 'externalist' explanation in all those cases in which multiple methodologies coexist within a field, the participants do not particularly disagree on theoretical issues, and logic, experiment, or epistemic criteria alone cannot settle the differences.

The difficulty in accounting for convergence, or consensus formation, is worth considering in detail. A functionalist explanation encounters a paradox if it holds that local scientific consensus is explainable in terms of its serving a positive social function. Since this function is realized only *ex post*, as an *effect* of the adoption of a method or theory, it cannot possibly be a reason for the scientists' implicit agreement. To account for consensus formation in terms of the social function the object of consensus would serve could only mean reintroducing intentionality into the picture. Conditions (3) and (4) would be violated, but then one would have renounced functional explanation.

This difficulty explains why in the sociological literature the scientist is often depicted as having no choice but to use the established method and assumptions of his own group (Barnes and Bloor 1982). Convergence, or consensus formation, is explained away by being taken as primitive. Conformism is taken as primitive too, as it is stated that we just 'prefer' our own methods and knowledge to all others, without further reasons. These methods, in turn, are a function of the local culture in which we happen to live. Thus when cultural influences are invoked, the actors tend to appear as 'cultural automata', their behaviour complying with and resulting from pre-established schemas provided by the available culture.[4]

In an account of this sort, then, scientific methodologies are relative only insofar as cultures are; within a local culture, there is no scope for arbitrariness. For this reason, individuals are presented as having no choice but to comply with whatever rule happens to be in existence. The rules they follow are not chosen; they are instead *external constraints* on their choices.

3. METHODOLOGICAL RULES AS CONVENTIONS

The identification of criteria for acceptability with the principles, precepts and techniques which characterize an entire discipline or a subgroup of it is not, *per se*, the cause of the above difficulties. There is little to object to the claim that what it means for a piece of work to be coherent with previous work in the field, or to provide an acceptable explanation, depends on the present state of the art. The fact that methodological rules are illustrated by the very ways in which we construct our theories, though, does not mean they have to be blindly followed, nor that they are not subject to rational scrutiny.

One must distinguish between the reasons for following a rule and the rationality of the rule itself. Even if the relativist's account does not provide reasons, it nonetheless offers an implicit definition of rule-rationality. Any existing methodological rule is rational insofar as it serves some 'positive' social function. Methodological rules, being means to an end, are thus at least instrumentally rational. Instrumental rationality, though, does not get us very far in accounting for convergence. If it is the case that a number of methodological rules may equally well serve the same function, local agreement in the absence of compelling reasons remains to be explained.

To say that many methodological rules may coexist does not imply that *any* set of such rules would be acceptable at a given time. There are

4 This form of determinism is already present in Ludwig Fleck (1979) who is in many respects one of its originators. In discussing how what he called 'thought styles' exert an almost absolute constraint on thought, Fleck states that 'the individual within the collective is never, or hardly ever, conscious of the prevailing thought style, which almost always exerts an absolutely compulsive force upon his thinking and with which it is not possible to be at variance' (Fleck 1979, p. 41). If we substitute 'thought style' with 'cultural influence', this statement could easily be attributed to a member of the Edinburgh School.

obvious constraints on what is acceptable and what is not; these limits are set by the history of the field, the availability of techniques and the individual creativity of the practitioners. Some of these constraints *are* epistemic. The important point to be made is that, although epistemic considerations play a role in restricting the class of available methods, there often remains a wide scope for variation. One cannot find, in other words, *conclusive reasons* for choosing one method over another.

In what follows, I shall provide reasons for adopting methodological rules which are neither epistemic nor functional. In so doing, I will model consensus formation by means of a methodological individualist explication of the concept of methodological rule. These rules are the result of a configuration of individual choices in which the choices are not isolated, but rather 'strategic'. Such reconstruction avoids the problem, encountered by the functionalist, of having to prove the existence of a causal link between social interests and scientific outcomes, since it admits the arbitrariness of choice. At the same time, by modelling methodological rules as the outcome of an interaction, we acknowledge the social element involved in the choice.

Before introducing into the picture the idea of choice, we need to better specify the concept of 'methodological rules'. These rules, for one, are not codified by some authority, nor are they enforced by sanctions. They are informal, yet we want to reconstruct their emergence as the result of individual choices. We want to claim that these rules are chosen, in some sense as yet unspecified; that neither is one forced to follow them, nor are they mere constraints on choice.

