Two-boxing is irrational

I

Philosophers debate whether one-boxing or two-boxing is the rational act in a Newcomb situation. The Newcomb set-up (Nozick 1997) is one in which there are two boxes, one opaque, one transparent. There is a (nearly) infallible predictor who has put a million dollars or nothing in the opaque box. The transparent box contains a thousand dollars. You have a choice of taking both boxes (two-boxing) or just the opaque box (one-boxing). The predictor (whom you know to be (nearly) infallible) has put nothing in the opaque box if he thinks you will two-box and a million if he thinks you will one-box. So your credence in the proposition that you will get a million conditional on one-boxing is very high and your credence in the proposition that you will get just a thousand conditional on two-boxing is also very high. However, you know that when you act the million will either already be in the opaque box or not. So you know that it is certain that two-boxing is the choice which *will get you more money than the alternative would no matter what.* You know that if you decide to one-box you will have to acknowledge 'I would get a thousand dollars more if I were to two-box' and that if you decide to two-box you will then be able to say 'I would get a thousand dollars less if I were to one-box'. That is why it seems to many (perhaps most) philosophers who have considered the question that two-boxing is the only rational option (though choosing to two-box brings the 'bad news' that you will almost certainly only get a thousand).

I shall argue that one-boxing is the only rational choice, given that your utilities vary linearly with the monetary values in the range of the problem.

This is so because there is no intelligible aim by reference to which you can justify the choice of two-boxing over one-boxing once you have come to think that you will two-box (whereas there is such an aim by reference to which you can justify one-boxing). The only aim by which the agent in the Newcomb situation can justify his two-boxing (once he has decided to two-box) is the *subjunctively described* aim of 'getting more than I *would* if I *were* to one-box'. But such a *subjunctively described* aim can justify an action only if it can be seen as generating, in conjunction with the agent's beliefs, an *indicatively describable* aim which justifies the action. In the case of the Newcomb agent, however, the aim of 'getting more than I would if I were to one-box' cannot be seen in this way.

My argument then appeals to the following principle:

(RATGEN) An aim whose content must be specified using a subjunctive conditional construction can *only* be intelligibly cited as justification for an action if it can be understood as rationally generating (in conjunction with the agent's beliefs) an aim which *can* be otherwise specified and which the action fulfils.

I shall come back to the meaning and justification for this principle in Section IV.

The argument I am going to give, in summary, then, is the following (in which by a 'committed' performer of an action of type A, I mean someone who believes that he is going to A because he has decided to, but has not yet A'd). In the Newcomb situation:

(1) Even the committed two-boxer cannot deny that he will get more than nothing if he one-boxes (all the evidence points to $M). So:

(2) He cannot justify his choice of two-boxing over one-boxing just on the ground that if he two-boxes he will get more than nothing

(3) But he cannot justify his choice of two-boxing over one-boxing on the *sole* ground that if he two-boxes then, no matter what he gets, he will get more than the amount he *would* get if he were to one-box (this offends against RATGEN)

(4) He cannot justify this choice in any other way

(5) If a committed agent cannot justify his choice, his acting cannot be rational. So:

(6) Two-boxing cannot be rational for a committed two-boxer. So:

(7) Two boxing is irrational.

II

We can think of the agent's participation in the Newcomb game as divided into two stages: before he forms a view on what he will do (when he is a merely 'prospective' but uncommitted one-boxer/two-boxer), and afterwards. Before he has a belief about whether he will one-box or two-box his credence in the proposition that he will get a million conditional on one-boxing is very high and his credence in the proposition that he will get a million conditional on two-boxing is very low. He therefore believes, entirely rationally, that almost certainly if he one-boxes he will get more than he will if he two-boxes. Since he assigns a high probability to getting $M conditional on one-boxing he also seems to have every reason to believe, at this stage, in the absence of a settled view about what he will in fact do, that almost certainly, if he *were* to one-box he *would* get a million. That is, it seems entirely reasonable for him to assert, in his uncertainty about what he will in fact do, 'If were to one-box (as I may or may not) I would get a $M' (his use of the subjunctive need not signal belief that the antecedent is false). On the other

