DAVID LEWIS

Prisoners' Dilemma Is a Newcomb Problem

Several authors have observed that Prisoners' Dilemma and Newcomb's Problem are related—for instance, in that both involve controversial appeals to dominance.[1] But to call them "related" is an understatement. Considered as puzzles about rationality, or disagreements between two conceptions thereof, they are one and the same problem. Prisoners' Dilemma *is* a Newcomb Problem—or rather, two Newcomb Problems side by side, one per prisoner. Only the inessential trappings are different. Let us make them the same.

You and I, the "prisoners," are separated. Each is offered the choice: to rat or not to rat. (The action of "ratting" is so called because I consider it to be *rat*ional—but that is controversial.) Ratting is done as follows: one reaches out and takes a transparent box, which is seen to contain a thousand dollars. A prisoner who rats gets to keep the thousand. (Maybe ratting is construed as an act of confessing and accusing one's partner, much as taking the Queen's shilling was once construed as an act of enlisting—but that is irrelevant to the decision problem.) If either prisoner declines to rat, he is not at all rewarded; but his partner is presented with a million dollars, nicely packed in an opaque box. (Maybe each faces a long sentence and a short sentence

1. Robert Nozick, "Newcomb's Problem and Two Principles of Choice" in *Essays in Honor of Carl G. Hempel*, ed. N. Rescher (Dordrecht: Reidel, 1969), pp. 130-131; Steven J. Brams, "Newcomb's Problem and Prisoners' Dilemma," *Journal of Conflict Resolution* 19 (1975): 596-612; Lawrence H. Davis, "Prisoners, Paradox, and Rationality," *American Philosophical Quarterly* 14 (1977): 321; and J. Howard Sobel, *Chance, Choice, and Action: Newcomb's Problem Resolved* (duplicated manuscript, July 1978), pp. 167-168.

Philosophy & Public Affairs 8, no. 3
0048-3915/79/030235-6$00.50/1

to be served consecutively; escape from the long sentence costs a million, and escape from the short sentence costs a thousand. But it is irrelevant how the prisoners propose to spend their money.) So the payoff matrix looks like this.

	I rat	I don't rat
You rat	I get $1,000 You get $1,000	I get $0 You get $1,001,000
You don't rat	I get $1,001,000 You get $0	I get $1,000,000 You get $1,000,000

There we have it: a perfectly typical case of Prisoners' Dilemma. My decision problem, in a nutshell, is as follows; yours is exactly similar.

(1) I am offered a thousand–take it or leave it.
(2) Perhaps also I will be given a million; but whether I will or not is causally independent of what I do now. Nothing I can do now will have any effect on whether or not I get my million.
(3) I will get my million if and only if you do not take your thousand.

Newcomb's Problem is the same as regards points (1) and (2). The only difference–if such it be–is that point (3) is replaced by

(3′) I will get my million if and only if it is predicted that I do not take my thousand.

"Predicted" need not mean "predicted in advance." Not so in English: we credit new theories with success in "predicting" phenomena already observed. And not so in Newcomb's Problem. While it dramatizes the problem to think of the million *already there*, or else already *not* there, in the opaque box in front of me as I deliberate, it is agreed all around that what really matters is (2), and hence that the "prediction" should be causally independent of my decision. Making the

prediction ahead of time is one good way to secure this causal independence. But it is not the only way.[2] Provided that I can have no effect on it, the prediction could just as well be made simultaneously with my decision or even afterwards, and the character of Newcomb's Problem would be unchanged.[3] Likewise in the case of Prisoners' Dilemma nothing need be assumed–and in my telling of the story, nothing was assumed–about whether the prisoners are put to the test simultaneously or one after the other.

Also it is inessential to Newcomb's Problem that any prediction–in advance, or otherwise–should actually take place. It is enough that some potentially predictive process should go on, and that whether I get my million is somehow made to depend on the outcome of that process. It could all be automated: if the predictive computer sends a pulse of current to the money-putting machine I get my million, otherwise not. Or there might be people who put the million in the box or not depending on the outcome of the process, but who do not at all think of the outcome as a prediction of my choice, or as warrant for a prediction. It makes no difference to my decision problem whether someone–the one who gives the million, or perhaps some bystander –does or doesn't form beliefs about what I will do by inference from the outcome of the predictive process.

Eliminating inessentials, then, Newcomb's Problem is characterized by (1), (2), and

> (3″) I will get my million if and only if a certain potentially predictive process (which may go on before, during, or after my choice) yields the outcome which could warrant a prediction that I do not take my thousand.

The potentially predictive process *par excellence* is *simulation*. To predict whether I will take my thousand, make a replica of me, put my replica in a replica of my predicament, and see whether my replica takes *his* thousand. And whether or not anybody actually makes a

2. And perhaps not an infallible way. See David Lewis, "The Paradoxes of Time Travel," *American Philosophical Quarterly* 13 (1976): 145-152.