We may say that they are supported by a special preference structure, such that my preference for compliance is *conditional* on the expectation that others will comply, too. This condition is stated by David Lewis (1969) as necessary to define conventions. However, it is not sufficient for claiming that methodological rules are conventions. For example, we may have conditional preferences in cases of conformism where the rule or behaviour conformed to is by no means arbitrary. There are people who vote for a party because their family or friends do, but this does not mean that the family or friends in question vote for that party for the same reasons.

A key feature of conventions is that they are *arbitrary*, they could have been otherwise. If, as I am suggesting, methodological rules *qua* conventions are the result of a choice, then we have to define what an arbitrary choice is. Intuitively, we say that a choice is arbitrary if—given some end we want to attain—another choice would have served it equally well. There are, in other words, many means to a given end.

In an interactive environment, where the attainment of a goal crucially depends on what a number of people do, a successful choice is a choice which is 'coordinated' with all the other people's choices. Arbitrariness here means that there exist many possible configurations of choices, all capable of attaining the given goal. A successful combination of actions is an 'equilibrium'; i.e., an equilibrium is a situation where no one would have any incentive to change her behaviour, given the others' actions.

Now suppose we have a common interest in achieving a result. This can be achieved only by our joint efforts. When there are at least two disjoint combinations of actions enabling the result to be achieved, we have a coordination problem. Since all agents must do their part to attain the goal, an equilibrium in this case is defined somewhat differently. A *coordination equilibrium* is a situation where no one would have any incentive to change his behaviour, given the others' actions, and no one wishes that the others would change either.[5]

Arbitrariness can thus be expressed as the existence of a multiplicity of equilibria. If a main feature of conventions is arbitrariness, and if conventions are social phenomena, in that they are the result of individual choices, then the outcomes of these arbitrary choices can be modelled as equilibria of coordination games.[6] The view that conven-

5 The main characteristics of coordination problems can be listed under the following heads: (*i*) there are $n \geqslant 2$ individuals, each of whom has to choose one from among several alternative actions; (*ii*) the outcome of each individual's action depends upon the action chosen by each of the others and all know that; (*iii*) there is coincidence, or quasi-coincidence, among the interests of the parties. Thus the gain (loss) of one is the gain (loss) of all, where the magnitude of the gain or loss may vary between players. Coordination problems can be represented as a species of games of strategy. When there is perfect coincidence of the parties' interests, this feature can be represented in payoff matrix notation, where the parties' payoffs are equal in every cell.

		II		
		A	*B*	*C*
	D	2, 2	0, 0	0, 0
I	*E*	0, 0	2, 2	0, 0
	F	0, 0	0, 0	2, 2

The matrix depicts a two-person coordination game, where the strategic problem for each agent is that she wants to coordinate her choice of strategy with the other player's choice, given the fact that both stand to gain if they succeed in concerting their choices. The diagonal cells *AD*, *BE*, *CF*, represent the combinations of strategies that have the best result for both players. These diagonal payoffs combinations are all coordination equilibria in the sense that, once one of them is attained, neither player has any incentive to change her behaviour, given the other players's behaviour, and neither player wishes that the other would change either. Given the multiplicity of equilibria, the problem of coordination consists in concerting the actions of all the parties involved so that one of them is aimed at by them all.

6 Margaret Gilbert (1981) has convincingly argued that Lewis' definition of nontriviality is inadequate. From the consideration that two equilibria (in Lewis' sense) are not a necessary condition for nontriviality, and that a trivial case (in Lewis' sense) may not be a case in which it is obvious what to do, she concludes that conventions need not arise out of coordination problems. What is inadequate, however, is only Lewis'

tions arise out of coordination problems is spelled out in Lewis (1969), and in what follows I will refer to his account.[7] Before discussing the implications of viewing methodological rules as conventions, however, I shall review Lewis' theory, since my account differs from his in some important respect.