hand, it seems that at this stage he can have no reason to assert 'If I were to one-box (as I may or may not) I would get nothing' (which, in fact, given his uncertainty about what he is going to do, it would be quite bizarre for him to assert). Again, given the high probability he assigns to getting just $K conditional on two-boxing, it seems that, at this stage, he has very good reason to assert 'If I were to two-box (as I may or may not), I would only get $K', and no reason to assert 'If I were to two-box I would get $M+K', which it would be quite bizarre for him to assert in his state of uncertainty about what he is going to do. Hence, it seems that at this stage, while uncertain about what he is going to do, he can reasonably affirm 'If I were to one box I would get more ($M) than I would if I were to two-box ($K)'. Though, of course, he knows that if he one-boxes he will be leaving on the table a thousand he would get if he were to two-box.

Gibbard and Harper (1978), however, say that of the two forward-looking subjunctive conditionals (i) 'If I were to one-box I would get a million' and (ii) 'If I were to two-box I would get just a thousand', utterable by the agent prior to his decision, one – the one whose antecedent is actually contrary to (future) fact – is false. This is because it must be interpreted in a non-backtracking way. It follows also that one of the two backward-looking subjunctive conditionals, utterable by the agent prior to his decision, (*) 'If I were to one-box the predictor would have predicted that I would' and (ii*) 'If I were to two-box the predictor would have predicted that I would', one – the one whose antecedent is actually contrary to future fact – is false, even if the predictor is, to use their word, 'infallible', that is, in fact always predicts correctly. How plausible this is as a claim about the English subjunctive conditional is dubious. Certainly, backward-looking counterfactuals can often be interpreted in a backtracking manner: 'If I had got up at 6am

(rather than 8am as I did) I would have gone to bed at 10pm (rather than midnight as I did)' is easily interpretable as expressing a truth in the mouth of someone known to be obsessive about getting his full eight hours (Bach 1987). But I do not wish to debate this. I think that Gibbard and Harper can simply *stipulate* that the conditional that defines U-maximization conforms to their semantic analysis. I wish merely to observe that it is only by interpreting in a non-backtracking way the subjunctive conditionals with antecedents that are actually contrary to future fact utterable by the agent prior to his decision that Gibbard and Harper can say that not *all* of these subjunctive conditionals – (i) and (ii) and (i*) and (ii*) – are true, as uttered by the agent still uncertain about what he will do. Consequently, it is only by doing this that they can say that if the undecided agent says 'if I were to one-box I would get more than I would if I were to two-box' he is saying something false.

III

Now consider the aims that might motivate the agent to one-box or two-box. In considering a course of action I can ask 'What *will* be my aim in acting in this way?' and after I have come to a decision to act I can ask 'What *is* my aim in acting in this way?' So as I try to decide whether to send my paper off to journal *A* or journal *B* I can ask 'What will be my aim in sending the paper to *A*?' (the answer might be: 'to get a paper in a top-ranked journal, albeit with a long delay before publication') and, having decided on journal *A*, I can ask, 'What is my aim in sending this off to *A*?' So the undecided Newcomb agent can ask, 'What *will* be my aim in one-boxing/two-boxing?' and having decided on one of these he can ask, 'What *is* my aim in one-boxing/two-boxing?'

In the case of the prospective one-boxer the answer is easy. His answer to 'What *will* be my aim in one-boxing?' must be: to do what *will* get me the most money. He believes, entirely rationally, that he *will* get more money if he one-boxes than he *will* if he two-boxes.

