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

PHILIP PETITT

PRESEVING 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. 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.

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2,2</td>
<td>0,-2</td>
</tr>
<tr>
<td>D</td>
<td>-2,2</td>
<td>3,1</td>
</tr>
</tbody>
</table>

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.

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 this is so, and so on.2 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 unremittingly 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 represent-

...
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 Laurence Sowa 1983, ‘That There Is a Dilemma in the Prisoner Dilemma’, Synthese 55. Davis and Sowa 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 indeterminate 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 satisfied 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 defect. 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.

I am grateful to Alastair Walea for his spirited defense of a version of the argument rejected here. I am grateful to him and Geoffrey Brennan for discussion of some of the points that I raise; 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 351, 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, . . . , a_n ∈ A (n ≥ 1), a_1 < a_2 < . . . < a_n < a_1. All the results mentioned for safe contraction continue to hold under this weakened assumption.


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 modus Ponens, 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. 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.