TWO BOXING IS NOT THE RATIONAL OPTION

Harold W. Noonan

Abstract

In the standard Newcomb scenario two-boxing is not the rational act and, in general, in Newcomb-style cases the 'two-boxing' choice is not the rational act. Hence any decision theory which recommends two-boxing is unacceptable.

I

I shall argue that in the standard Newcomb scenario two-boxing is not the rational act and, in general, in Newcomb-style cases the 'two-boxing' choice is not the rational act.[1] Hence any decision theory which recommends two-boxing in the standard Newcomb scenario is unacceptable. I end with a suggestion about why this is so.

My argument that two-boxing is not the rational option is as follows. (1) If two-boxing is the rational option in the standard Newcomb scenario, then in some 'Eganized' cases – the *Psychopath Button* and *Newcomb's Firebomb*—the analogues of two-boxing in the standard Newcomb case, i.e., pressing the 'kill all psychopaths' button and pressing the 'two-box' button, are likewise rational options.[2] (2) But in these cases these acts are not rational options (as Egan says). (3) So two-boxing is not the rational option in the standard Newcomb scenario (as Egan does not say).

The form of the supporting argument for premiss (1) of this argument is taken from Ahmed's 'Push the Button', but I illustrate with a different example and my conclusion is different.[3] As the title of Ahmed's paper indicates, he wishes to argue that the options Egan rejects are the rational ones. I argue for premiss (2), in support of Egan's claims about the cases he describes, by showing that the *only* reason there is for thinking two-boxing the rational option in the standard Newcomb scenario does not carry over to Egan's cases.

II

I begin, then, with the argument for premiss (1).

The standard Newcomb scenario (Fig. 1) is one in which there are two boxes, one opaque, one transparent.

{INSERT FIGURE ONE HERE}

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 either 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 that 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 then 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

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

[2] See Andy Egan, 'Some Counterexamples to Causal Decision Theory', *The Philosophical Review* 216 (2007), pp. 93-114.

[3] See Arif Ahmed, 'Push the Button', *Philosophy of Science* 7 (2012), pp. 386-395.

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 get only a thousand).

Now consider this *first variant* of the standard case (with details from Egan's case of 'Newcomb's Firebomb') (Fig. 2).

{INSERT FIGURE 2 HERE}

There are two boxes, one opaque, one transparent (in this case the opacity has no role to play, but I keep this feature for ease of reference). In the opaque box there is a million dollars. There is also a firebomb. In the transparent box there is $900. There may or may not also be, in a secret compartment, a triggering device connected to the firebomb in the opaque box in such a way that opening the transparent box will detonate the bomb and incinerate the million dollars. You know all this. You are also very, very confident that there is no such a triggering device in the transparent box, giving its presence a credence of less than one in 200 million. You are certain that whether this is so is causally and probabilistic independent of your choosing to two-box (so you have no reason to say 'If I two-box I will incinerate the million'). You know that no predictor has assessed whether you will one-box or two-box and set up the triggering device accordingly to punish greed. Your choice, as before, is to two-box or one-box.

In this case it is absolutely clear that two-boxing is the rational choice. You will almost certainly get an extra nine hundred if you two-box, even though if triggering is in place you will get more by one-boxing than by two-boxing. There is hardly any chance that you will get only $900 if you two-box, so the slight risk you take by two-boxing of losing the million is well worth it (otherwise you might as well stay indoors forever for fear that a time-travelling pterodactyl might swoop down and carry you off if you step outside). And you are almost certain that two-boxing is the option which satisfies the description 'the choice which will get me more money than the only alternative would'.

In short, in this case everyone must agree that two-boxing is the only rational choice.

Now consider a *second variant* (Fig. 3)

