# *Prisoners' Dilemmas are Newcomb Problems?*

*Ryan Doody*

## ***The Prisoners' Dilemma***

David Lewis argues that the Prisoners' Dilemma *is* a Newcomb Problem ("or rather, two Newcomb Problems side by side, one per prisoner.")

Here is Lewis' example of a Prisoners' Dilemma (which involves winning different sums of money):

| | You rat | You don't rat |
|---|---|---|
| I rat | I get \$1,000<br>You get \$1,000 | I get \$1,001,000<br>you get \$0 |
| I don't rat | I get \$0<br>You get \$1,001,000 | I get \$1,000,000<br>You get \$1,000,000 |

In general, a Prisoners' Dilemma is a game with the following structure:

Player 2

| Player 1 | | $c$ | $d$ |
|---|---|---|---|
| | $C$ | 3, 3 | 0, 4 |
| | $D$ | 4, 0 | 1, 1 |

Player 1 has the following preferences:

$$(D \wedge c) \succ (C \wedge c) \succ (D \wedge d) \succ (C \wedge d)$$
$$4 > 3 > 1 > 0$$

And Player 2 has the following preferences:

$$(C \wedge d) \succ (C \wedge c) \succ (D \wedge d) \succ (D \wedge c)$$
$$4 > 3 > 1 > 0$$

Both players have a *dominant strategy*: a strategy that is guaranteed to result in a better payoff no matter what the other player does. But the result of both players playing their dominant strategy is an outcome that is *Pareto-dominated* by some other.

Player 1: $D$ dominates $C$.

Player 2: $d$ dominates $c$.

Outcome $(C \wedge c)$ pareto-dominates outcome $(D \wedge d)$.

### *Is the Prisoners' Dilemma a Newcomb Problem?*

Lewis think it is. Here's is argument. First, he characterizes the Prisoners' Dilemma as follows.

**The Prisoners' Dilemma**

(1) I am offered \$1,000—take it or leave it.

(2) Perhaps also I will be given \$1,000,000; but whether I will or not is causally independent of what I do now.

(3) I will get my \$1,000,000 if and only if you do not take your \$1,000.

He then points out that the Newcomb Problem is almost identical—it just switches out (3) for (3′):

**The Newcomb Problem**

(1) I am offered \$1,000—take it or leave it.

(2) Perhaps also I will be given \$1,000,000; but whether I will or not is causally independent of what I do now.

(3′) I will get my \$1,000,000 if and only if it is predicted that I do not take my \$1,000.

He then points out that it is inessential to the Newcomb Problem that the prediction be carried out in advance. And so, we could characterize the Newcomb Problem with (1), (2), and:

(3″) I will get my \$1,000,000 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 \$1,000.

Lewis then says that the potentially predictive process *par excellence* is *simulation*. So, imagine that the predictor makes a replica of you in order to figure out what you will do. Then, we have a special case of (3″):

(3‴) I will get my \$1,000,000 if and only if my replica does not take his \$1,000.

And, because the replica needn't be an exact replica (and because the prediction needn't be *that* reliable in order to generate a conflict between EDT and CDT), we have a special case of (3‴):

(3) I will get my \$1,000,000 if and only if you do not take your \$1,000.

But (1), (2), and (3) are how we characterized the Prisoners' Dilemma.