Paradoxes—Reading group 4

Chapter 4: Acting rationally

In this chapter, Sainsbury considers two *paradoxes of rational action*: 'Newcomb's paradox', and the 'prisoner's dilemma'. These are central in the philosophy of probability and decision theory, as well as in other disciplines, such as the foundations of economics.

Newcomb's paradox

Sainsbury begins by presenting Newcomb's paradox.[1] Here's how he puts the problem:

> *You are confronted with a choice. There are two boxes before you, A and B. You may either open both boxes, or else just open B. You may keep what is inside any box you open, but you may not keep what is inside any box you do not open. The background is this.*
>
> *A very powerful being, who has been invariably accurate in his predictions about your behavior in the past, has already acted in the following way:*
>
> - *He has put $1,000 in box A.*
> - *If he has predicted that you will open just box B, he has in addition put $1,000,000 in box B.*
> - *If he has predicted that you will open both boxes, he has put nothing in box B.*
>
> *The paradox consists in the fact that there appears to be a decisive argument for the view that the most rational thing to do is to open both boxes; and also a decisive argument for the view that the most rational thing to do is to open just box B. The arguments commend incompatible courses of action: if you take both boxes, you cannot also take just box B. Putting the arguments together entails the overall conclusion that taking both boxes is the most rational thing and also not the most rational thing. This is unacceptable, yet the arguments from which it derives are apparently acceptable.* (Sainsbury, p. 69)

What's the argument for opening both boxes? Sainsbury articulates this as follows:

[1] The first philosophical paper on this was: Robert Nozick, "Newcomb's Problem and Two Principles of Choice", in N. Rescher *et al.* (eds.), *Essays in Honor of Carl G. Hempel*, 1969.

> *The powerful being—let us call him the Predictor—has already acted. Either he has put money in both boxes or he has put money in just box A. In the first case, by opening both boxes you will win \$1,001,000. In the second case, by opening both boxes you will at least win \$1,000, which is better than nothing. By contrast, if you were to open just box B, you would win just \$1,000,000 on the first assumption (i.e. that the Predictor has put money in both boxes) and nothing on the second assumption (i.e. that the Predictor has put money just in box A). In either case, you would be \$1,000 worse off than had you opened both boxes. So opening both boxes is the best thing to do.* (Sainsbury, p. 69)

And here's the argument for opening just box B:

> *Since the Predictor has always been right in his previous predictions, you have every reason for thinking that he will be right in this one. So you have every reason to think that if you were to open both boxes, the Predictor would have predicted this and so would have left box B empty. So you have every reason to think that it would not be best to open both boxes. Likewise, you have every reason to think that if you choose to open just box B, the Predictor will have predicted this, and so will have put \$1,000,000 inside. Imagine a third party, who knows all the facts. He will bet heavily that if you open just box B you will win \$1,000,000. He will bet heavily that if you open both boxes you will get only \$1,000. You have to agree that his bets are rational. So it must be rational for you to open just box B.* (Sainsbury, p. 70)

Maximising expected utility

The second argument above is based upon the principle that it is rational to act so as to *maximise expected utility*—call this principle MEU. In light of our knowledge of the previous success of the Predictor, MEU recommends opening just box B. Why? We can compute this quantitatively as follows.

In general, the expected utility $EU(A)$ of some action A is calculated as follows, where O_i is some outcome, $\text{prob}(O_i/A)$ is the probability of outcome O_i given action A, and $U(O_i)$ is the expected utility of outcome O_i:[2]

[2]Here, A is general—it doesn't necessarily refer to opening just box A in the above example.

$$EU\left(A\right) = \text{prob}\left(\frac{O_1}{A}\right) \cdot U\left(O_1\right) + \text{prob}\left(\frac{O_2}{A}\right) \cdot U\left(O_2\right) + \ldots$$