{INSERT FIGURE 3 HERE}

There are three boxes: (1) the opaque box in which there is a million dollars and a firebomb, (2) a second, transparent, box in which there is $900 and possibly, in a concealed compartment, a triggering device as before, as well as possibly, in a second concealed compartment a second triggering device and (3) a third, also transparent, box in which there is a thousand dollars, and, possibly, in a concealed compartment a triggering device as in the second box as before, as well as possibly, in a second concealed compartment a second triggering device. You think it is immensely unlikely that there is a triggering device in the first concealed compartment in either of the transparent boxes and certain that whether this is so is causally and probabilistically independent of any choice you make. You do not have the choice of one-boxing. You can either open the opaque box plus the transparent box with $900 in (call this 'two-boxing(900)') or the opaque box plus the transparent box with a thousand dollars in (call this 'two-boxing(K)'). You know that there is a (nearly) infallible predictor. If he believes that you will two-box(K) he has set up triggering devices in the second concealed compartment in *both* the $900 box and the thousand box so that if you two-box *in any* way you will incinerate the million, i.e., in the circumstance in which two-boxing(K) causes the destruction of the million in the opaque box two-boxing(900) would too. But if he believes that you will two-box(900) he has not set up *any* triggering device *in either box* (however, as noted, there is still a small chance that the boxes contain triggering devices anyway (in the first concealed compartments) as in the *first variant*). Your credence in the proposition that there is no triggering device at all in the $900 box conditional on your not choosing to two-box(K) is equal to your credence in the *first variant* in the proposition that there is no triggering device in the transparent box (you only think the predictor nearly infallible, but your credence in the proposition that there is a triggering device in the *first* concealed compartment in the $900 box in the *second variant* is even lower than your credence in that proposition in the *first variant*).

In this situation evidential decision theory recommends two-boxing (900) (given an appropriate assignment of utilities). But two-boxing(K) is unproblematically the rational option if two-boxing is the rational option in the original Newcomb scenario because this *second variant* has *precisely* the relevant structure of a standard Newcomb scenario. Two-boxing(K) is evidence that the predictor has put triggering devices in the two transparent boxes but does not cause him to do so and two-boxing(K) dominates two-boxing(900). Whether the predictor has put triggering devices in the transparent boxes or not you will get $100 more if you two-box(K) than you would if you were to two-box(900). Before you decide how to act you know that if you decide to two-box(900) you will then have to acknowledge that you would get $100 more if you were to two-box(K) and you know that if you decide to two-box(K) you will then be able to say that you would get a hundred dollars less if you were to two-box(900). You know that of the two actions, two-boxing(900) and two-boxing(K), two-boxing(K) is *the* action which will get you more money than the alternative would, no matter what.

It is important to recognise here that in the standard Newcomb case, just as in the *second variant* , the sole aim by reference to which the Newcomb agent can justify his choice of two-boxing is the *subjunctively* specifiable aim of performing the action which will get more money that the alternative *would*, no matter what. It is part of the definition of the Newcomb situation that the Newcomb agent's *sole* aim is monetary gain. Now if you are in the Newcomb scenario you have to accept the indicative conditionals: (a) 'If I take only the opaque box I will leave *some* money (in the transparent box) on the table' and (b) 'If I take both boxes I will leave no money on the table'. So if you have the aim of *leaving no money on the table* you have a justification for two-boxing in the standard Newcomb scenario which you do not have for two-boxing(K) in the second variant (as a referee suggests). But the aim of *leaving no money on the table* is *not* included in the aim of achieving maximum monetary gain. There is no sense of 'getting more money' (becoming richer) in which (c) 'If I two-box I *will* get more money than I *will* if I one-box' is entailed by the conjunction of the two conditionals (a) and (b). All (a) and (b) entail is that if you one-box you will leave some money (the money in the transparent box) on the table that you will not leave on the table if you two-box. But also if you one-box you will take away some money that you will not take away if you two-box. It is true that if you two-box you will get everything it is *possible* to get and so you will get more than you *would* if you were to one-box, just as if you two-box(K) in the *second variant* you get more than you would if you were to two-box(900). But there is no sense of (c) in which it is entailed by (a) and (b). So if your sole aim as Newcomb agent is maximum monetary gain you do not have a *non-subjunctively* specifiable aim which justifies two-boxing anymore than the agent in the *second variant* has a non-subjunctively specifiable aim which justifies two-boxing(K). Of course, the two-boxer in the standard Newcomb case will get everything that is in fact on the table, including what is in the transparent box, and the one-boxer will not, and the Newcomb agent could have that aim (and not care less that in achieving it he will get almost a million dollars less than he will if he one-boxes). But he is standardly defined as aiming solely at maximum monetary gain, which is the only interpretation which makes the case interesting.