What can the answer of the prospective two-boxer be to the question 'What *will* my aim be in two-boxing?' It cannot be: 'to do what *will* get me the most money', for he is very confident that he will get more money ($M) if he one-boxes than he will if he two-boxes (K). Nor can it be: to get at least $K, which he must believe, is most likely all he will achieve by two-boxing. For, of course, he thinks it very probable that he will get $K, in fact $M, if he one-boxes. But the Newcomb agent knows that two-boxing dominates, so that whatever is in the opaque box, if he does one-box he *would* have got more if he had two-boxed instead. So if he decides to two-box his aim must be the subjunctively specified aim: 'to get more money than I *would* if I *were* to one-box' or 'to perform *the* action which will get me more money than the alternative *would*, no matter what'. And on Gibbard and Harper's understanding of the subjunctive conditionals utterable by the agent before he has decided what to do, he can indeed say, correctly, in justification of two-boxing: 'two-boxing is *the* action which will get me more money than the only alternative (one-boxing) *would*, no matter what is in the opaque box'.

IV

This should be entirely unsurprising. It accords with Lewis's view that being guided by indicative conditionals leads to one-boxing, and that two-boxing can *only* be motivated by attending to subjunctive (counterfactual) conditionals. He writes: 'One-boxers are convinced by indicative conditionals: if I take one box I will be a millionaire, but if I take both boxes I will not…. We [two-boxers] … 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' (Lewis 1981: 377).

I shall now argue for the principle RATGEN, and hence argue that an action is rationally intelligible *only* if it is rational in the light of an aim which can be specified without the use of the subjunctive conditional construction – though that aim may itself only be generated by an antecedent, only-subjunctively-specifiable, aim. A subjunctively described aim can justify an action only if it can be seen as generating, in conjunction with the agent's beliefs, an indicatively describable aim which justifies the action.

What do I mean by an aim justifying an action? An aim is that *P*, that the world be a certain way, e.g., that I will retain healthy gums. I perform an action by doing something, e.g., flossing. An aim that *P* justifies an action *A* if I can reason as follows:

(A) If I perform action *A* I will make it the case that *P*

I want it to be the case that *P*

So (ceteris paribus) I should perform action *A*.

For example,

If I floss I will make it the case that I will retain healthy gums

I want it to be the case that I will retain healthy gums

So (ceteris paribus) I should floss.

If what replaces '*P*', as in this example, is wholly indicative this is a justification of *A*-ing by reference to an indicatively described aim. Another example (refer to the scenario below) is the following.

If I go for the $2M deal I will make it the case that I will have more than $M

I want it to be the case that I will have more than $M

So (ceteris paribus) I should go for the $2M deal.

If what replaces '*P*' contains a subjunctive construction this is a justification of *A*-ing by reference to a subjunctively described aim. An example (refer again to the scenario below) is the following.

If I go for the $2M deal I will make it the case that I will have more money than my rival would have if he were negotiating instead of me

I want it to be the case that I will have more money than my rival would have if he were negotiating instead of me

So (ceteris paribus) I should go for the $2M deal.

I can certainly act to achieve a goal I can only specify using a subjunctive conditional, as the following scenario makes clear. I am obsessed with doing better financially than my business rival. In fact, I not only want to do better than he does, I want to do better in any business situation than he *would* have done if he had been involved in my place. In any business deal the only question for me is 'If I were to act in this way would I do better than my rival would if he were doing the negotiating?' I am currently involved in a transaction in which I can make $900K easily, two million honestly but with effort or three million dishonestly. I believe that if my rival were involved instead he would only make a million. My monomaniacal aim of doing better than my rival would (in Lewisean terms, of outperforming even his *non-actual* counterparts in close worlds in which they are involved in the negotiation), together with this belief, rationally generates the aim of making more than a million. Thus it decisively rules out the easy $900K deal. But it does not generate any indicatively describable aim which justifies going for the $3M deal rather than the $2M deal or vice versa. Thus, in

the light solely of the aim of doing better than my rival would neither action is justifiable over the other.

This scenario makes it clear how an aim specifiable only with a subjunctive conditional construction can justify an action, but it also makes it clear, I think, how it can do so only by generating, in conjunction with the agent's beliefs, an indicatively describable aim – in this case making more than $M – which justifies the action. That is what RATGEN says.