While I accept Lewis' claim that conventions only exist in the context of coordination problems, I reject his equation of conventions with regularities in behaviour. It is useful to treat the emergence and the establishment of a convention as two different things, so that the arbitrariness feature and the regularity feature are distinguished.[8] Also, while Lewis requires that the participants in a convention know what they are doing, I drop this requirement in the case of descriptive accounts of conventions. While the participants' knowledge is required in a rational reconstruction, it is not necessary in a descriptive account of conventions.

4. CONVENTIONS AS REGULARITIES IN BEHAVIOUR

Lewis claims that a regularity R, in action or in action and belief, is a convention in a population P if and only if, within P, the following six conditions hold:

(1) Almost everyone conforms to R.

(2) Almost everyone believes that the others conform to R.

(3) This belief that the others conform to R gives almost everyone a good and decisive reason to conform to R himself.

(4) There is a general preference for general conformity to R rather than slightly-less-than general conformity—in particular, rather than conformity by all but one.

(5) There is at least one alternative R' to R such that the belief that the others conformed to R' would give almost everyone a good and decisive practical or epistemic reason to conform to R' likewise; such that there is a general preference for general conformity to R' rather than slightly-less-than general conformity to R'; and such that there is normally no way of conforming to R and R' both.

(6) (1)-(5) are matters of common knowledge (Lewis 1969, p. 78).

There are at least two crucial objections to this account of conventions. The first has to do with conditions (2) and (5), the second with condition (6).

definition of a 'proper equilibrium'. As I show in a forthcoming paper, other equilibrium concepts allow us to properly identify trivial cases.

7 See also Schelling (1960) and Ullmann-Margalit (1977).

8 Margaret Gilbert has shown that (*a*) there are conventions which are not regularities, such as one-shot signalling conventions (1983*a*), and (*b*) that conventions *do not imply* any regularity in behaviour (1983*b*). I argue along similar lines in a recent paper (Bicchieri 1988).

What is stated here is that, in order to have a convention, we need: (*a*) conditional preferences, such that we prefer to conform on condition that others do conform, too; (*b*) at least two alternatives or, in Lewis' terms, two coordination equilibria; and (*c*) common knowledge. The conditional preferences, coupled with a belief in the conformity of others, elicit one's conformity to a convention. Expectations of other people's behaviour are therefore crucial. We may well ask if there is any ground for these reciprocal beliefs, if we are offered a reasonable account of how we form expectations about other people's conformity.

According to Lewis, condition (2) can be derived from the following chain of reasoning:

(i) I expect you to prefer to conform to *R* if I conform to *R*.
(ii) I expect that you expect that I conform to *R*.
(iii) Therefore, I have reason to expect that you have reason to prefer to conform to *R*.
(iv) I expect you to be rational.
(v) Therefore I have reason to expect you to conform to *R*.
(vi) Then I expect you to conform to *R*.[9]

Nothing, however, guarantees the correctness of premiss (ii). There is, indeed, an infinite regress of expectations, since I expect you to expect me to expect you . . . without an end. There is no more ground for expecting you to conform to *R* than expecting you to do something else. Lewis argues that somewhere, at a higher level, the regress would be truncated (p. 31). But for the regress to come to an end, there must be some ground somewhere to justify our reciprocal expectations.

What Lewis seems implicitly to assume, without justification, is that people have *equilibrium expectations*, i.e., correct expectations. What correctness means in this context is that my expectation about your conforming to *R* is correct, and vice versa. These expectations, coupled with our conditional preference for conformity, lead us to conform. Moreover, expectations have to be correct, or consistent, otherwise no convention could exist. The game-theoretic model of conventions depicting them as coordination equilibria shows this very clearly. It may of course happen that we entertain mutually consistent expectations, and are therefore able to coordinate, but this does not mean our expectations were rational in the first place. We may have had no ground to expect what we did (Bicchieri 1988).

The problem of grounding expectations is essentially the same as the problem of how a coordination equilibrium is attained. We can, of course, do away with expectations altogether, and admit that an equilibrium can be reached by chance. Since Lewis examines only one-shot games, this is a highly unlikely event. In fact, Lewis offers a number of possible grounds for equilibrium expectations. As we shall see, a close

9 Lewis (1969), pp. 28ff.

examination of them leads us to conclude that a game-theoretic account of conventions *does not imply* that they are regularities in behaviour.