Now consider a *mixed scenario* (Fig. 4), which fuses the *first* and *second variants*.
{INSERT FIGURE 4 HERE}
The set-up is the same as in the *second variant* and the credences in the location of triggering devices are the same. But you now have three options: you can now also choose to one-box

(as in the *first variant*). As you know, the predictor has again placed the triggering devices in the transparent boxes if and only if he predicts that you will two-box(K).

In this set-up evidential decision theory recommends that you two-box(900) given the same utilities as in the second variant. But if two-boxing is the rational option in the original Newcomb scenario and hence two-boxing(K) is the rational option in the *second variant*, two-boxing(900) cannot be a rational option in this triple-choice *mixed scenario*. The case for its definite inferiority is the same as in the *second variant*. The addition to the options in the *second variant* of one-boxing cannot promote two-boxing(900) to rationality in the *mixed scenario* if it is irrational in the *second variant*. It is also clear that in this *mixed scenario* one-boxing is not rational. In the *first variant* of the standard case one-boxing is uncontroversially inferior to two-boxing(900), even though if the triggering is in place you will get more if you one-box than you would if you were to two-box(900). The differences between the *first variant* and the *mixed scenario* cannot make the uncontroversially inferior option in the *first variant* (one-boxing) the *best* of the three in the *mixed scenario*. The availability of the additional choice of two-boxing(K) in the *mixed scenario* does nothing to prevent the transfer from the *first variant* to the *mixed scenario* of the argument for preferring two-boxing(900) to one-boxing.[4] It is still the case in the *mixed scenario* that nothing in the choice between two-boxing(900) and one-boxing has any evidential or causal relevance to the presence of triggering devices in the transparent box. One-boxing can be no more rational in this scenario than in the *first variant* given the credences specified, in particular, the very low probability of triggering devices in the $900 box conditional on two-boxing(900). There can be no reason for one-boxing rather than two-boxing(900) in the *mixed scenario* which the standard Newcomb case two-boxer can acknowledge which is not equally a reason for one-boxing in the *first variant*. If triggering is in place you will get more if you one-box in the *mixed scenario* than you would if you were to two-box(900). But the same is true in the *first variant*. In the *mixed scenario,* unlike the *second scenario,* the agent has a way of preventing incineration (if the triggering devices are in place), namely, one-boxing. But the same is true in the *first variant*. There is no reason for one-boxing rather than two-boxing(900) in the *mixed scenario* which is not equally a reason for one-boxing rather than two-boxing(900) in the *first variant*. But one-boxing is definitely inferior to two-boxing(900) in the *first variant,* so it is definitely inferior to two-boxing(900) in the *mixed scenario*. To put it diagrammatically, the following cannot be:

First Variant	*Mixed Scenario*
One-boxing – BAD	One-boxing – GOOD
Two-boxing(900) – GOOD	Two-boxing(900) – BAD
	Two-boxing(K) – BAD[5]

[4] Notwithstanding the fact that an agent in the *mixed scenario* who is persuaded that he will two-box (K) must think 'Since I am going to two-box(K), if I were (as I am not) to one-box, I would do better than I would if I were (as I am not) to two-box(900)'. An agent in the *mixed scenario* who does not consider two-boxing(K) has the same argument for preferring two-boxing(900) to one-boxing as one in the *first variant*; an agent who considers the option of two-boxing(K) cannot cease to have this argument available to him just because he does so, nor even if he comes to believe that he will two-box(K).

[5] Note carefully that this is undeniably so only given the specification of credences in the location of triggering devices given in the text, viz. that your credence that the predictor has placed no triggering device in the $900 box conditional on your choosing to two-box(900) is very high. If in the *mixed scenario,* contrary to the specification of credences given in the text, you assign, before you make a decision what to do, a significant probability to there being a triggering device in the $900 box conditional on your choosing to two-box(900), as you do not in the *second variant* or the *first variant*, then it may be uniquely rational for you to choose to one-box consistently with thinking it rationally required to two-box(K) in *the second variant* and to two-box(900) in *the first variant*.

So unless every option in the *mixed scenario* is irrational,[6] i.e., definitely inferior to another available in the same scenario (in the sense of being the inferior member of a pair from the available choice set), two-boxing(K) must be a rational option in the *mixed scenario* if two-boxing is the rational choice in the standard Newcomb scenario and two-boxing(K) the rational choice in the *second variant*. For then two-boxing(900) is definitely inferior to two-boxing(K) and one-boxing definitely inferior to one-boxing (900).[7]

Now consider the final scenario, which is 'Newcomb's Firebomb' (Fig. 5).

