SYNTHESE

AN INTERNATIONAL JOURNAL

FOR EPISTEMOLOGY, METHODOLOGY AND

PHILOSOPHY OF SCIENCE

Volume 68 (1986)

D. REIDEL PUBLISHING COMPANY

A MEMBER OF THE KLUWER ACADEMIC PUBLISHERS GROUP

DORDRECHT / BOSTON / LANCASTER / TOKYO

McCall, W. A.: 1923, *How to Experiment in Education*, Macmillan, New York.
Meehl, P. E.: 1978, 'Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald and the Slow Progress of Soft Psychology', *Journal of Consulting and Clinical Psychology* **46**, 806–834.
Michotte, A.: 1963, *The Perception of Causality*, 1st English edn., Basic Books, New York.
Mill, J. S.: 1906, *A System of Logic*, Longman Greene's & Co., London.
Millikan, R. G.: 1984, *Language, Thought, and Other Biological Categories*, Broadford Books of MIT Press, Cambridge, Mass.
Moyer, D. F.: 1979, 'Revolution in Science: The 1919 Eclipse Test of General Relativity', in A. Perlmutter and L. F. Scott (eds.), *On the Path of Albert Einstein*, Plenum Press, New York, pp. 55–150.
Popper, K. R.: 1959, *The Logic of Scientific Discovery*, Basic Books, New York. (Originally *Die Logik der Forschung*, 1935.)
Popper, K. R.: 1972, *Objective Knowledge: An Evolutionary Approach*, Clarendon Press, Oxford.
Putnam, H.: 1983, *Realism and Reason*, Cambridge University Press, Cambridge.
Quine, W. V.: 1963, 'Two Dogmas of Empiricism', in *From a Logical Point of View*, Harper & Row, New York, pp. 20–46 (originally 1951).
Quine, W. V.: 1969, 'Epistemology Naturalized', in *Ontological Relativity and Other Essays*, Columbia University Press, New York. pp. 90–99.
Quine, W. V.: 1974, *The Roots of Reference*, Open Court, LaSalle, Illinois.
Quine, W. V.: 1975, 'The Nature of Natural Knowledge', in S. Guttenplan (ed.), *Mind and Language*, The Clarendon Press, Oxford, pp. 67–81.
Rosenberg, A.: 1983, 'Causation and Recipes: The Mixture as Before', *Philosophical Studies* **24**, 379–385.
Scriven, M.: 1971, 'The Logic of Cause', *Theory and Decision* **2**, 49–66.
Scriven, M.: 1975, 'Causation as Explanation', *Nous* **9**, 3–16.
Scriven, M.: 1976, 'Maximizing the Power of Causal Investigation: The Modus Operandi Method', in G. V. Glass (ed.), *Evaluation Studies Review Annual*, Vol. 1, Sage Publications, Beverly Hills, CA.
Suppes, P.: 1970, *A Probabilistic Theory of Causality*, North-Holland, Amsterdam.
van Fraassen, B.: 1980, *The Scientific Image*, Clarendon Press, Oxford.
Weber, S. J. and T. D. Cook: 1972, 'Subject Effects in Laboratory Research: An Examination of Subject Roles, Demand Characteristics, and Valid Inference', *Psychological Bulletin* **77**, 273–295.
Whitbeck, C.: 1977, 'Causation in Medicine: The Disease Entity Model', *Philosophy of Science* **44**, 619–637.
Wright, G. H. von: 1971, *Explanation and Understanding*, Cornell University Press, Ithaca, NY.

Department of Psychology
Northwestern University
Evanston, IL 60201
U.S.A.

PHILIP PETTIT

PRESERVING THE PRISONER'S DILEMMA

There is an argument in circulation to the effect that the rational thing for parties to do in an idealized prisoner's dilemma is to cooperate.[1] This runs counter to the orthodox view that since defecting dominates cooperating—it is better for each whether the other cooperates or defects—then it is the rational strategy. I wish to show that the argument, at least under my reconstruction, is fallacious. I suspect that every version of the argument is similarly fallacious but I shall not try to demonstrate that.