Now let us return to the Newcomb scenario. In this case two-boxing can be justified only by reference to the subjunctively described aim of getting more money than I would if I were to one-box, of getting more money than my non-actual one-boxing counterpart.[i] But given the only relevant belief I have once I have decided to two-box, i.e., that since the predictor is good at his job there is nothing in the opaque box, there is no indicatively describable aim in conjunction with my beliefs by reference to which two-boxing can be justified over one-boxing in the way in which going for the arduous $2M deal can be justified over going for the easy $900K deal in the business transaction. Since I believe, once I have decided to two-box, that the opaque box is empty (because I believe that the predictor is good at his job), I then believe that the amount I would get if I were to one-box is $0. This is comparable to the belief that my rival would make $M if he were involved in the business transaction. But the non-subjunctively describable aim of (actually) getting more than $0 cannot justify two-boxing over one-boxing, as the aim of making more than $M justifies going for the $2M or $3M deal rather than the easy and honest $900K deal, for I still cannot deny, what I have believed all along, that I *will* get more than nothing if I one-box and hence cannot deny that I could achieve *this* aim by

one-boxing.[ii] For when I decided to two-box, and so came to believe (assign a probability of one to the proposition) that that was what I was going to do, the probability of getting a million *conditional* on one-boxing did not become less for me; it just became undefined.[iii] So once I have decided to two-box I cannot cite a non-subjunctively specifiable aim which the one-boxer cannot also cite to justify his action.

So, given RATGEN, there can be no justification for two-boxing.

To contest this argument the defender of the rationality of two-boxing must reject RATGEN. He must say that the two-boxer can justify his action simply by saying, 'My sole aim is to get more than I *would* if I were to one-box. I cannot deny that I will get a million if I one-box and that most probably I will only get a thousand by two-boxing. I cannot specify any indicatively describable aim that I will achieve by two-boxing and will not achieve by one-boxing. Nonetheless, by two-boxing I will certainly fulfil *this* subjunctively-specifiable aim, and that is all I care about'.

V

The one-boxer has no problem in citing a non-subjunctively specifiable aim which the two-boxer cannot cite to justify his action. His aim is to get a million, to make it the case that he *will* have a million, which is the most he believes he will get. If it is said to him, 'You will get $M+K if you two-box' his answer must be, 'I have believed all along that I will most likely get just $K if I two-box and my deciding to one-box has given me no reason to deny that' (the probability of his getting just a thousand *conditional* on two-boxing did not become less for him when he came to the decision to one-box, it just became undefined). To the challenge, 'Given that you are going to one-box, if you were to two-box you would get a $M+K' his answer can be, 'Indeed, if the subjunctive

conditional is interpreted in a non-backtracking way it is true that I *would* if I *were* to – but if I *do* I *will* most likely get just $K'. And if the challenge is put without counterfactual language, 'Since you are in fact going to one-box if you two-box you will get $M+K' his answer must be that this is simply unintelligible as an explanation of why he should prefer two-boxing to one-boxing. He cannot regard it as rational to aim to get $M+K by two-boxing. For he is virtually certain that he will *not* get $M+K, that is, he assigns an *unconditional* probability as close to zero as you like to getting $M+K. However, he is virtually certain that he will get $M conditional on one-boxing (that is, he assigns a probability close to one to getting $M conditional on one-boxing). So to regard it as rational to aim to get $M+K by two-boxing the one-boxer must regard it as rational to pass up the near certainty of something he values a great deal for a small chance of something he does not value a great deal more. So the one-boxer cannot regard it as rational to aim to get $M+K by two-boxing. (Of course, the two-boxer does *not* aim to get $M+K; he expects to get $K. His aim is *only* to get more than he would by one-boxing, which he is confident he will do whatever he gets.) Hence the one-boxer cannot regard it as rational to aim to get *at least* a million by two-boxing since he knows that he will do that by two-boxing *just in case* he gets $M+K. So whilst 'to get (at least) a million' is an aim the committed one-boxer can regard as a reason for one-boxing, it is not an aim he can regard as equally giving him a reason for two-boxing. On the other hand, the committed two-boxer cannot deny that he will get more than nothing if he one-boxes and cannot deny that he could aim to get more than nothing by one-boxing. So he cannot say that whilst getting more than nothing is an aim he can achieve by two-boxing it is not an aim he could achieve by one-boxing.