{INSERT FIGURE 5 HERE}

The set-up is the same as previously, so that there is definitely a million in the opaque box, your credences are the same and you know that the predictor has made his predictions and acted accordingly in exactly the same way. But you have only two choices: to one-box or to two-box(K). The option of two-boxing(900) does not exist. In this scenario the evidential decision theorist will endorse one-boxing, consistently with his recommendation in the *mixed scenario*, since his preferred option there (to two-box(900)) no longer exists. The philosopher who endorses two-boxing in the standard Newcomb scenario and two-boxing(K) in the *second variant*, and hence, by the argument above, must regard endorsing two-boxing(K) as a rational choice in the *mixed scenario*, must say that in this final case, too, to two-box(K) is a rational choice, i.e., not definitely inferior to one-boxing, since the only difference between this case and the previous *mixed scenario* is the absence from this case of an option (two-boxing(900)) which was definitely inferior to two-boxing(K) when it was available. Removing a definitely inferior option from a choice-set cannot make inferior an option that was not so previously.

Hence the philosopher who endorses two-boxing as the rational choice in the standard Newcomb scenario must, contrary to Egan's intuitions, endorse two-boxing as rational in this case also. Here is a quick recap of the argument. The *first variant* is one in which everyone agrees one-boxing is crazy. The *second variant* has precisely the relevant structure of a standard Newcomb scenario so one-boxers in the original Newcomb case will say in this case 'two-box(900)' and standard Newcomb case two-boxers will say 'two-box(K)'. The *mixed scenario* is just the *second variant* with the added option of one-boxing. The argument for the inferiority of one-boxing to two-boxing(900) carries over from the *first variant* to the *mixed scenario*, and the Newcomb case two-boxer's argument for the inferiority of two-boxing(900) to two-boxing(K) carries over from the *second variant* to the *mixed scenario*. The final case, Newcomb's Firebomb, omits from the situation in the *mixed scenario* the option of two-boxing(900), which Newcomb case two-boxers must regard as definitely inferior to two-boxing(K) in the *mixed scenario*.

In general the same reasoning obviously requires that if two-boxing is the rational choice in the Newcomb scenario then its analogue is rational in any of the 'Eganized' cases ('The Murder Lesion', 'The Psychopath Button' etc.). To see that this is so for the Psychopath Button scenario, for example, think first of a case in which there is a button to press to kill all psychopaths, pressing which is *not* correlated with having psychopathic tendencies ('the Blue Button'). So, *if* your credence that you are a psychopath is sufficiently low it is rational to press it, even though, we can suppose, for the privilege of doing so you are charged a small fee (and, of course, even though if you are a psychopath you will be better off if you refrain than you would be if you were to press). Now suppose that there is in addition a second button ('the Death's Head Button'), which is free to press, and will kill all

[6] Not merely: 'not uniquely rational'. With respect to one-boxing and two-boxing(K) the agent foresees in the *mixed scenario* that whichever he chooses he will be able to say afterwards: 'I would have done better if I had chosen the other' (as in Newcomb's Firebomb, also) – neither is causally ratifiable.

[7] Note that I am here aiming at a *reductio*. I am not asserting that two-boxing(K) is a rational choice in the *mixed scenario*.

psychopaths, but pressing which is strongly correlated with, though, of course, it does not cause, being a psychopath (this is Ahmed's *Psychopath Button A* case). Exactly parallel reasoning to that just gone through yields the conclusion that if you are *required* to press a button (this corresponds to the *second variant* of the Newcomb scenario) then if two-boxing is the rational choice in the standard Newcomb scenario, the rational choice is to press the Death's Head Button and not the Blue Button. So, by the reasoning already gone through, in the corresponding mixed scenario, in which your credences are the same but you also have the choice of refraining, pressing the Death's Head Button is a rational choice if two-boxing is the rational choice in the standard Newcomb scenario. Whence we can conclude in an exactly parallel fashion that pressing the (Death's Head) Button is rational in Egan's Psychopath Button case, in which the option of pressing the Blue Button does not exist. This, in fact, is Ahmed's conclusion, as his title indicates, as noted previously.