The prisoner's dilemma can be depicted in a matrix like the following, where "C" stands for the cooperative strategy and "D" for the strategy of defection.

	C	D
C	R, R	S, T
D	T, S	P, P.

The payoff to Row is given first in each case and the letters describing payoffs are interpreted in the usual way: "R" represents reward, "P" punishment, "T" temptation and "S" the sucker's payoff. We have a prisoner's dilemma if each prefers T to R, R to P, and P to S.[2]

In the idealized prisoner's dilemma, we make a strong assumption about the rationality and the knowledge of the participants. We assume that it is common knowledge between them that they are rational: they are each rational, they each know that this is so, they each know that they each know that this is so, and so on.[3] It need not matter how rationality is understood in the present context, as long as it entails the following: that if a rational agent knows he can obtain m by performing one of two alternative actions, n by performing the other, and m is better by his standards, then he performs the first alternative: he unerringly maximises preference satisfaction.

The assumption of known rationality, as we may dub it, plays an important role in the argument which I wish to criticise. So too does the assumption of symmetry, according to which the decision situation of the two participants is identical. This is implicit in the matrix represen-

Synthese **86** (1986) 181–184.
© 1986 *by D. Reidel Publishing Company*

tation given, for what it means is that each ranks T, R, P and S in the same order, and expects to get these payoffs under the same circumstances.

We can now formulate the argument which I reject. It is phrased in the first person, being allegedly a piece of reasoning by which each participant would deduce that he ought to cooperate.

(1) If we each act rationally, and our situations are symmetrical, then if I cooperate, the other cooperates too.
(2) If we each act rationally, and our situations are symmetrical, then if I defect, the other defects too.
(3) But we do each act rationally, and our situations are symmetrical.
(4) Therefore if I cooperate, the other cooperates; and if I defect, the other defects.
(5) I get the better result by cooperating: R rather than P; so, being rational, I should cooperate.

This argument is invalid. As a way of seeing why, contrast it with the following clearly acceptable argument.

(1′) If we each act rationally, and our situations are symmetrical, then if I cooperate it is rational for me to cooperate and if it is rational for me to cooperate then the other cooperates too.
(2′) If we each act rationally, and our situations are symmetrical, then if I defect it is rational for me to defect and if it is rational for me to defect then the other defects too.
(3′) But we do each act rationally, and our situations are symmetrical.
(4′) Therefore if I cooperate it is rational for me to cooperate and if it is rational for me to cooperate then the other cooperates too; and likewise if I defect it is rational for me to defect and if it is rational for me to defect then the other defects too.
(5′) I get the better result if I cooperate, assuming it is rational to cooperate; so, being rational, I should hope that it is.

It is clear that (1′) and (2′) differ respectively from (1) and (2) through

spelling out more explicitly the link between antecedent and consequent. In interpreting (1) and (2) therefore within the first argument we should be careful to take them in the sense of (1′) and (2′). We should not think for example that they mean that if I cooperate or if I defect, then the other cooperates or defects, regardless of the rationality of my action.

But if (1) and (2) have the sense of (1′) and (2′), then (3) does not enable us to infer (4), at least under the natural reading of (4). (3) allows us to deduce, as in the amended argument, that in the event of my cooperating rationally, the other cooperates too; and in the event of my defecting rationally, the other defects too. (4) suggests, more strongly, that whether it is rational or not, my cooperation will be matched by cooperation, my defection by defection.

Perhaps we should be charitable and understand (4) in the sense of (4′). But that will not help the cause of the argument. For (5) certainly does not follow from (4′). And under no construal, can (5) be made out to have the sense just of (5′). I conclude that wherever we choose to locate it, there is a flaw in the original argument. It is not a piece of reasoning by which participants in the idealized prisoner's dilemma would be led to cooperate.[4]

The flaw can be identified as follows. What (1′) and therefore (1) says has this form: if p then if q, r and if r, s. Similarly for (2′) and (2). (3) asserts that p and by detachment we get the result expressed in (4′): if q, r and if r, s. In words, if I cooperate then it is rational for me to cooperate and if it is rational for me to cooperate then the other cooperates too. (4) elides "r", saying that if q, s; if I cooperate, the other cooperates too. That is fine, so long as it is understood that "q" must continue to imply "r", that my cooperating must remain rational. However in the sense in which it yields (5), (4) is not subjected to this constraint. If it is sufficiently strong to deliver (5), then (4) says that if q, s, regardless of whether "q" implies "r": if I cooperate, the other cooperates too, whether or not I cooperate rationally.