One possible ground for equilibrium expectations is *agreement*.[10] Suppose we agree to meet at a given place and date. Then, given that we want to meet, are both rational, and have common knowledge about all that, we have reason to expect the agreement to hold. This agreement is a basis for common knowledge that we will meet at the agreed upon place. Because of this agreement, our mutual expectations are grounded, and thus rational, and always because of the agreement, they are also convergent, which guarantees coordination. There is little to object to this argument; however, the most interesting cases of conventions are those that are not founded on explicit agreements, and hence require grounding of another sort.

The most common proposal involves a concept of *salience*.[11] It is claimed that, in the absence of communication, we would try for salient equilibria, that is, equilibria that have some characteristic sorting them out as special. The interesting case is that in which salience is not incorporated into the utility information.[12] In this case, however, it is very difficult to imagine people having a unique criterion for salience, and this being common knowledge. If there are many criteria, we cannot be sure the other party assigns to them the same weights we do, or even shares the same set of criteria.

Even in the case of a unique criterion, however, nothing guarantees that all players will follow it. Their rationality and preference structure and common knowledge of both of them is not enough. To see why this is so, let us take the simplest case of a coordination game and follow the reasoning of the players.

		Col	
		A	*B*
Row	*C*	2, 2	0, 0
	D	0, 0	2, 2

10 Ibid., pp. 55ff.

11 We find this solution also in Schelling (1960), Ullmann-Margalit (1977), and Gauthier (1975).

12 Otherwise, we would have the following cases. Either the salient equilibrium is better than the others, and thus it would be chosen on that account. But this would mean that there would be one obvious thing to do for all the participants, and the case would be trivial. A trivial case is a non-arbitrary case and then, according to Lewis, it would not count as a convention. Or the salient equilibrium is worse than the others, but in this case it is not clear that it would be chosen on that account. If it is notably worse, there is no reason to expect the players to choose it, since the utility loss may loom large with respect to the risk of not coordinating. Another case is that in which the interests of the parties do not coincide, but in this case one would expect them to solve the underlying bargaining game before reaching a coordination solution.

If Row player thinks it is likely that Column player (Col) chooses strategy B, then it makes sense for her to choose strategy D. This, however, makes sense only if Col has reason to think that Row thinks it is very likely that he chooses B; then he may expect her to choose D. And so he should choose B. But Row thinks Col will probably choose B only if she can reasonably believe him to have a reason to do so. This reason is grounded on the salience of B. Since salience is not represented in utility terms, it can provide Col with a reason to choose B only if he has some reason to expect Row to be influenced by it. Or if he expects Row to expect Col to be influenced by it. Salience, in other words, provides me with a reason to choose a particular strategy only if I think you believe that salience provides me with such a reason. It gives me a reason only if it gives you a reason, but it gives each of us a reason only if it gives a reason to both.

To solve the regress, we need to assume there is a rule which says that, whenever we can use this or that salience criterion, we have to use it. Or maybe it is just a convention that we use it. In this case, we resort to a meta-convention to explain how a convention emerges. But then the meta-convention itself stands in need of explanation.

Another proposal is that equilibrium expectations be grounded upon *precedent*. According to Lewis, the most common case in which we attain an equilibrium is the case of a coordination problem that we (or people we know about) have already encountered many times before. This is a special case of salience, where the characteristic sorting out one equilibrium as relevant is the fact that it has been reached in the past. In this case, the source of mutual expectations would be the acquaintance with past solved instances of the present problem.[13] My claim is that common knowledge of the past solution and of each other's rationality and preferences is not enough to form a rational expectation that we will get to the same equilibrium as before.

Suppose, for example, that in the case discussed above the salient equilibrium is BD, since it has been reached before. The previous reasoning may be repeated leading to the conclusion that this knowledge is not enough to guarantee that the same strategies will be chosen the next time. In fact, it may happen that the participants think that now it is the turn of AC to be chosen, just for the sake of variety or fairness or any other criterion we may imagine. Unless there is some higher order convention indicating that most people follow the precedent in such circumstances, we are not able to form any rational expectation.

Expectations must be correct for a coordination equilibrium to be attained. For them to be correct, we need a fair amount of common knowledge, not only of our rationality and preferences but also of our reciprocal expectations. Agreement aside, salience and precedent do

13 Lewis (1969), p. 36.

not serve as adequate bases for such common knowledge. As Lewis always refers to one-shot games, it is difficult to see how an equilibrium can be attained at all.