III

But it cannot be right, pressing the Button is definitely inferior to refraining from pressing in the Psychopath Button Case and two-boxing is definitely inferior in Newcomb's Firebomb.

Let us go back to Newcomb's Firebomb. In this case I am virtually certain that if I take both boxes I will cause the incineration of the million dollars definitely in the opaque box and will end up with just a thousand dollars. So I cannot believe, before making the decision, as the agent does in the standard Newcomb scenario, that two-boxing is *the action which will get me more money than the alternative would no matter what I get* (though if I *am* confident that I will one-box I will believe that if I were to two-box I would get more money than I in fact will). I believe only that it will get me more money than the alternative would if no triggering device has been set up, but that it will get me less money than the alternative would if a triggering device has been set up. So I cannot explain beforehand why I will two-box if I will just by saying 'If I two-box I will get more money than I would if I were to one-box no matter how much I will get', as I can in the standard Newcomb scenario. I do believe that, as in the standard Newcomb case, if I decide to one-box I will then have to acknowledge 'If I were to two-box I would get more'. But I do not believe, as I do in the standard Newcomb case, that if I decide to two-box I will then be able to say 'If I were to one-box I would get less'. On the contrary, since I believe that there is definitely one million in the opaque box, I believe that if I decide to two-box I will then have to acknowledge 'if I were to one-box instead I would get more'. So the *only shadow of a reason* there is for thinking two-boxing rationally preferable in the standard Newcomb scenario, namely that, as the agent knows *beforehand*, two-boxing is *the action which will have a better result than the alternative would no matter what* (a million plus a thousand instead of a million, or a thousand instead of nothing) is not available in the case of Newcomb's Firebomb. That is to say, the only reason why anyone reasonable would *not* reject outright the rationality of two-boxing in the Newcomb case is that it meets a subjunctive aim which is *not* met by two-boxing in the case of Newcomb's Firebomb.

If I am very confident in Newcomb's Firebomb that I will one-box[8] and hence that no triggering device has been set up, I will believe that if I two box I will get $M+K and hence that if I two-box I will do better than I would if I were to one-box. But, given my beliefs about the predictor and the utilities I must be assumed to assign to $M and $M+K if Newcomb's Firebomb is to be, as required, a case in which (non-ratificationist)[9] evidential

[8] If I am not, then as in the alternative *mixed scenario* envisaged in note 5, one-boxing is uncontroversially the rational choice, as Egan notes ('Some Counterexamples', p. 110). Similarly, in the Psychopath Button scenario, as Egan also notes ('Some Counterexamples', p. 107), if I am sufficiently confident that I will press and so am a psychopath, refraining from pressing will be uncontroversially the rational choice.

[9] See Egan 'Some Counterexamples', p. 109.

decision theory unequivocally recommends one-boxing, I cannot regard it as rational to aim to get $M+K by two-boxing. For I am virtually certain that I will not get $M+K, that is, I assign an *unconditional* probability as close to zero as you like to getting $M+K. However, I am certain that I will get $M conditional on one-boxing (I assign a probability of one to getting $M conditional on one-boxing). So to regard it as rational to aim to get $M+K by two-boxing I must regard it as rational to pass up the certainty of something I value a great deal for a small chance of something I do not value a great deal more (to reiterate, if I do value $M+K a great deal more non-ratificationist evidential decision theory also recommends two-boxing). So I cannot regard it as rational to aim to get $M+K by two-boxing even if I am confident that I will one-box and that no triggering devices have been set up.[10] Hence I cannot regard it as rational to aim to get more by two-boxing than I would by one-boxing since I know that I will do that *just in case* I get $M+K by two-boxing.[11] So even if I am confident that I will, in fact, one-box and hence that no triggering device has been set up, I cannot regard it as rational to aim to get more by two-boxing that I would by one-boxing. So whether or not I am confident that I will one-box, I cannot regard it as rational to aim to get more by two-boxing than I would by one-boxing, which is the only aim by which I could justify two-boxing. On the other hand, the aim of getting at least a million provides a justification for one-boxing.

