The prisoner's dilemma

PHIL 20229
Jeff Speaks

March 25, 2008

1 The dilemma

You are one of two prisoners arrested for a crime. You, and the other prisoner, are each rational, and you each know that if you both stay silent, and don't confess, you will each be convicted of a fairly minor crime, and get 1 year in jail each. If you turn State's evidence and the other prisoner stays silent, then you will get off with nothing, and the other prisoner will get 10 years; exactly the opposite will happen if the other prisoner turns State's evidence, and you stay silent. If you both confess, you both get 5 years.

Is it rational for you to confess, or stay silent?

How could you use the Dominance Principle to argue that you ought to confess?

Consider the following counter-argument: you and the other prisoner are both rational; so, most likely, you will each evaluate the situation in the same way and come to the same decision. So, the only possibilities worth considering are the *matching possibilities*: possibilities in which you and the other prisoner do the same thing. Of these, there are two: you both stay silent, or you both confess. Of these two possibilities, the better one is that you both stay silent, since then you both get 1 year. So, you ought to stay silent. (This argument could also be put using the principle that you should act so as to maximize expected utility: the idea would be that we should assign the highest probability to the two 'matching' scenarios.)

How should the proponent of confessing respond to this argument?

How is this related to Newcomb's problem?

2 The tragedy of the commons

One view about the importance of the prisoner's dilemma links the dilemma to a familiar sort of conflict between individual interests and the interests of a group, which is sometimes illustrated by the example of the 'tragedy of the commons':

> Suppose that a number of dairy farmers live in a town. All have insufficient land for their purposes, so each would be better off if they could let their cows graze on the town common. But if each of them do this, the commons will be ruined for everyone.

The question in this case is: what is it rational for the individual farmer to do?

How is this supposed to be similar to the prisoner's dilemma? Can the dilemma here be reconstructed using dominance reasoning?

Cases like this are sometimes also called 'free rider' problems. There are many free rider problems in ordinary life. Can you think of any?

3 The multiple prisoner's dilemma

Suppose now that we have a case like the original prisoner's dilemma, but which involves multiple choices over time, and in which the relevant decision makers learn the results of the previous choices of the other decision makers. Does this change things, or should you just reason in the same way as previously each time?

How does this affect the questions above about the relationship between individual rationality and group interests, if at all?

4 A one-person prisoner's dilemma?

It seems at first as though the prisoner's dilemma essentially depends on their being more than one prisoner. And, in the version we discussed, it does depend on this. But we can come up with cases which in certain ways resemble the prisoner's dilemma, but involve only on person. Such a case is the 'paradox of the self-torturer,' in which different stages in the life of a single person can be thought of as different agents in a prisoner's dilemma.

Is the sort of series imagined in the paradox really possible? How many different increases in electric charge would we have to imagine in order for there to be a noticeable difference between the first and last, even though there was no noticeable difference between adjacent charges?