While retaining the definition of conventions as arbitrary solutions of coordination problems (i.e., as equilibria of coordination games), I propose a modified version of Lewis' account. Although conventions may arise out of agreements in one-shot games, the most common case is that of repeated coordination games. In this case, it is reasonable to expect an equilibrium to be attained *by chance* after a few unsuccessful attempts at coordination. Here by 'equilibrium' is only meant that actions are not regretted *ex post*, not that expectations turn out to have been correct. Expectations play no role in this case.

This account of conventions retains the feature of arbitrariness, in that it is acknowledged that different conventions might have been established. It does away, however, with the equation of conventions with regularities in behaviour. After an equilibrium is once attained, nothing guarantees that the same way of coordinating actions will be chosen in successive instances of the same problem. The emergence of conventions and their stability are thus two quite different issues.

Nevertheless, many instances of convention involve regularities. That is, once a solution to a coordination problem is attained, that solution tends to be repeated whenever the same problem arises. This can be accounted for in the following way. One may argue that the strategic situation faced by the agents is no longer the same after an equilibrium has once been attained. This situation can be depicted by means of a new game with only two possible strategies: either to take this fact into account or to play as if nothing had happened.[14] Ignoring means randomizing between strategies.

The original game will thus be transformed into the following 'salience game':[15]

	seek salience	ignore
seek salience	2, 2	1, 1
ignore	1, 1	1, 1

Here there is just one best equilibrium, and we can confidently expect rational players to choose it on that account.[16] Expectations in this case

14 David Gauthier (1975) first proposed this solution. His approach, however, has the same defect as Lewis's, in that he takes salience to be the main solution to a primitive coordination game.

15 The payoffs can be thus explained: if we both seek salience, we get the same payoffs as in the original game. If one or both ignore, our payoffs are the expected utilities of the respective strategies, with equal probabilities for each.

16 It can be shown that any salience game has a unique best outcome. This two-stage

are grounded on the assumed rationality of the players. This explains how a successful solution to a coordination problem can become a regularity in behaviour. A possible objection to this solution is that nothing guarantees that the players will 'seek salience' in the same way. That is, one player might understand salience as following a precedent, while the other might interpret it in the opposite way. But would there be a good reason for *not* sticking to the mutually agreeable solution already attained? I believe not. Indeed, I believe that what we usually call 'habit' is precisely the tendency to cast more weight upon something we know has obtained and is mutually satisfactory for the parties involved. The very fact of not having a good reason to expect otherwise justifies the parties in interpreting the 'seek salience' strategy in the unique way proposed above.

The reasons why a convention emerges are thus very different from those which account for its stability. The 'salience game' shows that a convention, once established, can become a regularity because it is in the interest of each and every participant to follow it. Since stable conventions are the outcome of rational choices, they do not need to be enforced. Conformity, not compliance, occurs, in that fear of sanctions is not the motive to follow a convention. The non-conformist is not punished. He simply opts out from the community (Ullmann-Margalit 1977).

5. COMMON KNOWLEDGE

The second objection to Lewis' account regards his requirement that conditions (1)-(5) be common knowledge. Common knowledge is indeed an essential requirement for obtaining a convention, in that agents must know of each others' rationality and preferences, and know that they all know, . . . ad infinitum. The same is true in maintaining a convention. However, as Tyler Burge (1975) has persuasively argued, 'the stability of conventions is safeguarded not only by enlightened self-interest, but by inertia, superstition and ignorance' (p. 253).

This consideration requires not so much a rejection of Lewis's theory of conventions, as the drawing of a distinction between a *rational reconstruction* and a *descriptive account*. Reconstructing certain practices as conventional does not involve describing the actual beliefs of the people involved in these practices. Many people participating to a convention may not be aware that it is a convention, they may indeed think their practices to be natural and eternal. If choices are involved, they may not believe that there exist equally good alternatives.

solution answers Gauthier's problem that, in games where there is no perfect coincidence of interests, we may have no unique best equilibrium. Uniqueness can be attained by playing the salience game as a secondary game only. Cf. Gauthier, ibid., pp. 214ff.