These considerations carry over mutatis mutandis to the case of the Psychopath Button. I believe that if I decide to refrain from pressing I will then have to acknowledge ‘If I were to press I would get a better outcome than I am going to, i.e., the death of all psychopaths as well as my survival’. But I do not believe that if I decide to press I will then be able to say ‘If I were to refrain I would get a worse outcome’ and I do believe that if I decide to press I must then acknowledge ‘If I were to refrain I would get a better outcome’, i.e., to live. So in this case also I do not believe before making my decision that pressing will get me a better outcome than the only alternative would no matter what. So I cannot explain *beforehand* why I will press the button if I will by saying ‘No matter what, if I press the button I will get a better outcome than I would if I were to refrain’. Nor, even if I am confident that I am not a psychopath and will refrain, can I regard it as rational to aim to get a better outcome by pressing than by refraining (knowing that by doing so I will be passing up the certainty of a lesser but still acceptable outcome), given the utilities I must assign to life in a psychopath-containing world and life in a psychopath-free world if the case is to be one in which (non-ratification) EDT unequivocally recommends refraining. For I know that I will achieve this better outcome by pressing just in case I ensure thereby both the death of all psychopaths and my own survival. But this is something which I am (virtually) certain will not happen whatever I do, given my beliefs about the correlation between psychopathy and pressing. So whether or not I am confident that I will refrain from pressing, I cannot regard it as rational to aim to get a better outcome by pressing than by refraining, which is the only aim by which I could justify pressing. On the other hand, the aim of surviving provides a justification for not pressing.

Egan is right, then, to say that in the cases he describes the analogues of two-boxing are not rational options, i.e., are definitely inferior to the other available options. But if so two-boxing is not the rational choice in the standard Newcomb scenario.

IV

[10] Recall that in the standard Newcomb case the two-boxer is *not* aiming 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.

[11] By contrast, in the standard Newcomb case the two-boxer knows that he will get more by two-boxing than he would by one-boxing *no matter what*.

But how can this be? I tentatively suggest that the explanation is that the only aim by which the agent in the Newcomb scenario 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.

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 to 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.

To see what matters here consider the following scenario. 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 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 achieve at most a million dollars. My monomaniacal aim of doing better than my rival would (in Lewisean terms, of outperforming even his *non-actual* counterparts in close worlds),[12] together with this belief,[13] rationally generates the aim of making more than a million. Thus it decisively rules out going for the easy $900K. But it does not generate any indicatively describable aim which justifies going for the three million deal rather than the two million 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. In the standard Newcomb scenario two-boxing can be justified only

[12] Compare this to the aim of outdoing all people of similar hair colour to my rival, which is a possible aim.

[13] Of course, what this monomaniacal aim is, and what my belief that my rival would achieve only $M amounts to (and so what evidence I would need for it), depends on how mad I am. Is it the aim of doing better than my rival would if he were to engage in the transaction in his usual way? Or even if he were to first take performance-enhancing drugs or a crash course on Bayes'theorem? Or even if he were first to make a pact with the Devil? That is, which of his counterparts I will want to outperform will depend on how mad I am.

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.[14] 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 which is rationally generated by this subjunctively described aim in conjunction with my beliefs by reference to which two-boxing can be justified over one-boxing, just as a choice between the honest two million deal and the dishonest three million deal cannot be justified in the business transaction case. Since I believe, once I have decided to two-box, that the opaque box is empty (because I believe that 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 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 $0 if I one-box and hence cannot deny that I could achieve *this* aim by one-boxing .[15] For when I decided to two-box (and so came to believe that I was going to do that rather than one-box) the probability of getting a million *conditional* on one-boxing did not become less for me than it was before, it just became undefined.

However, it is not important for my present purposes whether this suggested explanation of why two-boxing is not the rational option in the standard Newcomb scenario is acceptable. Even if it is not my main conclusion remains. Two-boxing is not the rational option in the Newcomb scenario. So two-boxing cannot be justified as satisfying the aim of doing what will get more money than the alternative would. Whether this is because this is a subjunctively described aim which cannot generate in conjunction with the agent’s beliefs an indicatively describable aim by reference to which two-boxing is justifiable, or for some other reason, is another issue.

Department of Philosophy
University of Nottingham
Nottingham
UK
Harold.noonan@nottingham.ac.uk

[14] In more precise Lewisean terms, of doing better than any non-actual person who is importantly similar to me, and lives in a world most closely resembling this one apart from his making the one-boxing choice.

[15] Since my evidence indicates that the opaque box is empty I must also of course believe, so long as I continue 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.