In our example, we can now compute the expected utilities of opening box B, or both boxes ($A\&B$), as follows:

$$\begin{aligned} EU\left(B\right) &= \text{prob}\left(\frac{B \text{ empty}}{B}\right) \cdot U\left(B \text{ empty}\right) + \text{prob}\left(\frac{B \text{ full}}{B}\right) \cdot U\left(B \text{ full}\right) \\ &= (1-h) \cdot 0 + h \cdot 1,000,000. \end{aligned}$$

$$\begin{aligned} EU\left(A\&B\right) &= \text{prob}\left(\frac{B \text{ empty}}{A\&B}\right) \cdot U\left(B \text{ empty, } A \text{ full}\right) + \text{prob}\left(\frac{B \text{ full}}{A\&B}\right) \cdot U\left(B \text{ full, } A \text{ full}\right) \\ &= h \cdot 1,000 + (1-h) \cdot 1,001,000. \end{aligned}$$

h is a measure of how much we trust our past evidence about the Predictor: assuming that we are inductive agents, and so trust our past evidence about the Predictor, this should be a number close to one. Say we put $h = 0.9$. Then we have

$$\begin{aligned} EU\left(B\right) &= 900,000 \\ EU\left(A\&B\right) &= 101,100 \end{aligned}$$

—a nearly nine-fold advantage to taking just box B! So MEU recommends opening box B. But this isn't yet sufficient to resolve the paradox, because we need to understand what was wrong with the argument for opening both boxes.

The dominance principle

The argument for opening *both* boxes ($A\&B$) follows from the *dominance principle* (DP): (p. 74)

> *According to DP, it is rational to perform an action α if it satisfies the following two conditions:*
>
> - *Whatever else may happen, doing α will result in your being no worse off than doing any of the other things open to you.*
> - *There is at least one possible outcome in which your having done α makes you better off than you would be had you done any of the other things open to you.*

DP recommends the *first* strategy: opening *both* boxes. As Sainsbury says, "DP has commonsensical appeal. If you follow it you will act in such a way that nothing else you could do would have resulted in your faring better, except by running the risk of your faring worse."

> **One way to understand Newcomb's paradox is precisely as the tension between the principles MEU and DP!**

JR claim: The two principles can be rendered compatible, in the following way. Note that if $h = 0.5$, so that we have no knowledge of whether the Predictor is indeed a perfectly accurate predictor or not ($h = 0.5$ corresponds to complete agnosticism about the Predictor's predictive abilities), then the rational thing to do is to open both boxes (just work it out, using MEU). This seems to account for the plausibility of DP—for the advocate of DP *decouples* what the Predictor has done from my decision now.[3]

The prisoner's dilemma

Sainsbury now moves on to consider a different conundrum of rational action: *the prisoner's dilemma*. This is a very well-known example from game theory, which purports to show why two completely rational individuals might not cooperate, even if it appears that it is in their best interests to do so. Here's how Sainsbury puts the problem:

> *You and I have been arrested for drug running and placed in separate cells. Each of us learns, through his own attorney, that the district attorney has resolved as follows (and we have every reason to trust this information):*
>
> 1. *If we both remain silent, the district attorney will have to drop the drug-running charge for lack of evidence, and will instead charge us with the much more minor offence of possessing dangerous weapons. We would then each get a year in jail.*
> 2. *If we both confess, we shall both get five years in jail.*
> 3. *If one remains silent and the other confesses, the one who confesses will get off scot-free (for turning state's evidence), and the other will go to jail for ten years.*
> 4. *The other prisoner is also being told all of (1)-(4).*

[3] In this sense, Sainsbury seems correct when he writes "one appropriate restriction on DP is that it can be used only when there is no relevant difference in the probability of the various possible outcomes." (p. 81)

> *How is it rational to act? We build into the story the following further features:*
>
> 5. *Each prisoner is concerned only with getting the smallest sentence for himself.*
> 6. *Neither has any information about the likely behavior of the other, except that (5) holds of him and that he is a rational agent.*
>
> *There is an obvious line of reasoning in favor of confessing. It is simply that whatever you do, I shall do better to confess. For if you remain silent and I confess, I shall get what I most want, no sentence at all; whereas if you confess, then I shall do much better by confessing too (five years) than by remaining silent (ten years).* (Sainsbury, pp. 82-3)

But the problematic conclusion here is obvious: "By acting supposedly rationally, we shall, it seems, secure for ourselves an outcome that is worse for both of us than what we could achieve."[4] (p. 84) And note that one seems to reach this conclusion on both DP and MEU—so what has gone wrong?

There are some parallels between Newcomb's paradox and the prisoner's dilemma: one can see this if one treats the other prisoner as a Predictor.[5]

[4] A lesson for neoliberalism? The prisoner's dilemma experiment has actually been carried out: in reality, humans display a systemic bias towards cooperative behaviour in this and similar games.

[5] For a paper-length discussion of this, see: David Lewis, "Prisoner's Dilemma Is a Newcomb Problem", Philosophy and Public Affairs 8, pp. 235-240, 1979.