Our interpretation of the scientific past is a case in point. What we now see as relative and conventional was certainly not perceived in these terms at the time. The authors of the memoirs appearing in the first scientific societies' bulletins were preoccupied with chronicles of observations, experiments, descriptions of how experiments are set up. They were unaware of what we now see: the growing homogenization of scientific language, or the fact that scientific language was becoming more and more identified with the use of certain experimental techniques and mathematical modelling. What we now observe, looking backward, is the establishment of a regular pattern of scientific behaviour; doubtless long ago these practices were not regarded as arbitrary, as one among other equally good means to do science. An historical account cannot overlook these beliefs, the actors' sense that their activities were in some sense determined by the nature of things.

In a rational reconstruction, we may discount those beliefs. But a rational reconstruction, to be of any use, *must not be completely false*. We may depict people as if they were perfectly rational, had certain preferences and expectations, and had common knowledge of all that. This may be an inaccurate and idealized account, not a false one. It must be true that *if* the participants to a convention were made aware of the functions it fulfills, and were made available an equally good means to coordinate upon, and if they were to believe that a shift in allegiance was taking place, they would prefer to shift to the new practice too. It must also be true that those who prefer to retain the old convention would do so because coordinating with others, and thus participating and sharing in their community's life, would be an inferior choice.[17] We may indeed prefer to stick to a given practice for emotional reasons or otherwise, even at the cost of exclusion. Conventions exist even if not all the members of the population follow them.

It might be further objected that methodological conventions, being a sort of habitual regularity, manifest no more rationality than does instinctive behaviour. While scientists keep using the same procedures and follow the same rules, we would not want to say they keep solving the same problem rationally.[18] David Lewis has addressed this issue in later articles (1975, 1976), correctly concluding, in my view, that 'an action may be rational, and may be explained by the agent's beliefs and desires, even though that action was done by habit. . . . A habit may be under the agent's rational control in this sense: if that habit ever ceased to serve the agent's desires according to his beliefs, it would at once be overridden and corrected' (1975, p. 25). An explanation in terms of habit does not compete with one in terms of expectations and preferences. The habit persists without interference precisely because of certain

17 Cf. Tyler Burge's discussion of the 'language of the Gods', ibid., p. 251.
18 For a similar criticism, see Dale Jamieson (1975) and Eike von Savigny (1985).

expectations and preferences: if I ever wanted to be different, or if I expected others to do something different, I would probably overcome the force of habit.

Very much like the sociologist of science, we impute something to the agents of which the agents themselves might be unaware. While the sociologist imputes interests, we are imputing preferences and desires. People do not need to be conscious of their motives all the time, in that these motives are *dispositions* to act in certain ways in given circumstances. What is needed is that such motives be ready to manifest themselves in the relevant circumstances.

In conclusion, arbitrariness may not be represented in the participants' beliefs, and the basis for their participation may be either misunderstood or ignored. But it must be true that—whatever people believe—they can always learn and adapt to a new convention. This condition allows the rational reconstruction to be inaccurate, but not false.

6. METHODOLOGICAL CONVENTIONS

The above considerations apply to a rational reconstruction of scientific methodologies as systems of conventions. This account seeks to explain how a 'local' homogeneity of method is attained, under what circumstances a consensus is likely to emerge, and not how actual historical processes took place.

If one accepts the egalitarianism of criteria principle, one has to admit that methodological convergence is often an arbitrary affair. This arbitrariness is well depicted by the structure of a repeated coordination game. The idea that an equilibrium is attained by chance may be disturbing, but it is the logical consequence of *denying* that there exists a superior rationality criterion regulating our choices and *affirming* that the outcome of these choices is epistemically, culturally, and socially underdetermined.

It might be objected that this account is too restrictive, in that methodological rules fulfill all sorts of functions, and they are chosen on that account. I do not deny that methodological rules fulfill many other functions—cognitive and social alike—beyond that of coordination. All these functions are, however, secondary, in that they all depend upon coordination, or a previous agreement (however implicit) as to what set of rules is to be adopted.