3. That is noted by Nozick, "Newcomb's Problem," p. 132, and I have not seen it disputed.

prediction about me by observing my replica, still my replica's decision is a potentially predictive process with respect to mine. Disregarding predictive processes other than simulation, if such there be, we have this special case of (3″):

(3‴) I will get my million if and only if my replica does not take his thousand.

There are replicas and replicas. Some are the same sort of thing that I am, others are less so. A flesh-and-blood duplicate made by copying me atom for atom would be one good sort of replica. A working scale model of me, smaller perhaps by a ratio of 1:148, also might serve. So might a pattern of bits in a computer, or beads on an abacus, or marks on paper, or neuron firings in a brain, even though these things are unlike me and replicate me only by way of some complicated isomorphism.

Also, some replicas are more reliable than others. There may be grounds for greater or lesser degrees of confidence that my replica and I will decide alike in the matter of the thousand. A replica that matches me perfectly in the respects relevant to my decision (whether duplicate or isomorph) will have more predictive power than a less perfect replica; but even a poor replica may have some significant degree of predictive power.

As Newcomb's Problem is usually told, the predictive process involved is extremely reliable. But that is inessential. The disagreement between conceptions of rationality that gives the problem its interest arises even when the reliability of the process, as estimated by the agent, is quite poor—indeed, even when the agent judges that the predictive process will do little better than chance. More precisely, define *average estimated reliability* as the average of (A) the agent's conditional degree of belief that the predictive process will predict correctly, given that he takes his thousand, and (B) his conditional degree of belief that the process will predict correctly, given that he does not take his thousand. (When the predictive process is a simulation, for instance, we have the average of two conditional degrees of belief that the agent and his replica will decide alike.) Let r be the ratio of the value of the thousand to the value of the million: .001

if value is proportional to money, perhaps somewhat more under diminishing marginal value. We have a disagreement between two conceptions of rationality if and only if the expected value[4] of taking the thousand is less than that of declining it, which is so if and only if the average estimated reliability exceeds $\frac{(1+r)}{2}$. (That is .5005 if value is proportional to money.) This is not a very high standard of reliability. So there can be a fully problematic case of Newcomb's Problem in which the predictive process consists of simulation by some very imperfect and very unreliable replica.

The most readily available sort of replica of me is simply another person, placed in a replica of my predicament. For instance: you, my fellow prisoner. Most likely you are not a very exact replica of me, and your choice is not a very reliable predictive process for mine.[5] Still, you might well be reliable enough (in my estimation) for a Newcomb Problem.[6] So we have this special case of (3‴):

(3) I will get my million if and only if you do not take your thousand.

Inessential trappings aside, Prisoners' Dilemma is a version of Newcomb's Problem, *quod erat demonstrandum.*

Some who discuss Newcomb's Problem think it is rational to decline the thousand if the predictive process is reliable enough. Their reason is that they believe, justifiably, that those who decline their thousands will probably get their millions. Some who discuss Prisoners' Dilemma

4. As calculated according to the non-causal sort of decision theory presented for instance in Richard Jeffrey, *The Logic of Decision* (New York: McGraw-Hill, 1965).

5. On the other hand, you might be an extremely perfect and reliable replica, as in the Prisoners' Dilemma between twins described by Nozick, "Newcomb's Problem," pp. 130-131.

6. If you do not meet even the low standard of estimated reliability just considered, either because you are unlike me or because you and I alike are apt to choose at random or because the payoffs are such as to set r rather high, then we have a situation with no clash between conceptions of rationality; on *any* conception, it is rational to rat. But even this non-problem might legitimately be called a version of Newcomb's Problem, since it satisfies conditions (1), (2), and (3″).

think it is rational not to rat if the two partners are enough alike.[7] Their reason is that they believe, justifiably, that those who do not rat will probably not be ratted on by their like-thinking partners. These two opinions are one opinion in two guises.

But some—I, for one—who discuss Newcomb's Problem think it is rational to take the thousand no matter how reliable the predictive process may be. Our reason is that one thereby gets a thousand more than he would if he declined, since he would get his million or not regardless of whether he took his thousand. And some—I, for one—who discuss Prisoners' Dilemma think it is rational to rat no matter how much alike the two partners may be, and no matter how certain they may be that they will decide alike. Our reason is that one is better off if he rats than he would be if he didn't, since he would be ratted on or not regardless of whether he ratted. These two opinions also are one.

Some have fended off the lessons of Newcomb's Problem by saying: "Let us not have, or let us not rely on, any intuitions about what is rational in goofball cases so unlike the decision problems of real life." But Prisoners' Dilemmas are deplorably common in real life. They are the most down-to-earth versions of Newcomb's Problem now available.

7. For instance Davis, "Prisoners, Paradox, and Rationality." He considers the case in which the partners are alike because they are both rational; but there is also the case where they are alike because they are given to the same sorts of irrationality.