The argument presented does not show that the idealized prisoner's dilemma is resoluble. The participants may dwell wistfully on the benefits to be got from both cooperating. But if defection is rational – as the argument from dominance has it – then the thought can never issue in action. Neither will he have reason to believe that if he were to cooperate – per impossibile, if defection is rational – then the other would cooperate too.[5]

NOTES

[1] The fullest version of the argument is in Lawrence H. Davis: 1977, 'Prisoners, Paradox and Rationality', *American Philosophical Quarterly* **14**. It is subjected to a reductio argument in Lanning Sowden: 1983, 'That There Is a Dilemma in the Prisoner's Dilemma', *Synthese* **55**. Davis and Sowden both give references to other statements of the argument.

[2] This is a sufficient condition but not, by some accounts, a necessary one. Some authors allow that each may be indifferent between P and S. On the condition given here, D dominates C. On the alternative requirement, DD is an equilibrium but D may not dominate C.

[3] For a useful discussion of this assumption see Davis, 'Prisoners, Paradox and Rationality, page 320. Notice that the assumption is not always spelled out, and may not even always be endorsed, in the game-theory literature.

[4] David Lewis has pointed out to me that instead of the interim conclusion (4), we might have had this: either I cooperate and the other cooperates or I defect and the other defects. This raises similar problems. Read in its (natural) weak sense, it fails to license (5); read in a strained sense in which it might be held to license (5), it fails to follow from the premises.

[5] I am grateful to Albert Weale for his spirited defence of a version of the argument rejected here. I am grateful to him and Geoffrey Brennan for discussion of some of the points that it raises; and to Frank Jackson, David Lewis, and an anonymous referee for written comments.

Research School of Social Sciences
Australian National University
Canberra, A.C.T. 2601

DAVID MAKINSON

HOW TO GIVE IT UP: A SURVEY OF SOME FORMAL ASPECTS OF THE LOGIC OF THEORY CHANGE

ERRATA

As the author was not given the opportunity to see the proofs of the survey (*Synthese* **62** (1985), 347–363), he has requested a page to indicate some of the updating and errata that would normally be taken care of at that stage.

1. The most important updating is that in the definition of a *safe* element of a set A, on page 361, we need not assume that the ordering $<$ over A is irreflexive and transitive. All we need to assume is the weaker condition that $<$ is ***non-circular***, in the sense that for no $a_1, \ldots, a_n \in A$ $(n \geqslant 1)$, $a_1 < a_2 < \cdots < a_n < a_1$. All the results mentioned for safe contraction continue to hold under this weakened assumption.

2. The paper of Alchourrón, Gärdenfors and Makinson 'On the Logic of Theory Change: Partial Meet Contraction and Revision Functions', listed as forthcoming, has appeared in *The Journal of Symbolic Logic* **50** (1985), 510–530. Also, the paper of Alchourrón and Makinson, 'On the Logic of Theory Change: Safe Contraction and Revision Functions' appeared in *Studia Logica* **44** (1985), 405–422.

3. It is also possible to construct explicit maps to reveal close interconnections between the 'partial meet' contractions and the 'safe' contractions. For example, for a theory A finite modulo Cn, the class of all relational partial meet contraction functions over A turns out to be identical with the class of all the safe contraction functions over A that are determined by a non-circular relation $<$ over A that continues both up and down $\vdash$. Such maps and equivalences are established in C. E. Alchourrón and D. Makinson 'Maps Between Some Different Kinds of Contraction Function: The Finite Case', to appear in *Studia Logica* **45** (1986).

4. There is a discussion of the difficulties of basing a semantics for conditionals upon the concept of theory revision in P. Gärdenfors, 'Belief Revisions and the Ramsey Test for Conditionals', forthcoming in *Philosophical Review*.