Suppose convention C fulfills function F. Why not say that the participants follow C because of F? C is a means to attain F, and it is followed on that account. F however can be fulfilled only if all (or almost all) the participants follow C. Thus each follows C because each wants F, and C is a suitable means, *and* because each expects the others to want F too, and thus to make the same choice. Without this expectation, no one would have any reason to follow C. Expectation of conformity is a *precondition* for any function F to be fulfilled. This expectation, though, cannot be grounded on F, since it is a precondition for its existence.

If C were the only means to attain F, then one's expectation would be grounded upon the knowledge that all others would likewise conform to C, since one knows that they all want F. However, if C is only one of the possible means to attain F, all equally good, knowledge that all want F is not enough, since multiplicity deprives one of the grounds for forming a rational expectation. It follows that reciprocal expectations can be grounded only after some C has been attained, for reasons quite different from its fulfilling F. C is thus arbitrary; it is, indeed, a coordination equilibrium.

Intuitively, it seems likely that the epistemic and social functions we commonly attribute to methodological rules could not be fulfilled without previous coordination on some set of such rules. Of course, our shared cognitive aims, background knowledge and experience constrain our possible choices, but if we admit of multiplicity, we have to acknowledge that such functions are necessary but not sufficient conditions for convergence. They are necessary, since methodological rules that do not embody, for example, accepted scientific knowledge would not be even considered. Experimental rules in physics embody well established theories and principles, testing rules in economics embody accepted statistical theories and procedures. In many (if not all) cases, however, epistemic acceptability does not indicate a unique choice.

The same is true with social functions. An interest in social engineering certainly requires a concern for predictability, but there still remain many ways to interpret this general goal. Mendelians, for example, were as interested in prediction as their opponents. Predictability in the Mendelian case implied selected populations, controlled settings, and manipulative intervention, which made it more cumbersome and problematic than biometrical techniques. These latter, however, could be found unreliable if one accepted the Mendelian distinction between variations that contributed to evolution and variations that did not, since the statistical aggregates of biometricians could not distinguish between the two.

Even if the relativist provides us with a rule-rationality criterion, this never becomes *the* reason for choice. His rationality criterion (i.e., a methodology is rational if it fulfills some social and/or cognitive goals) is too weak to fully account for the 'local' convergence to a set of methodological rules.

Modelling scientific methodologies as systems of conventions has the advantage of making them *individually rational*. Being equilibria of coordination games, they are the outcome of social interaction and individual choice, hence they are not mere constraints on choice. While no set of rules is chosen because it is the uniquely best means to a given end (otherwise, there would be no arbitrariness), it becomes optimal once it has been attained. Methodological rules thus become collectively rational—in a purely instrumental sense—only after they have been established. This explains why the reasons for the emergence of a scientific methodology differ from the reasons for its stability; moreover, since stability—and thus abiding by the rules—is representable as the result of a rational choice, we may expect these rules to be resistant to change, or to change only very slowly, which is in fact what we observe within scientific practices.

REFERENCES

Barnes, B. (1974) *Scientific Knowledge and Sociological Theory*. London.
———— (1977) *Interests and the Growth of Knowledge*. London.
Barnes, B. and Bloor, D. (1982) 'Relativism, Rationalism and the Sociology of Knowledge' in M. Hollis and S. Lukes (1982).
Bicchieri, C. (1988) 'Strategic Behavior and Counterfactuals', *Synthèse*, **76**, 135-69.
Bloor, D. (1976) *Knowledge and Social Imagery*. London.
Burge, T. (1975) 'On Knowledge and Convention', *Philosophical Review*, **84**, 249-55.
Elster, J. (1979) *Ulysses and the Syrens*. Cambridge.
Fleck, L. (1979) *Genesis and Development of a Scientific Fact*. Chicago. (1st ed., 1935.)
Gauthier, D. P. (1975) 'Coordination', *Dialogue*, **14**, 195-221.
Gilbert, M. (1981) 'Game Theory and Convention', *Synthèse*, **46**, 41-93.
———— (1983*a*) 'Agreement, Conventions and Language', *Synthèse*, **54**, 375-404.
———— (1983*b*) 'Some Limitations of Rationality', paper presented to the A.P.A. meetings. December.
Hollis, M. and S. Lukes, (eds.) (1982) *Rationality and Relativism*. Cambridge, Mass.
Jamieson, D. (1975) 'David Lewis on Convention', *Canadian Journal of Philosophy*, **5**, 73-81.
Kuhn, T. S. (1962) *The Structure of Scientific Revolutions*. Chicago.
———— (1977) *The Essential Tension: Selected Studies in Scientific Thought and Change*. Chicago.
Lewis, D. K. (1969) *Convention*. Cambridge, Mass.
———— (1975) 'Languages and Language', in K. Gunderson (ed.), *Minnesota Studies in the Philosophy of Science*, vol. 7. Minneapolis.
———— (1976) 'Convention: Reply to Jamieson', *Canadian Journal of Philosophy*, **6**, 113-20.
MacKenzie, D. and Barnes, B. (1975) 'Scientific Judgement: The Biometry-Mendelism Controversy,' in B. Barnes and S. Shapin (eds.), *Natural Order*. London.
McMullin, E. (1984) 'The Rational and the Social in the History of Science', in J. R. Brown (ed.), *Scientific Rationality: The Sociological Turn*. Dordrecht.
Mulkay, M. (1979) *Science and the Sociology of Knowledge*. London.
Schelling, T. C. (1960) *The Strategy of Conflict*. Oxford.
Ullmann-Margalit, E. (1977) *The Emergence of Norms*. Oxford.
von Savigny, E. (1985) 'Social Habits and Enlightened Cooperation: Do Humans Measure Up to Lewis Conventions?', *Erkenntnis*, **22**, 79-96.