So the committed two-boxer has no reason to two-box which is not an equally good reason to one-box. Or, if he has, this is an aim only specifiable by a subjunctive – contrary to RATGEN. However, the committed one-boxer *does* have a reason to one-box which is not an equally good reason to two-box – to get a million. So, given RATGEN, whatever the agent commits to doing, he has no reason to two-box which is not an equally good reason to one-box. Two-boxing is rationally indefensible.

VI

It may be illuminating to contrast the Newcomb Problem with another, Hume's Farmer's Dilemma. Here idleness is rationally justifiable because it can be justified by an indicatively describable aim:

> Your corn is ripe today; mine will be so tomorrow. 'Tis profitable for us both, that I shou'd labour with you to-day, and that you shou'd aid me to-morrow. I have no kindness for you, and know you have as little for me. I will not, therefore, take any pains on your account; and should I labour with you upon my own account, in expectation of a return, I know I shou'd be disappointed, and that I shou'd in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security. (Hume 1740: 520–521)

I know that whatever I do the other farmer will not help me out tomorrow, so the only rationally selfish thing for me to do is not to help him out today. My aim in doing so is to make it the case that I will spend today in leisure. No appeal to a subjunctively specified aim is necessary. The Farmer's Dilemma is thus not a dilemma but merely an unfortunate situation in which cooperation is irrational.

Lewis is right that two-boxing, if justifiable at all, can be justified only by reference to (non-backtracking) subjunctive/counterfactual conditionals. But it cannot be justified solely by reference to such conditionals. So it cannot be justified.

References

Bach, K. 1987 'Newcomb's Problem: The $1,000,000 solution', *Canadian Journal of Philosophy* 17, pp. 409 - 425.

Gibbard, A. and Harper, W.L. 1978 'Counterfactuals and Two Kinds of Expected Utility', in *Foundations and Applications* of *Decision Theory,* ed. C. A. Hooker, J. I. Leach, and E. F. McClennen, 2 vols., vol. 1: *Theoretical Foundations,* The University of Western Ontario Series in Philosophy of Science 13, Dordrecht, Holland, D. Reidel, pp. 129-52.

Hume, David, 1740 [1888 1976], *A Treatise of Human Nature,* L. A. Selby-Bigge (ed.), rev. 2nd. edition P. H. Nidditch (ed.), Oxford: Clarendon Press.

Lewis, D. 1981: 'Why Ain'cha Rich', *Nous* 15, pp. 377- 80.

Nozick, R. 1997: 'Newcomb's Problem and Two Principles of Choice' in *Socratic Puzzles* Cambridge University Press, Cambridge, Mass., pp. 45-73. Reprinted from N. Rescher et al. (edd.) *Essays in Honour of Carl G. Hempel,* D.Reidel Publishing Company, Dordrecht, Holland, 1969, pp. 114-146.

[i] Of course, an agent in the Newcomb scenario might have indicatively describable aims *in addition* to his aim of making the maximum monetary gain. So he might simply wish to take both boxes, to take the money in both, to take all the money there in fact is, as an end in itself. But the problem is of interest only if monetary gain is specified as his only aim. And he is very confident that he *will* gain more if he one boxes (a million) than he *will* if he two-boxes and takes all the money – though, in fact, in doing so he will get more than he *would* have done by one-boxing.

[ii] Since my evidence indicates that the opaque box is empty, I must, of course, also believe, so long as I continues to believe that I will two-box, that I will get $0 if I one-box. Of course, I can do this consistently since I do not believe that I will one-box.

[iii] Of course, if my deciding to two-box did not result in my assigning a value of one to the proposition that I was going two-box, and hence zero to the proposition that I was going to one-box, the conditional probability remains as high as before.