'Why Ain'cha Rich?'

DAVID LEWIS

PRINCETON UNIVERSITY

Some think that in (a suitable version of) Newcomb's problem, it is rational to take only one box. These one-boxers think of the situation as a choice between a million and a thousand. They are convinced by indicative conditionals: if I take one box I will be a millionaire, but if I take both boxes I will not. Their conception of rationality may be called *V-rationality*; they deem it rational to maximize V, that being a kind of expected utility defined in entirely non-causal terms. Their decision theory is that of Jeffrey [2].

Others, and I for one, think it rational to take both boxes. We two-boxers think that whether the million already awaits us or not, we have no choice between taking it and leaving it. We are convinced by counterfactual conditionals: If I took only one box, I would be poorer by a thousand than I will be after taking both. (We distinguish nornal from back-tracking counterfactuals, perhaps as in [4], and are persuaded only by the former.) Our conception of rationality is *U-rationality*; we favor maximizing U, a kind of expected utility defined in terms of causal dependence as well as credence and value. Our decision theory is that of Gibbard and Harper [1], or something similar.

The one-boxers sometimes taunt us: if you're so smart, why ain'cha rich? They have their millions and we have our thousands, and they think this goes to show the error of our ways. They think we are not rich because we have irrationally chosen not to have our millions.

We reply that we never were given any choice about whether to have a million. When we made our choices, there were no millions to be had. The reason why we are not rich is that the riches were reserved for the irrational. In the words of Gibbard and Harper [1],

> we take the moral. . . to be something else: if someone is very good at predicting behavior and rewards predicted irrationality richly, then irrationality will be richly rewarded.

Rationality will not.

(Let us say that irrationality will be richly *pre*-rewarded. That cancels the suggestion, which of course we do not intend, that the irrationality causes the "reward".)

What is the status of this moral? Is it

(1) one more piece of two-boxist doctrine that one-boxers may consistently deny?

Or is it

(2) common ground, something that ought to be uncontroversial?

Can all agree that no matter whether true rationality is *V*-rationality or *U*-rationality—indeed, even if it is some undreamt-of third sort of rationality—still the predictor can see to it, if he is so inclined and good enough at predicting, that irrationality is richly pre-rewarded and the smart ain't rich?

I regret to say that alternative (1) appears to be correct. At any rate, the obvious way to argue for alternative (2) is a failure. So it's a standoff. We may consistently go on thinking that it proves nothing that the one-boxers are richly pre-rewarded and we are not. But they may consistently go on thinking otherwise. For it is impossible, on their conception of rationality, to be sure at the time of choice that the irrational choice will, and the rational choice will not, be richly pre-rewarded. *V*-irrationality cannot be richly pre-rewarded, unless by surprise. (And we did not plead surprise. We knew what to expect.) The expectation that only one choice will be richly pre-rewarded—richly enough to outweigh other considerations—is enough to make that choice *V*-rational.

Try to imagine that the predictor in Newcomb's problem changes sides. Hitherto, his announced policy has been to pre-reward *U*-irrationality. He has left a million just when he predicted that the subject was going to make the *U*-irrational choice of taking only one box. But from now on he will create a new kind of problem. His announced policy henceforth will be to pre-reward *V*-irrationality. He will leave a million just when he predicts that the subject is going to make the *V*-irrational choice, whichever that is. (If neither choice in the new problem is *V*-irrational, he will never leave a million.) He is just as good at predicting as he was before; and he sees to it that the subject is convinced (or close to convinced) that a correct prediction has been made. Now that someone is very good at predicting behavior and rewards predicted *V*-irrationality richly, it seems that *V*-irrationality will be richly rewarded (and not by surprise). Why not?

Answer: because the story just told is self-contradictory. The new problem, unlike the Newcomb problem, is impossible. The predictor announces, convincingly, that he will pre-reward a certain choice. Thereby he makes the choice V-rational. But the choice to be thus made V-rational is the V-irrational one, whichever that is. That is, it is whichever one is V-irrational given, *inter alia*, his announcement. So the story says that the predictor makes it the case that one and the same choice is V-rational and V-irrational. Whatever he may do, he cannot do that.

To reach a *reductio* against the supposition that the new problem is possible, let us ask which choice (if either) is V-irrational in the new problem. Let C be the subject's credence function at the time of deliberation; let M be the proposition that the predictor has left a million; let A_1 be the proposition that the subject takes only one box, declining his thousand; and let A_2 be the proposition that he takes both boxes. Let the utility of the payoffs be measured by money. Then we have three cases.

Case 1: $C(M/A_1) < C(M/A_2) + .001$. Then taking only one box is V-irrational, and taking both boxes is not. But if so, $C(M/A_1) \approx 1$ and $C(M/A_2) \approx 0$. Contradiction.

Case 2: $C(M/A_1) = C(M/A_2) + .001$. Then the choices are tied, so neither is V-irrational. But if so, $C(M/A_1) \approx 0$ and $C(M/A_2) \approx 0$. Contradiction.

Case 3: $C(M/A_1) > C(MA_2) + .001$. Then taking both boxes is V-irrational, and taking one box is not. But if so, $C(M/A_1) \approx 0$ and $C(M/A_2) \approx 1$. Contradiction.

All three cases are impossible. Yet if the new problem is possible, one of the three must hold. The new problem is impossible, *quod erat demonstrandum*.

In discussion it has been suggested that the new problem *is* possible; that in the new problem it is V-rational to take both boxes and V-irrational to take only one (so that in this problem V-rationality and U-rationality agree); that the one-boxer must concede that on his view also, predicted irrationality may be richly rewarded, and not by surprise; and that my *reductio* fails because in Case 1, the correct case on this proposal, $C(A_1) = 0$ and $C(M/A_1)$ is undefined. The V-rational subject is imagined to deliberate as follows:

> I'm going to do the V-rational thing. That makes it almost certain that there's no million for me. Then the V-rational thing is to take both boxes and get my thousand, and that is what I'll do.

I object that if the subject is still deliberating, then he is not yet sure (even implicitly) what he will do. If he is, for instance if $C(A_1) = 0$ and $C(A_2) = 1$, then his decision problem collapses as described in Jeffrey [3]; in which case the distinction between V-rational and V-irrational actions in his situation is undefined. But if he is not sure (even implicitly) what he will do, he must be lacking in self-knowledge. He must be uncertain either about his credences, about his utilities, or about the standards of rationality (or irrationality) to which he is going to conform. In this case the third sort of lack of self-knowledge is most plausible. It is therefore inadmissible to suppose him to be deliberating, and yet suppose him already to be certain that he will do the V-rational thing.[1]

REFERENCES

[1] Allan Gibbard and William Harper, "Counterfactuals and Two Kinds of Expected Utility", in C. A. Hooker, J. J. Leach, and E. F. McClennen, eds., *Foundations and Applications of Decision Theory*, Volume 1 (Dordrecht, Holland: D. Reidel, 1978), pp. 125-62.

[2] Richard C. Jeffrey, *The Logic of Decision* (New York: McGraw-Hill, 1965).

[3] ———, "A Note on the Kinematics of Preference", *Erkenntnis* 11(1977), pp. 135-41.

[4] David Lewis, "Counterfactual Dependence and Time's Arrow", *Noûs* 13(1979), 455-76.

NOTES

[1]This paper is based on a talk given at a conference on Conditional Expected Utility given at the University of Pittsburgh in November 1978. I thank Paul Benacerraf, Jane Heal, Calvin Normore, and Robert Stalnaker for valuable discussion.