Philosophy & Social Criticism
Volume 13 No. 4

Philosophy & Social Criticism P.O. Box 368, Lawrence, KS 66044

EDITORIAL CORRESPONDENCE AND BOOKS FOR REVIEW should be sent to the Managing Editor, Department of Philosophy, York University, North York, Ontario, Canada M3J 1P3.

Toute correspondence éditoriale et tous livres soumis pour critique doivent être adressé au rédacteur gérant, Department of Philosophy, York University, North York, Ontario, Canada M3J 1P3.

PHILOSOPHY OF THE SOCIAL SCIENCES is supported by a grant from the Social Sciences and Humanities Research Council of Canada #441-86-0168.

PHILOSOPHIE DES SCIENCES SOCIALES est subventionnée par le Conseil de recherches en sciences humaines du Canada #441-86-0168.

PHILOSOPHY OF THE SOCIAL SCIENCES is now being made available in microform. Enquiries and orders should be sent to University Microfilms Ltd., St. John's Road, Tyler's Green, High Wycombe, Bucks. HP10 8HR, England—or for orders from the Americas, to Xerox University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106, U.S.A.

La revue est enregistrée sur microfilm. Prière de s'adresser à University Microfilms Ltd., St. John's Road, Tyler's Green, High Wycombe, Bucks. HP10 8HR, England, ou à Xerox University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106, U.S.A.

SUBSCRIPTION RATE: $30.00 per year for 4 issues (**Canadian or U.S. funds only**). All correspondence concerning subscriptions and address changes should be sent to **Wilfrid Laurier University Press, Wilfrid Laurier University, Waterloo, Ontario, Canada N2L 3C5**. Second-class mail registration #3827.

Abonnement annuel $30.00 pour quatre numéros. Prière de faire les chèques ou ordres de paiement en dollars canadiens ou américains à l'ordre de: **Wilfrid Laurier University Press, Wilfrid Laurier University, Waterloo, Ontario, Canada N2L 3C5.** Régistration postale deuxième classe #3827.

BACK NUMBERS

A limited number of back issues are available for $8.00 each. Complete volumes for $30.00. These are available from: **Wilfrid Laurier University Press, Wilfrid Laurier University, Waterloo, Ontario, Canada N2L 3C5.**

Volumes 1-8 are out of print. These volumes are being reprinted by: **Topos Verlag AG, Aulestrasse 74, P. O. Box 668, FL-9490 Vaduz, Leichtenstein**, to whom all orders and inquiries in respect to these volumes should be directed.

PHILOSOPHY OF THE SOCIAL SCIENCES

Volume 18, Number 4
December 1988

PHILOSOPHIE DES SCIENCES SOCIALES

Volume 18, Numéro 4
Décembre 1988

CONTENTS

Continued on page 588 . . .

Printed in Canada/Imprimé au Canada